Feature Extraction and Classification Phishing Websites Based on URL

Mustafa AYDIN, Nazife BAYKAL
CyDeS, Cyber Defense and Security Laboratory of METU-COMODO
Informatics Institute, Middle East Technical University (METU)
Ankara, Turkey
maydin@bddk.org.tr, baykal@metu.edu.tr

***Abstract*—In this study we extracted websites' URL features and analyzed subset based feature selection methods and classification algorithms for phishing websites detection.**

***Index Terms*—classification; cyber security; data mining; feature extraction; phishing detection**

I. Introduction

Phishing is a malicious form of online theft that aims at stealing users' personal information, such as online banking passwords, credit card numbers and other financial data. In the last decade, many users suffered monetary losses as a result of the increasing number of phishing attacks. The motivation of our study is to propose a safer framework for detecting phishing websites with high accuracy in less time.

The detection of phishing can be achieved by either increasing user awareness or using software based detection. Although there are several software detection techniques that address the problem of phishing detection, phishing has become more and more complicated and sophisticated, and can bypass the filter set by anti-phishing techniques. In this study, we extracted more URL features and analyzed subset based feature selection methods which have not been used previously for the purpose of phishing websites detection based on URL.

II. Description of the Study

To successfully identify a wide variety of phishing pages, we extracted and analyzed a number of features related to these pages. We created our feature set relying on our analysis and various existing literatures on phishing attack detection. The objective behind a feature-based approach is to make the technique of phishing attack detection as unsophisticated as possible. One of the main goals of our approach is to make the framework flexible and simple to extend the feature set by incorporating new and emerging phishing strategies as they are encountered.

A. Data Collection

Phishing attacks have different scopes. For our study dataset focuses on phishing websites which are related to most targeted brand names. We determined most targeted brand names and their real phishing URLs from PhishTank website. This website is a collaborative clearing house for data and information about phishing on the Internet which is operated by OpenDNS platform.

After having determined most targeted brand names, we used Google to obtain legitimate URLs related to these brand names. First of all we wrote the brand names on Google search engine and obtained the first link from search results as a legitimate website's main page link of the searched brand name. After getting these main page links of most targeted brands we started to get more links by Google site search. We analyzed 8538 URLs including 3622 legitimate and 4919 confirmed phishing URLs.

B. Feature Extraction

We extracted features about the URL of the pages and composed feature matrix. We categorized features into five different analyses as shown in Fig. 1. Most of these features are the textual properties of the URL itself and others based on third parties services. To obtain textual properties we wrote codes with using C# programming language at Microsoft Visual Studio program. We did some online processing steps to obtain some features from third party service providers. We then wrote R programming language script for getting "whois record" and finally we collected some data by manual work. As a result, we obtained 133 different features about the URLs in our dataset.

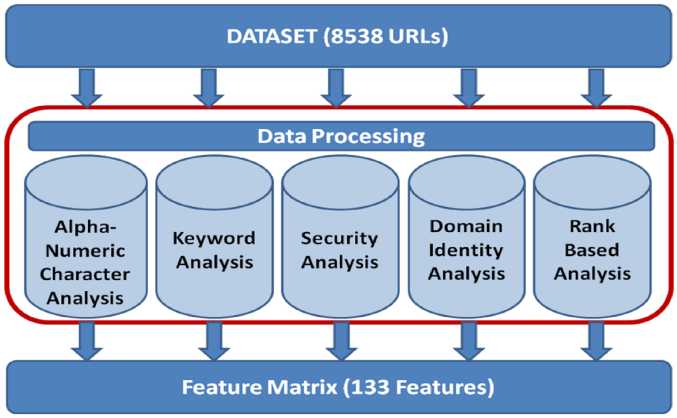

Fig. 1. Data processing categories.

978-1-4673-7876-5/15/$31.00 ©2015 IEEE

C. Feature Selection

After creation of the feature matrix, we used subset based feature selection methods to detect most prominent features. Feature selection methods are applied to optimize the dataset dimension by removing redundant and irrelevant features with respect to learning phase in the analysis. In this study, we evaluated CFS Subset based and Consistency Subset based feature selection methods with their performance contribution to classification algorithms. These two feature selection methods were separately run on the dataset, which included different attributes of 8538 URLs. After applying these methods we obtained two new feature matrices with different number of features. In this study, this process is analyzed by WEKA data mining and classification software tool.

