International Journal of Research in Advent Technology, Vol.7, No.1, January 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Performance Evaluation of Classification Techniques for Phishing Attack Detection

Dr. Anshu Chaturvedi [1], Prof. D.N. Goswami [2], Manali Shukla[3]
[1] *Department Of Computer Applications,MITS, Gwalior 474005, India,*
[2]*S.O.S.in Computer Science And Applications,Jiwaji University, Gwalior 474001, India,*
[3] *S.O.S.in Computer Science And Applications,Jiwaji University, Gwalior 474001, India*
Email: anshu_chaturvedi@yahoo.co.in, goswamidn@yahoo.com , shukla_manali@rediffmail.com

Abstract- Phishing attack has been a concerning threat for security experts over the years. The rapid increase and advancement of phishing methods produce a vast challenge in the field of web security. Although several research works has already done and various security mechanisms has been implemented in this field but still people are becoming victim of this attack. Therefore, still there is a need of some productive techniques which can prevent phishing attacks. Broadly Phishing attack exists into two forms first is in the form of phishing emails and secondly in the form of phishing websites. This paper evaluates and studies various classification algorithm performances for the detection of phishing websites. Experimental work is carried out using the data set of phishing websites from UCI Machine Learning Repository. Performance comparison among different classification algorithms is done by using Weka 3.8. tool over the dataset.

Index Terms- Phishing; Datamining; Weka; Malicious; Phishers; Blacklist; Whitelist

1. INTRODUCTION

PHISHING is an illegal action by the Phishers to steal user credentials e.g. passwords username, financial ID's and personal data with a aim to use user's data for the accomplishment of malicious activities and make money. PHISHING is implemented by making an illegitimate website which is look alike or replica of a legitimate website. Phishers make all effort to fool users in believing that he/she is using the original website and eventually user get trapped by the phishers, submit their passwords to a malicious website unknowingly. Phishing detection mechanism can be divided into two categories list based and heuristic based techniques [1]. List -based phishing detection mechanism simply classify a website as phishing or trusted through a visit in database. List-based approaches further can be divided into blacklist and white list. Blacklists contains URLs that refer to websites that are considered phishing or malicious and White - list encompasses trusted or legitimate websites. Black list based mechanism works on the principle of list lookup , the browser queries the blacklist while loading the webpage to find whether that webpage is blacklisted or not ? In case if it is on the list then, the website is considered as illegitimate Otherwise, the page is considered legitimate. On the other hand Heuristic-based techniques test some feature of a website. These features includes uniform resource locater (URL) based features e.g its length , domain registration length, spelling errors, embedded links, and host based features etc. In contrast to list based mechanism heuristic based scheme could detect new phishing websites which black list approach could not. Therefore, the features of Phishing websites can be used for the detection of malicious websites. Previous research has significantly illustrates the impact of Data Mining techniques in phishing attack detection. Classification is one of the prominent data mining technique which is also known as supervised learning where for a given input value the preferred output is known. "Classification is the process of finding a model that describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown"[2]. The model which is derived is dependent on the critical study of a training data. Thus classification can be implemented to classify a website as phishing or non phishing site. Numerous classification methods are there which includes, Bayesian, decision tree, k-nearest neighbor classifier, case-based reasoning etc. Classification is implemented with two types of data one is called training data and second is called testing data. Training data needs to be preprocessed if it is noisy and not suitable for the mining task for e.g. if it contains missing values, inconsistent values etc. In such a case, data preprocessing task is to be carried out to remove all sort of noise as well as required filters can be applied on it. Classifier model is build by applying appropriate algorithm on training data and then the precision of a classifier is evaluated by using test data in the form of percentage of test data tuples that are rightly identified by the classifier.

