International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-10, August 2020

A Predictive Classification Method for Email Phishing Attacks using Random Forest and A-R trees

Sanjitha M, Sri Lakshmi J, S Sameeksha, Ravi V

Abstract: ***Cyber-attacks are the attempts made by an individual or an organization deliberately, to breach the information system mainly computers of another individual or organization. These attacks have risen in recent years due to various reasons posing the need for systems that can use adaptive learning techniques to detect and mitigate these attacks at an early stage. Phishing is one of the significant cyber-attacks. According to global security report 2019, phishing was the major cause of attacks in corporate networks. Phishing attack uses disguised email to achieve its goal. In this attack, attacker masquerade himself as a trusted individual or a company and trick the email recipient into clicking malicious links or attachments. The proposed method provides a testbed for detecting and mitigating various types of phishing attacks. Machine learning techniques are used to build an intelligent system which can detect phishing attacks. This application uses random forest algorithm with AR-Trees (acceptance-rejection tree algorithm) to determine the attacks by considering various datasets available online and new datasets dynamically constructed for making the system ready to mitigate future phishing attacks.***

Keywords: AR-Trees, Random Forest.

I. INTRODUCTION

Phishing attacks [1-5] are the ones in which carefully targeted digital messages are transmitted to fool people into clicking on a link that can expose sensitive data. Technical vulnerabilities can be used by attackers to construct far more persuading socially-engineered messages. This makes phishing attacks a major problem, and effective mitigation system is necessary. An example of Phishing; an email which appears to be from a legit website or a user's bank account or an internet service provider is sent to the user. Usually, confidential information such as Credit/Debit card number or password is asked to update user's accounts. As soon as the user clicks the URL link present in the email, it redirects the user to a website requesting confidential information which will be forwarded to the attacker. The main goal is to steal sensitive data such as credit card details or to install malicious software on the victim's machine. Phishing is a common cyber-attack that everyone must be aware of and they must learn to protect themselves from it. In this type of attack, the email/message appears to come from a legit source. Sometimes malware is downloaded in the target computer. Attackers get benefitted financially by having the victim's credit card information or other personal data. Sometimes, phishing e-mails are also sent to retrieve login details or other details of employees which can be used to attack against reputed companies. A phishing email is sent to the targeted victim. Our developed software will be able to predict that the email received by the victim is phishing and alarms him against the attack. The developed software makes use of machine learning techniques to predict a phishing email by using Random forest algorithm combined with AR-algorithm which helps the system to detect phishing email at high accuracy. Organization of the paper: Section II discusses the literature survey regarding various standard techniques as well as different approaches that have been taken for the prediction of the phishing email. In section III, we discuss the preliminary concepts behind this work and purpose of our techniques. In section IV, we discuss simulation experiments and in Section V, we discuss the results and in Section VI we conclude our work and give light to the future work.

II. LITERATURE SURVEY

M. Khonji et.al [6] has surveyed the various features needed to detect, correct and prevent a phishing attack and proposed various ways for detecting the phishing attacks using software systems. Srishti Rawal et.al [8] considered, a group of features that are frequently used by phishing attackers and collected from different literatures are used to classify emails into either phishing or non-phishing. For the implementation and testing of a machine learning algorithm datasets from ham corpora and phishing corpus was used. Random Forest classifier was used to test the algorithm. 10 fold cross-validations was used to train and test the classifier. Adwan Yasin et.al [7] presented an intelligent classification model for detecting phishing emails using knowledge discovery, data mining and text processing techniques

Revised Manuscript Received on August 30, 2020.
* Correspondence Author
Sanjitha M*, Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumakuru, Karnataka, India. E-mail: sanjsamith@gmail.com
Sri Lakshmi J, Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumakuru, Karnataka, India. E-mail: srilakshmij98@gmail.com
S Sameeksha, Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumakuru, Karnataka, India. E-mail: sameekshamakam@gmail.com
Ravi V, Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumakuru, Karnataka, India. E-mail: ravi@sit.ac.in

In this they have considered five classifiers namely J48, BayesNet, SVM, MLP, and Random Forest. They have also used a 10-fold cross-validation technique to avoid overfitting.

