Machine Learning based Phishing Website Detection System

Pradeep Tiwari
Master of Technology Scholar
Department of CSE
SITM, Lucknow

Ravendra Ratan Singh
Assistant Professor
Department of CSE
SITM, Lucknow

***Abstract*— Over the years there have been many attacks of Phishing and many people have lost huge sums of money by becoming a victim of phishing attack. In a phishing attack emails are sent to user claiming to be a legitimate organization, where in the email asks user to enter information like name, telephone, bank account number important passwords etc. such emails direct the user to a website where in user enters these personal information. These websites also known as phishing website now steal the entered user information and carries out illegal transactions thus causing harm to the user.**

Phishing website and their mails are sent to millions of users daily and thus are still a big concern for cyber security. Social engineering has come up with many educational and training programs to make users be aware of phishing website and avoid users to become victim of such attacks.

Usually a phishing website can be easily identified by its URL, its email links or HTML code. Thus many automatic phishing classifier are been built to classify whether the given mail or website is a phishing website or not. Data mining techniques, Machine algorithms techniques and programming can help in developing a system capable enough to classify whether a website is a phishing website or not. In this research work I use the dataset of phishing website of UCI machine learning dataset and data mining concepts to understand the pattern of phishing website. I select some classifiers compare their results over the given dataset and select among them the best classifier to make a machine learning base phishing website detection system. I make use of R Script interfaced with WEKA 3.2 to help in detecting phishing website.

Keywords*— Phishing; machine learning; data minin;, R Scrip; WEKA.

I. INTRODUCTION

Phishing website and their mails are sent to millions of users daily and thus are still a big concern for cyber security. Social engineering has come up with many educational and training programs to make users be aware of phishing website and avoid users to become victim of such attacks.

The process of data mining is used to analyze and help in developing a phishing website detection system.

II. DATA SELECTION

The dataset is downloaded from UCI machine learning dataset. The dataset contains 31 columns where in 30 are features on the basis of which they are classified as Phishing website and 1 as target. The dataset has 2456 observations.

The attributes of the dataset with their column names is given in table 1 below. All the attributes are represented with binary values which mean that the attribute is present of absent. Some attributes have 3 values which represent its strength ranging from low, medium and high.

TABLE I. ATTRIBUTES AND COLUMN NAMES OF PHISHING WEBSITE DATASET

Attribute	Values	Column Name
Having IP Address	{ 1,0 }	has_ip
Having long url	{ 1,0,-1 }	long_url
Uses ShortningService	{ 0,1 }	short_service
Having '@' Symbol	{ 0,1 }	has_at
Double slash redirecting	{ 0,1 }	double_slash_redirect
Having Prefix Suffix	{ -1,0,1 }	pref_suf
Having Sub Domain	{ -1,0,1 }	has_sub_domain
SSLfinal State	{ -1,1,0 }	ssl_state
Domain registeration length	{ 0,1,-1 }	long_domain
Favicon	{ 0,1 }	Favicon
Is standard Port	{ 0,1 }	Port
Uses HTTPS token	{ 0,1 }	https_token
Request_URL	{ 1,-1 }	req_url
Abnormal URL anchor	{ -1,0,1 }	url_of_anchor
Links_in_tags	{ 1,-1,0 }	tag_links
SFH	{ -1,1 }	SFH
Submitting to email	{ 1,0 }	submit_to_email
Abnormal URL	{ 1,0 }	abnormal_url
Redirect	{ 0,1 }	Redirect
on mouseover	{ 0,1 }	Mouseover
Right Click	{ 0,1 }	right_click
popUp Window	{ 0,1 }	Popup
Iframe	{ 0,1 }	Iframe
Age of domain	{ -1,0,1 }	domain_age
DNS Record	{ 1,0 }	dns_record
Web traffic	{ -1,0,1 }	Traffic
Page Rank	{ -1,0,1 }	page_rank
Google Index	{ 0,1 }	google_index
Links pointing to page	{ 1,0,-1 }	links_to_page
Statistical report	{ 1,0 }	stats_report
Result	{ 1,-1 }	Target

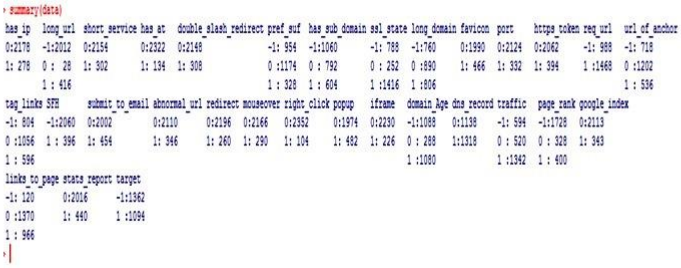
In this prediction system I have chosen all the dataset to have good training and testing data. I have divided the training set as 75% of the given observation. Thus training data set has 1843 observations and testing dataset has 613 observations.

