Impact Factor (SJIF): 4.542

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 4, Issue5,May-2017

A Novel Approach for Phishing Website Detection using Hybrid Technique of Rule Mining and SVM

Binal Masot[1] , Riddhi Kotak[2], Mittal Joiser[3]

[1] Research scholar, Computer Engineering MEFGI, Rajkot
[2] Assistance prof., Information technology MEFGI, Rajkot
[3]Assistance prof., Computer Engineering MEFGI, Rajkot

Abstract- ***With the quickly increase development of website and web application, internet user utilize that benefits. They make their all day to day daily life activity like reading a newspaper, online shopping, online payment etc. Hence the chances of the users to get caught in the web threat its called phishing attack. There for the phishing detection is necessary. There is no conclusive solution to detect phishing. In this paper we present novel technique to detect phishing attack and compare with the other existing technique. Our proposed framework work on combine algorithm of rule mining and SVM.***

Keywords: Data mining, feature extraction, legitimate, machine learning, SVM, phishing

I INTRODUCTION

Now a day the most profitable threat is phishing. Phishing word come from the words fishing + phreaking fishing means use bait to induce the target and phreaking. The word "phishing" was first used in 1996 over the internet by a group of hackers who stole America online (AOL) accounts. By tricking unaware AOL users into disclosing their passwords [1]. The main target of phishing attack is to steal private information of victim. Last few years phishing quickly spread posing a real threat to universal security.

Phishing is a form of web threat so indirectly get the information of user like username, password, and credit card details etc. phisher will create a replica of legitimate site as a target website and disclosing personal data or credentials. The different technique of Phishing by send email of fake site URL hyperlink, instant message, website and SMS.

In this paper, include overview of phishing attacks, methodology used for detection an attack, proposed solution with result and comparison with other technique.

II RELATED WORK

B.B.Gupta et al.[1] propose the survey on fighting against phishing attack. They give the various challenges and available solution.

Jeeva et al.[3] propose an approach is based on the association rule mining to detect phishing URL. This approach in two phase in first phase they search URL and in second phase they extract the features. The result show that the proposed method achieved overall 93% accuracy.

Ramesh et al.[5] proposed an anti-phishing technique using target domain identification algorithm. In this algorithm they take a groups the domain from hyperlinks having direct or indirect association with the given suspicious webpage. The result show that the proposed Method achieved 99.65% accuracy on google.com search engine, 99.6% on aol.com search engine, 99.55% on hotbot.com search engine, 99.45% on bing.com search engine.

Mahmoud et al.[11] proposed a URL based phishing detectors. In this paper they detect the fraud domain name from the URL and compare with the page rank of the URL. If the page rank difference of the original domain and phishy domain is lagre then the URL categorized in form of phishing. They used FP ratio of the evaluation matrix.

Shrestha et al.[10] use a multi-modal feature classification algorithm to classify phishing and legitimate sites. A multi-modal classification algorithm use two types of features visual base and text base for classification. They take a visual feature from the snapshot of the fraud website and text features from the code of the website. In this paper used a Map Reduce framework. In this paper also give the various challenges and available solution.

III METHODOLOGY

3.1. SVM Theory

Support vector machine (SVM) used as the classifier for the software defect detection. Support vector machine is supervised learning algorithm.

SVM (Support vector machine) are supervised learning algorithm that used for classification and regression analysis. It is a binary linear classifier. A SVM model is a depiction of the examples like points in space, so that the examples of the different analysis are divided by a relevant gap that is as broad as possible. SVM are very helpful in text mining. SVM uses mapping into a space so that cross product can be easily computed in terms of original space. SVM (Support Vector Machine) is a useful technique for data classification. Its classify data into two class.