1.1 Related Work

Several research works has been done in this field, numerous methodologies are being implemented to detect phishing *attack.* Another classification of phishing detection mechanisms is content based and non – content based approaches. Content based approach detects malicious attack by investigating content based features of website which include links, spelling errors, keywords, fields of password, embedded links, etc. in conjunction with uniform resource locator and host based features as proposed by [3]. Phishing attempts and malware can also be detected by Google's anti-phishing filter by determining webpage uniform resource locator, rank, WHOIS information and contents of a webpage along with JavaScript, HTML, iframe, images, etc. as given in [3]. Non-content based approaches suggested in [3] are mainly based on uniform resource locator and classification of host information. An anomaly detection system proposed by the author in [4] based on the combinational approach of K -means and ID3 algorithm for classifying the two clusters for classifying the normal and anomalies activities. A completely unique HTML content method is proposed by [5] for determining malicious websites. It evaluates the code of a webpage and uses TF-IDF to find the keywords which have highest rank. In their approach Google search engine takes those keywords as input and identify a match for the domain name of uniform resource locator with the top search result, On finding a match it will be considered as legitimate. This technique is completely based on google search engine. The upgraded version of CANTINA was proposed by the author in [6] where additional features are added to get improved results. In their proposed work the author make use of the HTML Document Object Model, third party and Google search engines along with machine learning technique to determine malicious web pages. The author in [7] proposed use of lexical and host-based features of the associated Uniform resource locators in their online learning technique for identifying malicious Web sites. Their work is specially suitable for online algorithms. SVM based technique was proposed by the author in [8] to identify malicious uniform resource locator and used features such as structural, lexical and brand names that present in the Uniform resource locators.

Effective feature selection procedure was proposed in [9] for improved phishing detection. In the approach proposed by the author in [10], when the user submits the user login details for the first time, system gives warning to the user, although the current website is a legitimate website. This happens because the information about the authentic website is not managed. The author in [11] proposed Clustering and Bayesian approach and implemented at client side. In this approach database is clustered using K-Means Clustering and Naive Bayes Classifier prediction technique to find the probability of the web site in the form of Valid Phish or Invalid Phish. For a given site first the URL features are extracted, then k-. means clustering algorithm is applied on them to categorize them into Phish or non Phish. Two new contributions was made by the approach suggested by [12] for determining malicious uniform resource locator in which they proposed to extract lexical patterns from uniform resource locators dynamically instead of examining pre-existing features or fixed delimiters for feature selection, A hybrid approach was proposed by the author in [13] that combines extraction of key phrase, textual, financial data to discover the maliciousness of phishing attack using classification methods based on supervised technique. Pshark is an approach proposed by the author in [14] to determine and remove the identified malicious web page from host server. Information of the webpage is retrieved from the WHOIS database. In order to inform the host server about the presence of phishing web page in that server a notification is sent.

In the research work done by the author in [15] used domain feature enhanced model of classification for the identification of Chinese phishing e- business websites.

1.2 Data Set

This research work is carried out using the data set of phishing websites from UCI Machine Learning Repository[16].This dataset was collected from mainly phishtank archive ,Millersmiles archive and google's searching operators[17].This dataset contains 31 phishing websites attribute including one result attribute to classify a website phishy or non – phishy. The dataset has 2456 observations. All instances are categorized using binary values 1 for legitimate, 0 for suspicious and -1 for illegitimate.

1.3 Phishing Website Attributes

The data set which is used has 30 attributes of phishing websites and 1 result attribute which is used to classify a website phishi or non – phishi. This segment describes the most common features that are used in the security research domain to distinguish between a legitimate and phishing websites as suggested in [18].The impact of these features for identifying Phishing attack is discussed below -

. 1.3.1 Attributes of Phishing Websites Dataset

1. Modifying number of characters limits of Uniform Resource locator: Generally attackers make changes in the length of uniform resource locator in order to cover the suspicious section of the uniform resource

locator. Attackers use redirection links which takes users's webpage to suspicious and harmful domains. Previous studies in this field identified that acceptable length of URL is not more than 56 characters.

***2. Occurrence of links in Uniform resource locator*:**
Attackers use this feature that redirect web users to a suspicious domain or malicious domain from the legitimate domain which was specified in the address bar of uniform resource locator by the web user. This feature is also knows as Anchor.