All the attributes of the dataset are to be considered for an accurate Phishing website prediction system. Thus all 31 attributes are chosen in the dataset.

III. DATA PREPROCESSING

The .arff file of the dataset is converted and data is stored in .csv format. Thus the dataset is Phishing.csv file. But the .csv file only has observation patterns. It does not has any column names. But there shall be problems in using classifier like SVM or any if the dataset has no column names as these algorithms shall create pseudo names like V1, V2, V3 etc. Hence I rename each column and arranged them in the Phishing .csv file.

Dataset with new column names is now ready for evaluation



```
> summary(data)
has_ip  long_url  short_service has_at  double_slash_redirect pref_suf  has_sub_domain ssl_state long_domain favicon  port    https_token req_url  url_of_anchor
0:2178  -1:2012   0:2154        0:2322  0:2148                -1: 954   -1:1060        -1: 788   -1:760      0:1990   0:2124  0:2062      -1: 988  -1: 718
1: 278  0 :  28   1: 302        1: 134  1: 308                0 :1174   0 : 792        0 : 252   0 :890      1: 466   1: 332  1: 394      1 :1468  0 :1202
        1 : 416                                               1 : 328   1 : 604        1 :1416   1 :806                                           1 : 536
tag_links SFH     submit_to_email abnormal_url redirect mouseover right_click popup   iframe  domain_Age dns_record traffic  page_rank google_index
-1: 804   -1:2060 0:2002          0:2110       0:2196   0:2166    0:2352      0:1974  0:2230  -1:1088    0:1138     -1: 594  -1:1728   0:2113
0 :1056   1 : 396 1: 454          1: 346       1: 260   1: 290    1: 104      1: 482  1: 226  0 : 288    1:1318     0 : 520  0 : 328   1: 343
1 : 596                                                                                       1 :1080               1 :1342  1 : 400
links_to_page stats_report target
-1: 120       0:2016       -1:1362
0 :1370       1: 440       1 :1094
1 : 966
>
```

Fig. 1. Summary of loaded dataset of Phishing.csv

IV. DATA TRANSFORMATION

Data transformation includes normalization and is used when comparison is done of observations of different sizes. For distance based classification it is necessary to normalize each feature value. Thus in the given dataset the accepted range is [-1,1] or [-1,0, 1] but if a feature has a range [-100,100] it shall give huge differences among two vectors. Thus normalization of the data is carried out. In the given dataset all the data was normalized earlier and hence there was no reason to carry out normalization.

The last column is named as target which is numeric; it is to be transformed to factor values. If one uses numeric values SVM classifier or any other could assume it to be a regression thus the dataset is transformed to factor values to store the classification output.

V. DATA MINING

To perform data mining over the dataset the dataset is split into training and testing dataset. The split ratio is 70-30. Where in 70% accounts to training set. 10 folds cross validation is performed over the dataset for each classifier.

Now the training set is used to train the classifier. The classifiers chosen are:

1. Naïve Bayes classifier
2. J48 classifier
3. SVM radial kernel based classifier
4. Random forest based classifier
5. Tree bag based classifier
6. IBK lazy classifier

VI. PATTERN EVALUATION

1) *Naïve bayes classifier output:*

Time taken to build model: 0.02 seconds
Correctly Classified Instances 2281 92.8746 %
Incorrectly Classified Instances 175 7.1254 %

2) *J 48 Classifier:*

Accuracy: 0.9511
95% CI: (0.9309, 0.9667)
No Information Rate: 0.5546

3) *SVM radial kernel based Classifier:*

Accuracy: 0.9657
95% CI: (0.9481, 0.9787)

4) *Random forest based classifier:*

Accuracy: 0.963
95% CI: (0.952, 0.9812)
No Information Rate: 0.5546

5) *Tree bag based classifier:*

Correctly Classified Instances 2305 93.8518 %
Incorrectly Classified Instances 151 6.1482 %

6) *IBK lazy classifier:*

IB1 instance-based classifier using 1 nearest neighbor(s) for classification
Time taken to build model: 0 seconds
Correctly Classified Instances 2294 93.4039 %
Incorrectly Classified Instances 162 6.5961 %