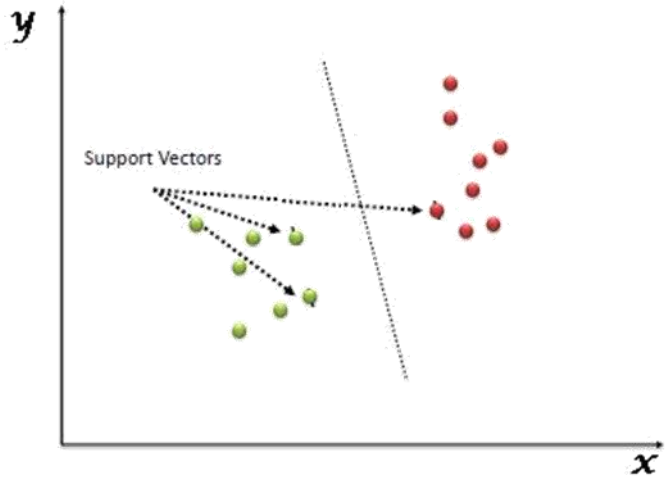

Figure 1. SVM

In the above figure we show that the data set are classified into two classes. The one classes is dissimilar to other class. So in same class there are similar data and other class have dissimilar data than one class.

Any of the data set SVM work same as above. In this classified using the hyper plan. The result is good if the maximum margin as broad as possible and relevant gap is maximum.

3.2. Multilable Rule Mining

In this data mining algorithm that extract information into 3 classes. This algorithm used if then else rules. This algorithm is same as the association rule mining algorithm. In this algorithm the confidence and support is count using of classification. Confidence is the conditional probability of C with reference to A or, in different words, the relative cardinality of C with reference to A. Apriori is an important algorithm for mining frequent item sets.

This algorithm uses past information of frequent item set properties. To select fascinating rules from the set of possible rules, constraints on numerous measures of significance and interest are used. Support and confidence are the measures of rule that replicate the quality and certainty of a rule. Apriori grades the rules with respect to confidence alone but predictive apriori deliberates the confidence and support together in ranking the rules. The support and confidence is joined in a single measure called accuracy. In our frame work it used for basic classification of mix data set to into three data set. The if-else rule used for features classification.

IV PROPOSED SOLUTION

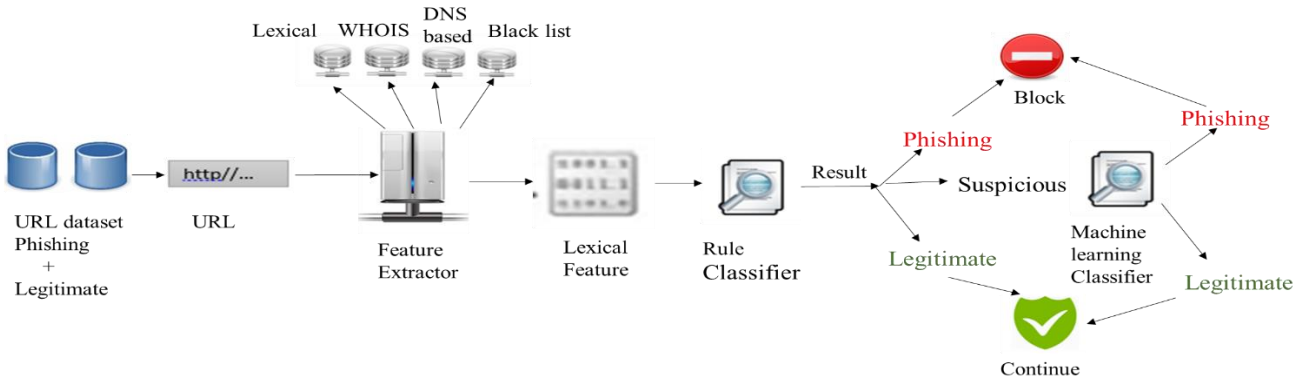

Figure 2. Proposed Framework

Figure 2 shows that the flow proposed system. This system have two parts in firsr part we classify the data into three lable phishing suspicious and legitimate and the second part we classify the suspicious URL into phishing and legitimate. In this system database taken from Phish Tank API and Alexa dataset.