The results were evaluated using various performance metrics.

In literature survey [11,12] authors describe the use of machine learning algorithms like Random Forest and SVM classifiers to classify an email as phishing or ham. The features like domain count, number of links, presence of javascript, presence of form tag, presence of HTML, number of action words, presence of word Paypal, presence of word bank, and presence of word account were considered. The paper cites to have obtained classification results with a whopping accuracy of 99.8%.

Software called "Anti Phishing Simulator" was developed by authors in [16] giving information about difficulty in detection of phishing emails and how to detect them. With this software, phishing and spam emails are detected by looking at mail contents. Bayesian algorithm is used to classify emails.

Several phishing detection methods have been proposed to detect phishing attacks at the early stages of occurrence, but those methods are static and also require static methods to analyse the data set generated by various websites which the system has to rely on. Related work carried out by various authors simplifies the detection of phishing attacks using a combination of methods such as blacklisting, heuristic methods, ML based approaches which results in high false positives and false negatives rates. In our method, all the features are extracted directly from the email itself and hence there is no need to send queries to larger dataset thereby reducing the space and time complexity for the proposed approach. In the proposed approach the data sets are derived from both static and generated dynamically from various sources, the algorithm reads the test data set and validates the output using real time data sets.

III. PROPOSED METHODOLOGY

To predict an email as phishing, several features of the email are to be considered such as URL, presence of words provoking urgency and domain related information which can help in predicting a phishing email. The following classification of features helps us to develop automated system to determine phishing email.

A. ***URL Based Features:*** It is believed that phishing emails will have at least one phishing URL. We have considered a few major features to decide whether an URL is phishing or not. These include host, number of dots in URL, largest domain, URL token count, average token length, largest URL token, average path token, average domain token length, length of URL, presence of IP address in URL, presence of security sensitive words in URL, ASN number, whether URL is safe to browse, length of a host, rank of a host in 'alexa.com' website among others.

B. ***Content Based Features*:** Some mail contains information which may provoke urgency or fear: Body of the phishing email often contains some terms which will try to provoke users to click on the links provided in the email. These may include words like 'important', 'immediately', 'urgent', 'expires', 'sudden' etc.,

C. ***Domain Based Features:*** There are some domains (from address of email) frequently involved in phishing activities. Hence, the address of the email can also be considered for phishing prediction.

The proposed system uses the above features to predict and analyze whether the emails is sent from legitimate or fraudulent users who may cause various attacks on users. 3In the proposed work, the system dynamically constructs the dataset for the random forest algorithm using alexa.com and various public data set available in [9,10] which makes the system more robust and reliable for phishing attack detection. The figure below shows the proposed architecture which is configured for email servers such as gmail and outlook email servers, the system also uses random forest algorithm with AR trees to classify the emails as phishing or legitimate with high accuracy.

