A Hybrid Model to Detect Phishing-Sites using Clustering and Bayesian Approach

Bhushan Dasharath Dhamdhere **Kaushal Sudhakar Dhonde** **Rohit Gopal Chinchwade**
Student, Department of Computer Engineering Pimpri Chinchwad College of Engineering –PCCOE Pimpri Chinchwad, Pune-411 044, India

Abstract
Phishing sites are the major attacks by which most of internet users are being fooled by the phisher. The replicas of the legitimate sites are created and users are directed to that web site by luring some offers to it. There are certain standards which are given by W3C (World Wide Web Consortium), based on these standards we are choosing some features which can easily describe the difference between legit site and phish site. We are proposing a model to determine the phishing sites to safeguard the web users from phisher. The features of URL along with the features of Web Page in HTML tags are considered to determine the attack. Here Clustering of Database is done through K-Means Clustering and Naive Bayes Classifier prediction technique is applied to determine the probability of the web site as Valid Phish or Invalid Phish. K-Means Clustering is applied on initial URL features and Validity is checked if still we are not able to determine the Validity of Web Site then Naive Bayes Classifier is applied onto URL as well as HTML tag features of Site and probability is evaluated based on training model.

Index Terms
Anti Phishing Technique, Bayesian Approach, Data Mining, Database Clustering, Hybrid Model, and Phishing Attack.

1. Introduction

Phisher is the community of hackers which creates the replicas of the legitimate web sites to retrieve user's personal information such as passwords, credit card number, and financial transaction information [1]. As per the survey done by RSA Fraud Surveyor, the Phishing attacks have been raised by 2% since the last December 2012 to January 2013 [2].

The W3C has set some standards that are followed by most of the legit sites but a phisher may not care to follow these standards as this site is intended to catch many fish in very small amount of time and bait [6]. There are certain characteristics of the URL's and source code of the Phishing site based on which we can guess the site is fake or not [3].

To detect and prevent the attacks from such phishing sites various preventive strategies are employed by anti-phishing service providers like Google Toolbar, an Anti-Virus service provider. These are the most common in the anti-phishing service providers [3]. These service providers are creating and maintaining the databases of blacklisted sites. Some of the anti-phishing organizations are available like PHISHTANK.COM who maintains the blacklist of the reported phishing sites and their current status if they are still online or not.

The phisher are creating sites at such a rate that there always will be some period in what the site is not reported as phish, in that case these techniques of maintaining online blacklist repositories fails. The major drawback or setback we have seen in this method is like the normal user will not always be taking caution about the phishing site, he may get tricked by overall look of site like legitimate site and it may happen like the site is not yet verified by the service providers and hence is not blocked.

Our aim is to overcome such loop-holes in the anti-phishing systems. We have proposed an efficient method to detect the phishing sites. Our model determines the phishing site and legit sites on the basis of URL features [3] and HTML features of the site with the combination of K-Means Clustering and Naive Bayes Classifier [4]. The K-Means Clustering is applied on the URL features of the web site and the feature set is plotted in one of the three clusters of database. If the feature set is nearest to Valid Phish then site is declared as Phishing, if site is nearest to Invalid Phish then it is a legit site but if the feature set not nearer to the Valid cluster or Invalid cluster it means it is nearer to cluster in between them which is suspicious sign cluster.

If the site is in the suspicious cluster then there is need of more feature extraction where we will extract HTML features by using DOM representation [5] of the HTML and features of different tags are observed. A Naive Bayes Classifier is employed if K-Means clustering is not that much useful, considering both URL and HTML features and the training datasets provided to predict the legit site or phish site.

2. Our Approach

This section describes architecture and our approach towards the design of the system. The system architecture is given below:

Manuscript received January 5, 2015
Manuscript revised January 20, 2015

A. System Architecture

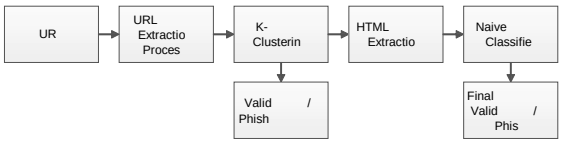

Fig. 1. System Architecture

The model we have proposed has four major parts or modules and using the pipes and filters architecture as the system is using output of process1 as input of process2.