Our system works on combination of rule mining [3] and SVM [4] algorithm. First using if-else mining to classify the URL in three form phishing, legitimate and suspicious. Then take the suspicious URL and applied the Machine Learning algorithm to classify the Suspicious URL is phishing or legitimate. So overall we classify the all the URL in two form phishing and legitimate. For classifier we used a WEKA for SVM algorithm implementation and MATLAB for Rule Mining algorithm.

V EXPERIMENTAL RESULT AND ANALYSIS

The implementation of data mining algorithms requires the more powerful software tools. The number of available tools for data mining is grow therefor the choice of the suitable tools becomes difficult. We have used WEKA 3.6 and MATLAB for implementing proposed work.

5.1. Features

Features are extract from the URL and content of webpage. There are 40 features to detect phishing website. In this 40 features we have used 7 URL based features from that. This 7 features are listed below:

- Having_IP_Address
- URL_Length
- Port
- HTTP_token
- Age_of_domian
- Number_of_dots
- Number_of_slashes

Features	Value
Age_of_domain	{-1,1}
IP_address	{-1,1}
Number_of_dots	{-1,1}
Number_of_slashes	{-1,1}
HTTP_token	{-1,1}
Port	{-1,1}
URL_length	{-1,1}

Table 1. Parameter & its Value

5.2. Analysis of the Result

Figure 3 show the instances VS accuracy graph of Naïve Bayes algorithm. In this graph we show that the accuracy is increase and decrease with changes in instances so this algorithm is depends on instances.

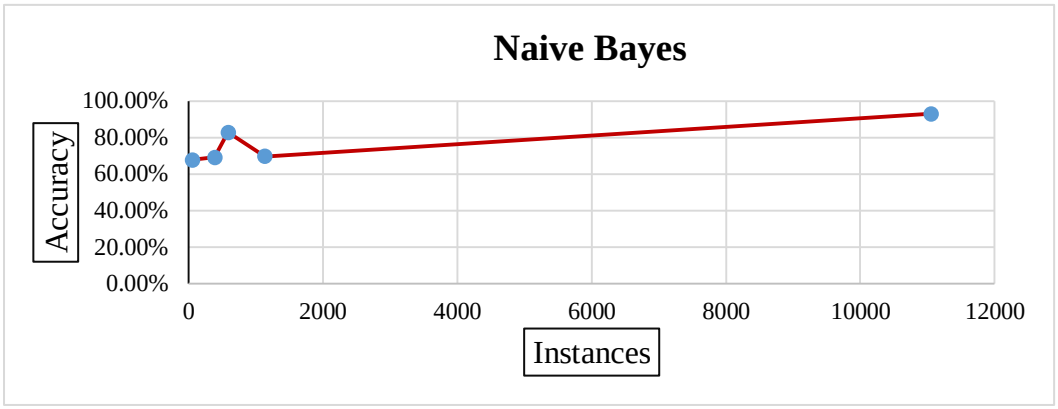

Figure 3. Naïve Bayes Accuracy VS Instances Graph

Figure 4 show the accuracy graph of the logistic region algorithm. In this graph we show that the accuracy is also increase in decrease with the change of data set instances so its algorithm is also depends on terms of instances.

Figure 4. Logistic Region Accuracy VS Instance Graph

Figure 5. show the accuracy graph of the SMO. In this graph we saw that the accuracy is change if we change the instances. It may goes up and down so we can't satisfied the result