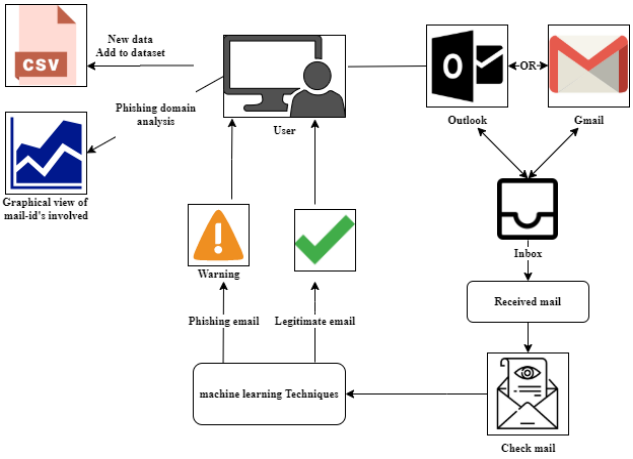

Fig.1. Proposed System

Random forest with AR Trees: Random forest is a well-known prediction and analysis technique used frequently in solving problems that contain interactions and missing observations with large number of mixed type variables which makes the algorithm computationally intensive. Hence we propose to use the variation of random forest algorithm proposed by Calhoun, P.[13] et.al where author constructs the random forest tree for the variable selection bias and applies the acceptance-rejection tree for the forest to improve the prediction rate. The proposed work is implemented completely by python and also uses the code available for AR trees in the github link [14] which can be called in the python code using the rpy2 [15] package which is used to call R inside python by installing the package using pip command ***pip install rpy2==2.3.0***.

IV. IV. SIMULATION EXPERIMENT

A. Data Used in the System:

a) A list of URLs consisting of both legitimate and phishing collected respectively from alexa.com and PhishTank websites are used for phishing URL prediction serving as a dataset.

b) A list of emails containing bag of words related to urgency or fear are sent to legitimate email to detect content based feature analysis.
c) A list of domain based featured mails are sent to legitimate email to detect domain based feature analysis.

B. Machine Learning Implementation:

The various features of a phishing URL, domain related features and content related are extracted programmatically for all the emails sent to a legitimate user and stored in the dataset and data set is dynamically updated for every new mail. The machine learning algorithm, Random Forest Classifier is used as a classifier. Marchal S et.al [16] data sets which provide standard dataset for classification of phishing mails are also used.

URL dataset: URL dataset consists of more than 22 features related to domain and URL giving various information which helps in detecting a phishing email.

Random forest with AR trees Implementation: Random forest algorithm constructs set of trees against a target attribute and more informative subset of features are extracted out of attribute selected. The attributes which results in best splits in the tree are considered for future analysis. Attribute score is calculated for each tree and most predictive attribute is determined based on highest score of the attribute. The extension for the random forest is the AR-trees (acceptance and rejection trees) which improves accuracy and reduce variable selection bias.

Suppose the data consists of independent identically distributed (i.i.d) observations $\{(x_i, y_i) : i = 1,...,n\}$ where y_i is the ith observed response and $x_i = (X_{i1}, ..., X_{ip})$ is a p-dimensional input variable vector. Consider the regression model:

$$y_i = \beta_0 + \beta_1 I(x_{ik} \leq c) + \epsilon_i \quad \text{for k } 1, \ldots, p$$

where I ($x_{ik} \leqslant c$) is the indicator function corresponding to the split based on the k^{th} input variable and i i.i.d.N $(0,\sigma^{2})$ The algorithm works on a randomly selected one variable as the tree grows and one cutpoint is randomly chosen for the variable at each internal node in the tree. If the randomly chosen split doesn't satisfy the acceptance-rejection criterion *(p)* then the above process is continued until the cutpoint results in a variable which is equal or less than *p*. Use of AR-trees helps to improve the accuracy since at every internal node the binary decision has to be taken by the algorithm to either accept or reject the variable.

V. RESULTS

This application helps users to detect a phishing email and thus alarming them against a phishing attack. It predicts whether an email is phishing using the machine learning model developed with a prediction accuracy of 98.3% using the combination of random forest and AR-trees.

VI. CONCLUSION AND FUTURE WORK

The global security and the economy have been suffering a lot from phishing attacks these days. There has been a fast rate of increase in phishing attacks affecting human life. This application helps in mitigating phishing attack. In this paper, firstly all the features of phishing email including URL based, content-based and domain based are considered. All these features considered are programmatically extracted from the email to be classified and stored in database which is used by Random Forest and A-R tress for classification, thus detecting a phishing attack. Detection paves way for mitigating such harmful attacks. Hence, this method helps in predicting a phishing email with a classification accuracy of 98.3%.