B. Procedure

Following are the steps that are followed during the execution of the system:

Step 1: Given the web site *X*.
Step 2: Extract the URL features from *X*.
Step 3: Apply K-Means Clustering on dataset of *X* and predict the cluster in which the *X* is nearer to centroid (-1, 0, +1).
// -1: Legit Site, 0: Suspicious, +1: Phishing Site
Step 4: If output is -1 or +1, predict the result.
If output is 0 then go to step 5.
Step 5: Download the source code of webpage and extract the HTML tag features and enter into *X*.
Step 6: Classify *X* using Naive Bayes Classifier and predict the output -1 or +1.

3. URL Features and K-Means Clustering

C. URL Feature Set

The main working of the model depends upon what features are to be used in the dataset to detect the phishing attack. After studying W3C standards we have chosen following four features which can effectively determine the phishing attacks:

1. **IP Address in URL:** The domain needs to be registered in order to obtain a specific URL address for the web site but the phishing sites do last only for few days hence the phisher may not register the web site. Legit sites have their domain registered and have the URL address. This will help us to determine the phish site.
2. **Dots in URL:** The dot in the URL represents the existence of the sub-domain in the URL. Some phisher may use the sub-domain to look the site address as the legit site hence causing the user to mislead to phish site. More the dots in URL more the sub-domain existing in URL to hide the web URL and look alike the legit site.
3. **Suspicious Characters:** The Phisher will use some special characters other than alpha-numeric character to trick the user. Special characters used maybe '@','&','-', and '_' in the web URL to create the pattern of the legit URL which the user easily click on.
4. **Slashes in URL:** The slashes in the URL shows existence of sub-folders in it. The sub-folders are added to hide the information in the web-pages.

D. K-Means Clustering Technique

After studying 1000 records in the Phishtank Online repository (www.phishtank.com) [8], we have determined a set of pattern as follow:

1. The phish URL have the higher values of above features.
2. The legit URL shows the lower values of above features.

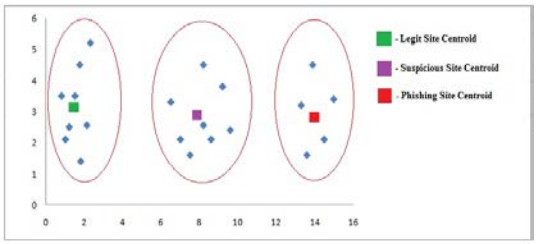

Fig. 2. Plotting of the feature set with respect to 3 clusters

Thus we can create the database with application of the clustering. The database can be divided into three clusters on the characteristics:

1. **Valid Phish:** This cluster contains the feature sets having higher values of features. These features show the properties of a phishing site.
2. **Suspicious Phish:** This cluster contains the feature sets that are some feature shows site is legit and some feature shows the site is phish site.
3. **Invalid Phish:** This cluster contains the feature sets whose values are relatively very less. These features indicate properties of the legit web sites.

Now that we know how the database is to be clustered in horizontal partitions, we employ K-Means Clustering algorithm. K-Means Clustering denotes the clustering of the database of n feature sets using k partitions, for our clustering k=3. For the initial clustering there is need of providing initial values for clusters [7].

For the given set of feature sets (fs_1, fs_2, fs_3 … fs_n), where each feature set is of 4 dimensional vectors, K-Means clustering with the help of following formula:

$$\arg\min_{\mathbf{s}} \sum_{i=1}^{3} \sum_{xj \in Si} \|x_j - \mu_i\|^2$$

We measure the distance between each centroid and feature set after every iteration and updating of centroid values.

4. HTML Features and NB Classifier

A. HTML Tags Features

Sometimes only the URL features are not enough predict the site is phishing or legit. Once the site is classified as suspicious phish by K-Means Clustering, more dynamic or HTML tag features are required to be extracted from the source code of the web page. For this feature extraction we will be using the HTML DOMTree Parser which will enable us to view the HTML code and extract various details very easily.