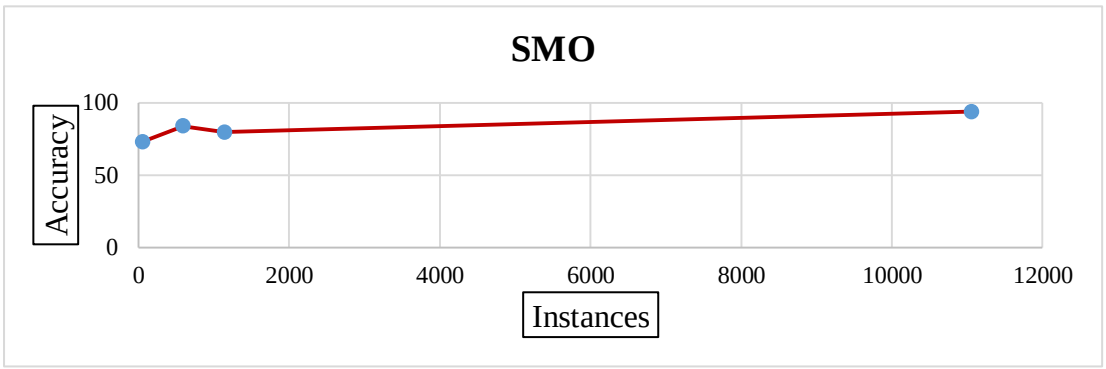

Figure 5 SMO Accuracy VS Instances Graph

Figure 6 show the accuracy graph of our proposed system. In this graph the accuracy is not depends on instances. If we increase instances accuracy also increase, so it's give a better result than other three algorithm.

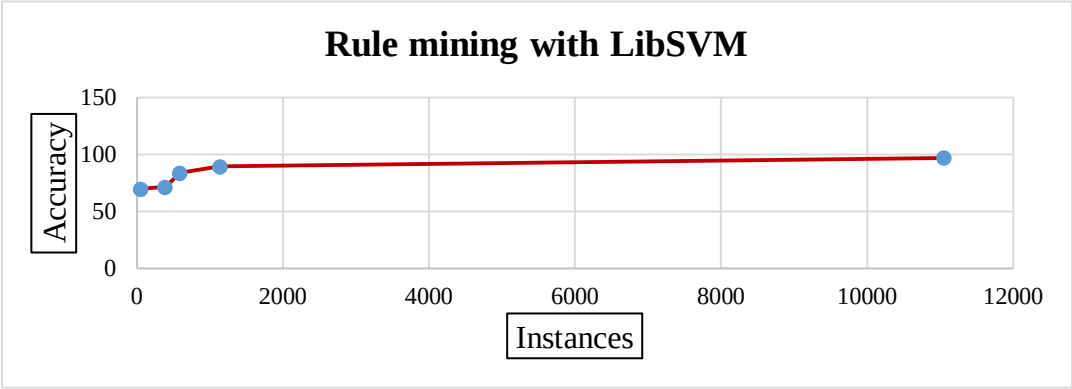

Figure 6 Rule Mining with LibSVM Accuracy VS Instances Graph

Figure 7 show the comparison graph of all the algorithm. In this result we take a same attributes and same instances, so we can say that our proposed is better than other three algorithm.