Images in the email can be used as a feature in predicting a phishing email which is open for future implements. The phishing email sometimes may include an attachment which attempts to steal sensitive data or sometimes may spread malware in victim's computer. This feature is not used since we could not find a good way to extract the attachments in the email, but can be implemented in future.

REFERENCES

1. Gupta BB, Tewari A, Jain AK, Agrawal DP. Fighting against phishing attacks: state of the art and future challenges. Neural ComputAppl. 2017;28(12):3629–54.
2. The Phishing Guide Understanding & Preventing Phishing Attacks By: Gunter Ollmann, Director of Security Strategy, IBM Internet Security Systems, 2007
3. Phishing: Cutting the Identity Theft Line Published by Wiley Publishing, Inc. 10475 Crosspoint Boulevard Indianapolis, IN 46256 www.wiley.com, 2005, Rachael Lininger and Russell Dean Vines.
4. Anti-Phishing Working Group (APWG), "Phishing activity trends report—first quarter 2013. http://antiphishing.org/reports/apwgtrendsreportq12013.pdf, accessed September 2014
5. Pierluigi Paganini (2014) Phishing: a very dangerous cyber threat. http://resources.infosecinstitute.com/phishing-dangerous-cyber-threat/ 2012. Accessed on Sept 2014
6. M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: a literature survey," IEEE Communications & Surveys Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013. View at: Publisher Site | Google Scholar
7. Adwan Yasin and Abdelmunem Abuhasan.An intelligent classification model for phishing email detection,2016.
8. Srishti Rawal, Bhuvan Rawal, Aakhila Shaheen and Shubam Malik .Phishing detection in emails using machine learning.International Journal of Applied Information Systems,12:21-24,10 2017.
9. https://web.cs.hacettepe.edu.tr/~selman/phish-iris-dataset/
10. https://www.phishtank.com/developer_info.php
11. Andronicus A. Akinyelu and Aderemi O. Adewumi. Classification of phishing email using random forest machine learning technique. Journal of Applied Mathematics, 2014:425731, Apr 2014.
12. Calhoun, P., Hallett, M.J., Su, X. et al. Random forest with acceptance–rejection trees. Comput Stat (2019). https://doi.org/10.1007/s00180-019-00929-4
13. https://github.com/pcalhoun1/AR-Code
14. https://sites.google.com/site/aslugsguidetopython/data-analysis/pandas/calling-r-from-python
15. https://research.aalto.fi/en/datasets/phishstorm-phishing-legitimate-url-dataset
16. Muhammet Baykara and Zahit Ziya Gürel. Detection of phishing attacks. 2018 6th International Symposium on Digital Forensic and Security (ISDFS).

AUTHORS PROFILE

Sanjitha M, Dept. of CSE, Siddaganga Institute of Technology, Tumakuru, Karnataka, India.

Educational details:

- B.E in Computer Science and Engineering from Siddaganga Institute of Technology, Tumakuru, Karnataka (2016-2020)
- PU from Mahesh PU College, Tumakuru, Karnataka (2014-2016)
- X from MaruthiVidya Kendra, Belagumba, Tumakuru, Karnataka (2014)

Projects:

- **"*Find Your Bread*",** mini project carried out under the guidance of Shruthi K, Assistant professor, Siddaganga Institute of Technology, Tumakuru (2019)
- **"*Smart Waste Management*",** an IoT project carried out in 2018.

Achievements:

- Participated in **WIE CODE**, a 12 hour **hackathon** conducted at Siddaganga Institute of Technology, Tumakuru on 12th October 2018.
- Was amongst top 60% teams in the coding competition, **CoDecode**, during Technorion, nationwide zonal competitions of Techfest, **IIT Bombay**, 2018-19.

Membership:

- Member of PATHFINDER, a student voluntary organization in Siddaganga Institute of Technology, Tumakuru.

.

Sri Lakshmi J Dept. of CSE, Siddaganga Institute of Technology, Tumakuru, Karnataka, India.