1.NULL Anchors: These are the Anchor tags in the web page which are not pointing to anything. When clicked on such links nothing happens or the links are redirected to the same page. After copying the source code of legit site the phisher may delete most of the links or replace with the link of the same page. More the NULL anchors are in page more the page is likely to be a phishing site.

2.Foreign Anchors: These are the Anchor tags in web page which are linking to the domain which is not a domain or sub-domain of the current web site. Web sites can have some foreign links on it but too many foreign links will obviously increase the suspicion about that site. Phisher may copy the source code of legit web site to his own web page and then modify the web page in order to achieve the higher similarity with site he is trying to attack, the phisher cannot always modify each and every anchor tag which are pointing to legit sites and hence increasing the suspiciousness of that web page.

3.SSL Certificate: It is nothing but Secure Socket Layer Certificate provided by some of the trusted authorities on W3C, the SSL Certificate covers the identity of the owner of the web page along with how it is encrypted and other information. Every legit page now has the SSL certificate 2.0 or 3.0 versions.

The SSL Certificate has validity of very short period and needs to be updated over period of time. Most of the browsers allow the web page access when SSL Certificate is present.

As the SSL Certificate is only provided to legit and verified owners of web pages, phisher has very less chances of obtaining SSL Certificate.

B. Naive Bayes Classifier

Bayesian classifiers find the distribution of attribute values for each class in the training data. To find the probability p(Cj|d) of the instance d being in class Cj, Bayesian classifiers use Bayes theorem which says:[ADB Book Korth]

$$p(C_j|d) = \frac{p(d|\ C_j)\ p(C_j)}{p(d)}$$

In other words the formula for Naive Bayes can be given as follows:

$$V_{nb} = \arg\max_{fsi \in V} P(fs_i) \prod P(fs_i|C_j)$$

We are using the Naive Bayes Classifier which estimate p(fs|C) using m-estimates as follows[NB Example]:

$$p(fs_i|C_j) = \frac{n_c + mp}{n + m}$$

n = no. of training examples for which $fs_i = fs$.
m = arbitrary value, equivalent sample size
n_c = no. of examples for which $f_s = fs$ and C = C_j.
p = 1/ no. of values attribute can have (2)

The advantages of using the NB classifier over the decision tree classifiers is they can classify the unknown or null attribute values by omitting from probability computation. Hence the results will be more accurate than that of the decision tree classifiers as they cannot handle null or unknown attributes meaningfully.

We need to provide the strong training data set in order prediction to be close to correct, we have studied 1000 records in www.phishtank.com out of which 600 are valid phishing sites and 400 are legit site records. This training set will be useful according to our assumptions in order to predict the result using NB classifier.

5. Estimated Results

Based on the training set taken from www.phishtank.com some predictions are evaluated and proposed model is verified. Following example will illustrate how the system will predict the results by fusion of the clustering and NB classifier.

Following is the illustration for the K-Means Clustering for our model:

WB	IP	DOT	SLAS	SCH	CLUSTER
A	0	2	2	1	**1**
B	0	1	1	2	**1**
C	0	2	3	4	**1**
D	0	5	4	9	**2**
E	1	3	5	10	**2**
F	1	4	7	20	**2**
G	1	5	9	4	**2**
H	1	8	13	15	**3**
I	1	9	9	16	**3**

INTIAL CLUSTERS CENTROID				
C1	1	1	1	5
C2	1	5	5	10
C3	2	10	10	20

MODIFIED CLUSTER CENTROID				
C1	0	1.67	2	2.34
C2	0.75	4.25	6.25	10.75
C3	2	8.5	11	15.5

Following is the illustration of NB Classifier:

TRAINING DATA SET FOR NAIVE BAYES								
WB	**IP**	**DOT**	**SLAS**	**SCH**	**SSL**	**FA**	**NA**	**CLUSTER**
A	0	2	2	1	0	1	0	**1**
B	0	1	1	2	0	2	1	**1**
C	0	2	3	4	0	1	0	**1**
D	0	5	4	9	1	5	2	**1**
E	1	3	5	10	1	7	0	**3**
F	1	4	7	20	0	1	5	**1**
G	1	5	9	4	1	7	4	**3**
H	1	8	13	15	1	5	7	**3**
I	1	9	9	16	1	9	8	**3**
J	1	5	9	10	1	7	2	

CALCULATION FOR VALID (CLUSTER=3)					
	N	**NC**	**M**	**P**	**PROB**
IP=1	5	4	7	0.5	0.625
DOT=5	2	1	7	0.5	0.500
SLAS=9	2	2	7	0.5	0.611
SCH=10	1	1	7	0.5	0.563
SSL=1	5	4	7	0.5	0.625
FA=7	2	2	7	0.5	0.611
NA=2	1	0	7	0.5	0.438
					0.017950

CALCULATION FOR INVALID (CLUSTER=1)					
IP=1	**N**	**NC**	**M**	**P**	**PROB**
IP=1	5	1	7	0.5	0.375
DOT=5	2	1	7	0.5	0.500
SLAS=9	2	0	7	0.5	0.389
SCH=10	1	0	7	0.5	0.438
SSL=1	5	1	7	0.5	0.375
FA=7	2	0	7	0.5	0.389
NA=2	1	1	7	0.5	0.563
					0.002617

- Here probability of feature set belonging to Cluster 3 is more than Cluster 1 (0.017950>0.002617) hence this feature set is classified as VALID phish.

6. Conclusions

- In this paper, we evaluated two phishing detection a system mechanisms out of which one is dependent on URL features of web-sites and second is based on HTML tags and Visual Features of web-sites. We have created a system which is a trail of combination of these two systems and using base techniques given by them.
- Application of clustering on this system generates the output faster but by compromising with the accuracy of results.
- Bayesian approach generates more accurate results but it requires analyzing the training data set provided and takes a very long time of execution.
- We have used a combination of these two algorithms resulting into a mechanism which is more efficient and reliable than these two separate techniques. Our mechanism uses K-Means Clustering which is efficient to generate output at higher throughput but with lack of efficiency and this lack of efficiency is recovered with the Naive Bayes Classifier.

Acknowledgment

We express our sincere thanks to our guide Prof. Rahul Patil, for his constant encouragement and support throughout or project, especially for the useful suggestions given during the development of this model.

We also thank all the web committees for enriching us with their immense knowledge. Finally, we take this opportunity to extend our deep appreciation to our families and friends, for all that they meant to us during the crucial times of the completion of our project.

References

[1] Rachna Dhamija, J. D. Tygar, and Marti Heast, "Why Phishing Works", CHI-2006, Conference on Human Factor in Computing Systems, April 2006.

[2] RSA Online Fraud Surveyor, "The phishing kit – the same wolf, just different sheep's clothing", RSA Surveys,vol-1, February-2013.

[3] Xiaoqing GU, Hongyuan WANG, and Tongguang NI "An Efficient Approach to Detect Phishing Web" Journal of Computational Information Systems 9:14(2013), 2013, pp. 5553-5560.

[4] Haijun Zhang, Gang Liu, Tommy W. S. Chow, Senior Member, IEEE, and Wenyin Liu, Senior Member, IEEE "Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach", vol-22, IEEE Transactions October-2011 pp. 1532-1546.

[5] Angelo P. E.Rosiello, Engin Kirda, Christopher Kruegel, Fabrizio Ferrandi, and Politecnico di Milano "A Layout-Similarity-Based Approach for Detecting Phishing Pages"- unpublished

[6] WIKIPEDIA.ORG- The Online Encyclopedia, http://www.wikipedia.org/

[7] Abraham Sillberschatz, Henry Korth, and S. Sudarshan, "Database System Concepts", 5th Edition, pp. 895-903.

[8] PHISHTANK.COM- The Online Valid Phish Sites Repository, http://www.phishtank.com