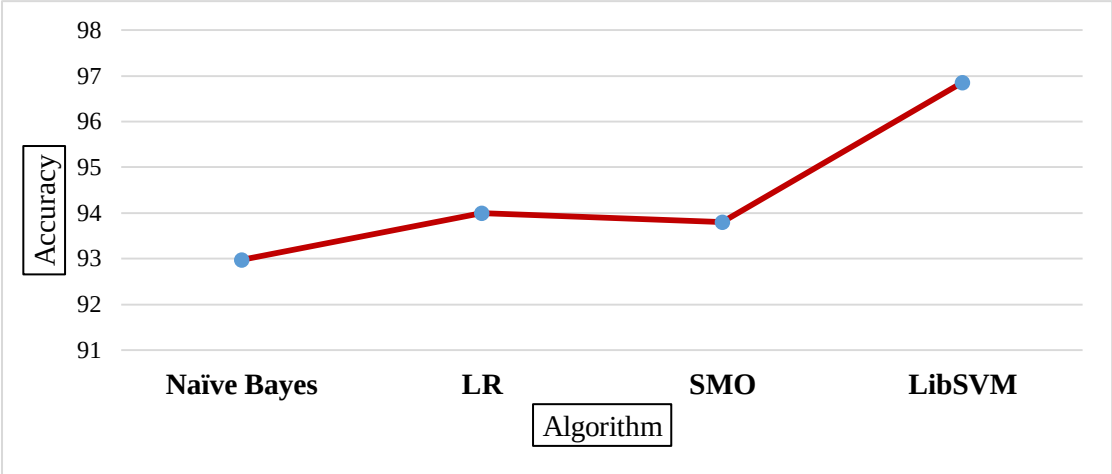

Figure 7. Comparison Graph

VI CONCLUSION

In This research present a introduction of phishing detection we analyzed the URL features using the if-else rules it is hybrid with SVM technique to solve the suspicious URL problem. Analyzed features are more sensible to phishing detection URL. Finally we conclude that Hybrid algorithm is better than other classifier.

VII REFERENCES

[1] B. B. Gupta, Aakanksha Tewari ,Ankit Kumar Jain Dharma P. Agrawal, " Fighting against phishing attacks: state of the art and future challenges" Neural Computing and Applications (2016)springer ,pp. 1–26, 2016.

[2] The Phishing Guide Understanding & Preventing Phishing Attacks By: Gunter Ollmann, Director of Security Strategy, IBM Internet Security Systems, 2007.

[3] Jeeva, Rajsingh , "Intelligent phishing url detection using association rule mining" Human-Centric Computing Information Sciences (2016) springer, pp. 1-19 , 2016.

[4] Huajun Huang; Liang Qian; Yaojun Wang "A SVM based technique to detect Phishing URLs ", Information Technology Journal;2012, Vol. 11(7), pp.921-925.

[5] Ramesh Gowtham, k. Sampath Sree Kumar, Ilango Krishnamurthi, " An efficacious method for detecting phishing webpage through Target Domain Identification " Decision Support Systems (2014) Elsevier , vol.61, pp.12–22, 2014.

[6] Dhamija R, Tygar JD, "Hearst MA (2006) Why phishing works,'' in proceedings of the 2006 conference on human factors in computing systems (CHI). ACM, Montre´al, Que´bec, Canada, pp 581–590.

[7] Anti-Phishing Working Group (APWG) (2014) Phishing activity trends report—first quarter 2014. http://antiphishing.org/reports/apwgtrendsreportq12014.pdf. Accessed Sept 2014.

[8] Anti-Phishing Working Group (APWG) (2014) Phishing activity trends report—fourth quarter 2013. http://antiphishing.org/reports/apwgtrendsreportq42013.pdf. Accessed Sept 2014.

[9] Anti-Phishing Working Group (APWG) (2014) Phishing activity trends report—second quarter 2013. http://antiphishing.org/reports/apwgtrendsreportq22013.pdf. Accessed Sept 2014.

[10] Niju Shrestha, rajan kumar kharel, Jason britt, ragib hasan, "High Performance classification of phishin URLs using a multimodel Approach with MapReduce", 2015 IEEE world congress on date of conference, pp. 206-212

[11] Mahmovd Khonji, Andrew Jones, Youssef Iragi, "A novel Phishing Classification based on URL Features", 2011 IEEE GCC Conference and exihibition(GCC),Dubai,United Arab Emirates, pp.19-22.

[12] Abdelhamid N, Ayesh A, Thabtah F (2014) "Phishing detection based associative classification data mining" Science- Direct , pp.5948–5959.

[13] Agrawal R, Imielinski T, Swami A (1993) "Mining association rules between sets of items in large databases" ACMSIGMOD, pp.207–216.

[14] Aburrous M, Hossain MA, Dahal K, Thabtah F (2010) "Predicting phishing websites using classification mining techniques with experimental case studies" Seventh international conference on information technology. IEEE Conference, Las Vegas, Nevada, USA, 2010, pp 176–181.

[15] Husna H, Phithakkitnukoon S, Palla S, Dantu R (2008) "Behavior analysis of spam botnets". In: Communication systems software and middleware and workshops, 2008. COMSWARE 2008. 3rd International Conference, Bangalore, India, 2008, pp246