TABLE II. FINAL SUMMARY OF ALL THE CLASSIFIER

Name of the classifier	Accuracy
Naïve Bayes	92.8746 %
J48	95.11 %
SVM	96.57 %
Random forest	96.3 %
Tree bag	93.85 %
IBK lazy classifier	93.4039 %

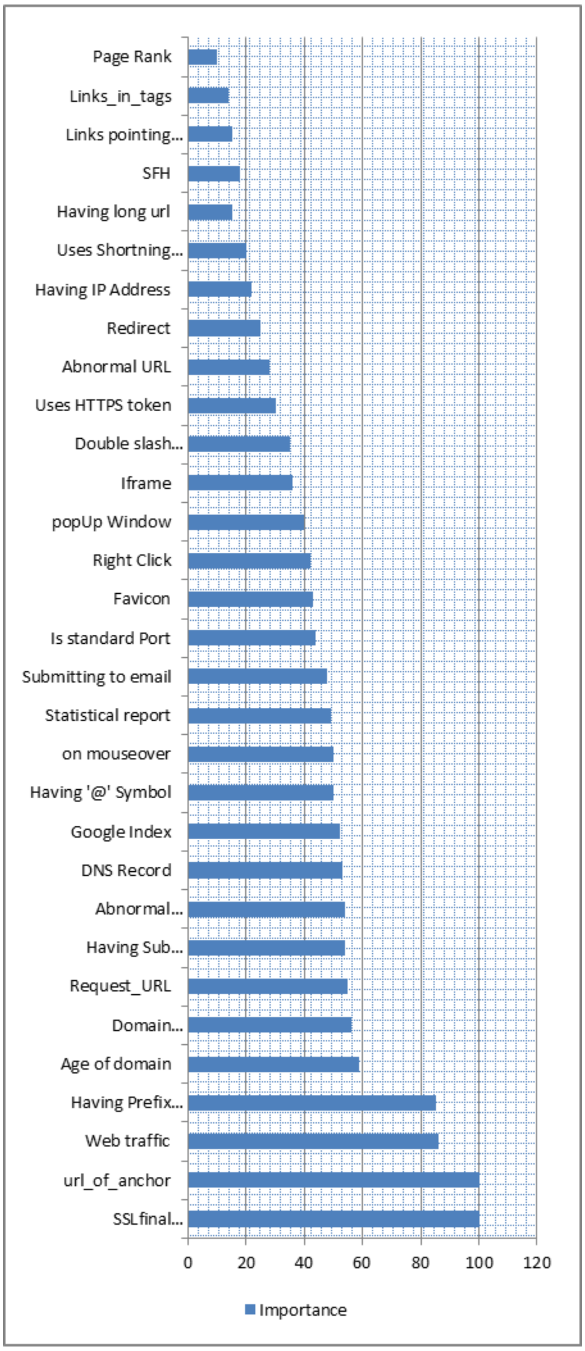


Fig. 2. Summary of results

Based on this it can be concluded that SVM radial kernel classifier is best among others for Phishing website detection and prediction.

VII. CONCLUSION

The aim of this research work is to predict whether the given URL is a phishing website or not. This work collects the dataset of UCI machine learning dataset and creates a R script and uses interface of WEKA to evaluate various types of classifier over the given dataset. The aim is to find out the best available classifier. It turns out in the given exploration that SVM based classifiers are the best classifier with good accuracy of 96.57% for the given dataset of phishing website. Now as a future work I shall extend the SVM based classifier and make an integrated R function for predicting the URL as phishing or not.

REFERENCES

[1] Anti-Phishing Working Group. Phishing Activity Trends.Report,http://antiphishing.org/apwg_report_final.pdf. 2007.

[2] B. Adida, S. Hohenberger and R. Rivest, “Lightweight Encryption for Email,” USENIX Steps to Reducing Unwanted Traffic on the Internet Workshop (SRUTI), 2005.

[3] R. Dhamija and J.D. Tygar, “The Battle against Phishing: Dynamic Security Skins,” Proc. Symp. Usable Privacy and Security, 2005.

[4] FDIC., “Putting an End to Account-Hijacking Theft,” http://wwfdic.gov/consumers/id/identity_theft.pdf, 2004.

[5] A. Y. Fu, L. Wenyin and X. Deng, “ Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover’s Distance (EMD) ,” IEEE transactions on dependable and secure computing, vol. 3, no. 4, 2006.

[6] A. Herzberg and A. Gbara, “Protecting Naive Web Users,” Draft of July 18, 2004.

[7] L. James, “Phishing Exposed,” Tech Target Article by: Sunbelt software, searchexchange.com, 2006.