Educational details:

- B.E in Computer Science and Engineering from Siddaganga Institute of Technology, Tumakuru, Karnataka (2016-2020)
- PU from Mahesh PU College, Tumakuru, Karnataka (2014-2016)
- X from Bishop Sargant School, Tumakuru, Karnataka (2014)

Projects:

- **"*Find Your Bread*",** mini project carried out under the guidance of Shruthi K, Assistant professor, Siddaganga Institute of Technology, Tumakuru (2019)
- **"*Smart Waste Management*",** an IoT project carried out in 2018.

Achievements:

- Participated in **WIE CODE**, a 12 hour **hackathon** conducted at Siddaganga Institute of Technology, Tumakuru on 12th October 2018.
- Was amongst top 60% teams in the coding competition, **CoDecode**, during Technorion, nationwide zonal competitions of Techfest, **IIT Bombay**, 2018-19.

S Sameeksha Dept. of CSE, Siddaganga Institute of Technology, Tumakuru, Karnataka, India.

Educational details:

- B.E in Computer Science of Engineering from Siddaganga Institute of Technology, Tumakuru, Karnataka (2016-2020)
- PU from Mahesh PU College, Tumakuru, Karnataka (2014-2016)
- X from Ankur High School, S.S.Puram, Tumakuru, Karnataka (2014)

Projects:

- **"*Find Your Bread*",** mini project carried out under the guidance of Shruthi K, Assistant professor, Siddaganga Institute of Technology, Tumakuru (2019)
- **"*Funfair park webapp*",** a Flask project carried out in 2019.
- ***"E-Library Management webapp",*** a Django project carried out in 2019.

Achievements:

- Industrial Training in Monkfox in field of Python, Flask, Django with Machine Learning.

Ravi V, Assistant Professor, Dept. of CSE, Siddaganga Institute of Technology, Tumakuru, Karnataka, India.

Educational Qualification:

- M.Tech (Computer Science and Engineering) from VTU.
- B.E. (Computer Science and Engineering) from VTU, SIT Tumakuru.
- Diploma in Computer Science and Engineering (Board of Technical Education)

Publications in National / International Conferences and Journals:

1. Dr. Chandrashekara and Ravi.V, "Agent Based group scheduling policy for Computational Grids", National Conference on Communications, Acharya Institute of Technology, Bangalore.
2. Premalatha.G and Ravi.V, "Availability for Efficient Audit Services in Cloud Computing", International Conference on Computational Intelligence and Engineering Applications ICIEA 24th march 2013 ,Thirupathi.
3. Sanjay.M and Ravi.V, "Enhanced ODMRP for MANET's to protect against Multicast Attacks", International Conference on Information Technology and Communication Engineering (ICITCE-2014), 5th July 2014, Bangalore.
4. Naveen A and Ravi.V"Client Side Deduplication Scheme for Secured Data Storage in Cloud environments" International Journal of Engineering Research and Technology (IJERT), May 2015.
5. Ravi.V and Aparna.R, "Security in RFID based Smart Retail System" , 10th INDIACom INDIACom-2016; IEEE Conference ID: 37465 3rd International Conference on "Computing for Sustainable Global Development, 16th- 18th March 2016BharatiVidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA).
6. SushmaVerma (DRDO), Dr. N.R.Sunitha and Ravi.V, "Strand Space Approach for Formal Verification of Security Protocols", Published in DRDO magazine "CHAKRAVYHA", 2016.

Research Proposals:

"***Formal verification of Security Protocols in Network Devices***" approved by DRDO, SAG, New Delhi.
Duration =2 yrs (2016-2018)
Amount = 22 Lakhs
Principal Investigator: Dr. N. R. Sunitha , Professor, Dept. of CSE, SIT.
Co-Investigator: Ravi.V, Assistant Professor, Dept. of CSE, SIT.

Professional Memberships:

- Life Time member of Computer Society of India (CSI).
- Student membership IEEE.org and ACCS.org