D. Classification

For the next step after feature selection, the two new datasets were used as an input to the classification step of the analysis. In this study, we focus on two types of classification algorithms, namely Naïve Bayes and Sequential Minimal Optimization (SMO). These algorithms were run on the each datasets and used for a performance comparison. The evaluation of the classification algorithms were performed by using the Overall Accuracy, True Positive (TP) Rate, False Positive (FP) Rate and Precision. These two classification algorithms were examined by WEKA with default settings. To improve the validity, we used 10-fold cross validation to divide the training and the test data in the datasets. This classification process is developed in an offline training and testing process.

III. Results and Evaluations

The number of features was specified as 17 and 25 for CFS Subset and Consistency Subset feature selection methods respectively. The feature selection methods applied in this study produced different prediction output values depending on the classification algorithm used. For instance, the Naïve Bayes and SMO algorithms revealed their best accuracy results when they were used with the CFS Subset and Consistency Subset feature selection methods respectively.

Naïve Bayes algorithm showed its high performance with the 88.17% accuracy. In addition the SMO algorithm revealed the best compatibility with the 95.39% accuracy. This is the highest overall accuracy value obtained in the analysis. The results of the analysis are presented in Table 1 below.

The Consistency Subset method exhibited the weakest performance in Naïve Bayes algorithm with the lowest accuracy. On the contrary, this method exhibits its best performance in SMO. The SMO algorithm showed better performance in both two feature selection methods when it is compared to the Naïve Bayes algorithm. The graphical representation of the results is given in Fig. 2.

The prediction results obtained by the evaluation of the classification algorithms revealed that the SMO algorithm might be preferred for the phishing detection based on URL properties. While the Naïve Bayes algorithm did not perform well.

TABLE I. The Results of the Analysis

Feature Selection Methods	Classification Algorithm	Overall Accuracy	TP Rate	FP Rate	Precision
CFS Subset (17)	Naive Bayes	88.17%	0,882	0,093	0,900
	SMO	94.67%	0,947	0,059	0,947
Consistency Subset (25)	Naive Bayes	83.69%	0,837	0,127	0,872
	SMO	95.39%	0,954	0,046	0,954

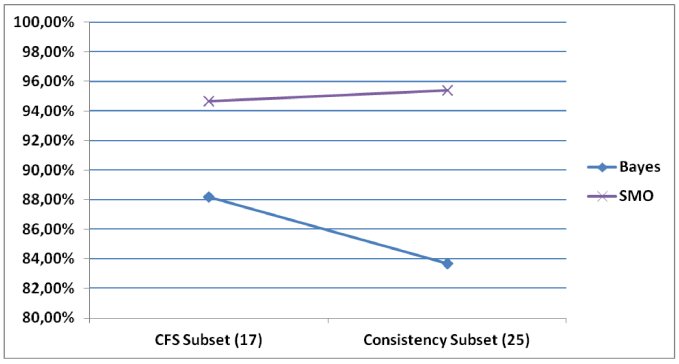

Fig. 2. Overall accuracy values of classification algorithms.

IV. Conclusion

We expect that the results of the current study might provide the basis for our future research in phishing detection. As a future work, we will work on different feature selection methods. Also, we will test and compare other classification algorithms to maximize the overall accuracy. Finally, we will test the verification process with random datasets to ensure the success of suggested algorithms.

References

[1] A. Le, A. Markopoulou, and M. Faloutsos, "PhishDef: URL names say it all," IEEE INFOCOM, 2011

[2] M. Khonji, A. Jones, and Y. Iraqi, "A novel phishing classification based on URL features," IEEE GCC Conf. and Exhibition, 2011

[3] T. Balamuralikrishna, N. Raghavendrasai and M. Satya Sukumar, "Mitigating online fraud by ant phishing model with URL and image based webpage matching," International Journal of Scientific and Engineering Research (IJSER), March, 2012

[4] A. Abunadi, O. Akanbi, and A. Zainal, "Feature extraction process: A phishing detection approach," International Conf. on Intelligent Systems Design and Applications (ISDA), 2013

[5] J. James, L. Sandhya, and C. Thomas, "Detection of phishing URLs using machine learning techniques," International Conf. on Control Communication and Computing (ICCC), 2013

[6] L. A. T. Nguyen, B. L. To, H. K. Nguyen and M. H. Nguyen, "A novel approach for phishing detection using URL-based heuristic," International Conf.on Computing, Management and Telecommunications (ComManTel), 2014

[7] PhishTank, http://www.phishtank.com/.