3. Turning up non requested windows: This type of feature is used by the intruders to get user's credentials data. Such types of windows simply appears like a pop up messages without any request by the web user and invites user to input data which will be submitted to malicious domain**.**

4. Number of double slashes for Redirection: Intruders uses Redirection links by adding number of slashes in the uniform resource locator. This technique used by them to spoof web users.

5. *Presence of internet address in Uniform resource locator:* Presence of internet address in the uniform resource locator may lead the web user to phishing attack.

6. Keeping empty form data: *Intruders generally keep server form data empty so* that whenever web user send data on that webpage, attacker modifies the domain in the empty section of form.

***7. Adding special characters to uniform resource locator*:** Attackers add some special characters in the beginning and ending of uniform resource locator with a aim to spoof users.

8. *Numerous additions of sub domains:* If a uniform resource locator contains multiple sub domains then that uniform resource locator cannot be considered as legitimate one and may result into a phishing website.

9.Domain Name System: To verify the legitimacy of a domain, domain name systems are used which determines whether a domain is live or not? Generally superfluous domains are unavailable on the Domain name system. Attackers quickly steal web user's data because their domain life span is almost less than three days.

10. **Presence of internet protocol & Certificate:** Hypertext transfer protocol plays an important role in giving the idea of website authenticity. This is not only sufficient because presence of certificate assigned to hyper text transfer protocol and its extent is also significantly important.

11. *Using special symbol:* Phishing website uses at the rate "@" symbol in uniform resource locator address due to which browser overlook any thing written before that symbol. Legitimate address generally written after the "@" symbol.

12.Request URL: The website can also be considered malicious when the object of existing webpage are found to be loaded from a server different as that of specified in the uniform resource locator.

13.Checking URL on WHO-IS *:* Irregular URL'S are not on the list of WHO-IS database. Therefore a test needed to be done to examine whether the current browsed website is inside the WHO-IS database or not to determine the legitimacy of the website.

14. Disabling facility of Right Click: Attacker hides the *legitimate* links and display malicious ones to deceive online users. This method can be implemented by chasing the mouse cursor movements and once it arrives to the fake link the status bar content is altered. When the property "Right Click" is disabled this is an indication of phishing.

15.Lifetime of Domain : Lifetime of domain plays an important role to identify whether a website is legitimate or illegitimate . This feature is considered for identifying malicious websites because lifetime of phishing domains are generally not so long.

16. *Website Access by Visitors :* A website which has higher number of users or visitors can be considered safe and users can browse safely On the other hand Phishing websites normally have low browsing traffic and which can be checked through the rank inside Alexadatabase.

17. *Using tiny uniform resource locator:* Attackers often shorten the standard length of uniform resource locator in order to execute their malicious WebPages.

18. *Registration time span of domain.:* *Registration of Legitimate domains are often deposited in advance on the other hand* illegitimate websites use a domain, which is currently registered , and their lifetime is not long.

19. Use of Graphic image*:* A graphic image associated to a specific webpage is known as fevicon. This graphic image is also used by several browsers like a visual reminder of the website identity in address bar. When this graphic image is loaded from a domain other than that displayed in the address bar, then the webpage may be considered as malicious.

20. Presence of internet protocol in domain*:* Attackers use internet protocol in the domain part of uniform resource locator to hide their malicious intentions.

21. Redirecting to an Email: Attackers uses special function of scripting languages in order to redirect web surfers into a desired email. A phisher might redirect the user's information to his/her personal email. This method of spoofing is based on a server-side scripting language.

22. *Count Of Webpage Redirecting* : When a webpage repeatedly redirect their user's to specific

uniform resource locator then that website can be considered as phishing website because genuine websites not have more than one redirection of web pages.

23.Change in the onMouseOver Event: If there is any change observed in the "onMouseOver "event on the status bar then there is a possibility of phishing.

24. Linking Of Tags: Generally tags of Legitimate websites are linked to the identical domain of the webpage, If these tags are linked to a domain different than that of a webpage then this can be considered as mark of malicious website.

***25. Redirection Of IFrame* :** This tag is used to show an additional webpage within the existing webpage. Web users are deceived when this frame is found missing.

26.Ranking: Ranking accounts the worth of a webpage in the World Wide Web. Generally malicious WebPages posses low rank.

27.Indexing: One of the aspects of website security is indexing by the Google. A malicious website can be identified by checking its status of being indexed by the Google.

28. ***Opening & Blocking Of Port:*** Servers are controlled by opening and blocking ports by the security administrators when these ports are open attackers can use this as a opportunity to achieve their malicious goals.

29. ***Pointing Links Of Website:*** Website legitimacy can also be determined by the number of links pointing to a website because malicious websites have at most one link or even zero due to their little existence in the network.

30. Phishing Forums Reports : There are number of communities and forums which produce list of malicious internet addresses and domains in their annual statistical reports. If any host found top ranking in these reports then that host can have malicious intentions.

2 DATA PREPROCESSING

The data set is in the form of .arff file i.e. attribute relation file format. This format is suitable for Weka tool for classification [19]. Thirty attributes of data set have been selected for building different classifiers. All attributes values are in binary form. Data set is divided into training and testing data in the ratio of 70-30 using splitting in weka. Training data contains 70% of data set which is used for building the classifier whereas 30% of dataset is used as testing data.

2.1 Data Transformation

Data Transformation step is applied for converting data into a form appropriate for Mining which includes -

- smoothing
- Aggregation
- Generalization
- Normalization
- Attribute Construction

Since data set is already in the appropriate form for Mining as well as normalized. Therefore it is in appropriate form, as well as ready for the application of different classification algorithms in order to assess their results. In the next segment, we analyzed the results obtained after the application of different classification algorithms.

2.2 Experimental Work

This segment analyses and compares the performance of different classification algorithms. 10 Cross fold validation is used as test option to overcome the problem of over fitting as suggested in [20]. This method makes predictions more general as compare to holdout method by reducing variance among data. Following algorithms are being applied on the data set -

- Naïve Bayes Classifier
- J48 classifier
- Random Forest Classifier
- IBK lazy Classifier

1. ***Naive Bayes Classifier output*:**

Model building Time: 0.05 seconds

Instances Classified Correctly: 10279 92.9806 %

Instances Incorrectly Classified: 776 7.0194 %

2. ***J48 classifier Output*:**

Model building Time: 0.59 seconds
Instances Correctly Classified 10599 95.8752%
Instances Incorrectly Classified 456 4.1248 %

3. ***Random Forest Classifier Output :***

Model building Time: 3.45 seconds
Instances Correctly Classified 10752 97.2592 %
Instances Incorrectly Classified 303 2.7408 %

4. ***IBK lazy Classifier output:***

Model building Time: 0.03 seconds
Instances Correctly Classified 10743 97.1777 %
Instances Incorrectly Classified 312 2.8223 %

Table 1. Summary of Result

Total Instances	Classifier	Accuracy
1155	Naïve Bayes Classifier	92.9806 %
1155	J48 classifier	95.8752 %
1155	Random Forest Classifier	97.2592 %
1155	IBK lazy Classifier	97.1777 %

Furthermore these four classifier are tested to compare their performance using t- test against two measures –

(i) Percent correct and
(ii) F – Measure.

Performance evaluation using Percent _ correct and F – measure as comparison field is given below in the following table -

Table 2. T-test result on Algorithms using Percent_Correct and F- measure as evaluation measure.

Algorithms	Percent_Correct	F-Measure
Naïve Bayes Classifier	92.94	0.92
J48 classifier	95.90 (v)	0.95
Random Forest Classifier	97.24 (v)	0.97 (v)
IBK lazy Classifier	97.14 (v)	0.97 (v)

These results clearly show that Naïve base classifier performance is not as much of other classifiers. Among the classifiers the performance of Random Forest classifier is most significant. The performance of IBK lazy Classifier is also good but it gives more false positives whereas the number of false positives of Random forest algorithm is also less than the other classifiers.

3. CONCLUSION

This research work has been carried out using dataset of UCI machine learning and uses interface of WEKA to evaluate the performance of various types of classifier over the given dataset. This performance comparative study is carried out to determine the best available classifier as well as to get deeper sight of their performance in the form of accuracy. In this process of exploration Random Forest Algorithm displayed remarkable performance with an accuracy of 97.2592 % over the phishing data set.

REFERENCES

[1] Mona Ghotaish Alkhozae and Omar Abdullah Batarfi: Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code. International Journal of Information and Communication Technology Research, Volume 1 No. 6, October 2011..

[2] Jiawei Han, Micheline Kamber, Jian Pei: Data Mining Concepts and Techniques .3rd ed, 22nd June 2011.

[3] S. Afroz and R. Greenstadt, PhishZoo: Detecting Phishing Websites by Looking at Them, *2011 IEEE Fifth International Conference on Semantic Computing(ICSC)*, Palo Alto, California USA, 2011,pp.368-375.doi:10.1109/ICSC.2011.52

[4] K. Hanumantha Rao, : Implementation of Anomaly Detection Technique Using Machine Learning Algorithms. International Journal of Computer Science and Telecommunications [Volume 2, Issue 3, June 2011].

[5] Zhang Y, Hong JI, Cranor LF (2007) CANTINA: a content based approach to detecting phishing web sites. In: Proceedings of the 16th International conference on World Wide Web, Banff, p 639 – 648.

[6] Xiang G, Hong J, Rose CP, Cranor L (2011) CANTINA+: a feature-rich machine learning framework for detecting phishing web sites. ACM Trans Inf Syst Secur 14:21.

[7] Justin Ma Lawrence K. Saul Stefan Savage Geoffrey M. Voelker,:Identifying Suspicious URLs An Application of Large-Scale Online Learning.Proceeding of 26th International Conference on Machine learning Montreal Canada 2009.

[8] Huang H, Qian L, Wang Y (2012) :A SVM based technique to Detect phishing URLs. Int Technol J 11(7):921–925.

[9]. R. Basnet, A. Sung, and Q. Liu, :Feature selection for improved phishing detection. Advanced Research in Applied Artificial Intelligence, pp. 252–261, 2012.

[10] Han W, Cao Y, Bertino E, Yong J (2012).:Using automated individual white list to protect web digital identities. Expert Syst Appl 39:11861–11869.

[11] Prof. Rahul Patil, Bhushan Dasharath Dhamdhere, Rohit Gopal model kaushal sudhakar dhonde and swapnil balasaheb mehetre : a hybrid to Detect Phishing-Sites using Clustering and Bayesian Approach.International Conference for Convergence of Technology 2014.

[12] Da Huang Kai Xu, and Jian Pei: Malicious URL Detection by Dynamically Mining Patterns

without. Pre-defined Elements .Simon Fraser University.

[13] Chen X, Bose I, Leung ACM, Guo C (2011): Assessing the severity of attacks: a hybrid data mining approach. Expert Syst Appl 50:662–672.

[14] Shah R, Trevathan J, Read W, Ghodosi H (2009): A proactive approach to preventing phishing attacks using Pshark. In Sixth international conference on information technology new generations. IEEE, Las Vegas, pp 915–921

[15] Zhang D, Yan Z, Jiang H, Kim T (2014) :A domain-feature enhanced classification model for the detection of Chinese phishing e- business websites. Inf Manag 51:845–853

[16] Lichman, M. (2013). :UCI Machine Learning Repository University of California, School of Information and Computer Science. [http://archive.ics.uci.edu/ml].

[17] Data for Fraud websites,https://www.phishtank

[18] Abdelhamid N., Ayesh A., Thabtah F. (2014): Phishing detection based associative classification data mining Expert Systems with Applications 41(13) pages 5948– 5959, Oct 2014

[19] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016): The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning and Techniques Morgan Kaufmann, Fourth Edition, 2016.

[20] S.Yadav,S.Shukla(2016): Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification Advanced Computing (IACC), 2016 IEEE 6th International Conference.