Clustering and Bayesian Approach-based Model for Detection of Phishing

Amitkumar Shinde
Dr.D.Y.Patil institute Of Engg, Ambi, Pune,
Pune University

Angad Pandey
Dr.D.Y.Patil institute Of Engg, Ambi, Pune ,
Pune University

Rahul Pawar
Dr.D.Y.Patil Institute Of Engg, Ambi, Pune,
Pune University

Vinayak Gangule
Dr.D.Y.Patil institute Of Engg,Ambi, Pune ,
Pune University

ABSTRACT

Phishing is an internet attack that aims to get users sensitive information by fraud websites. Website phishing is one of the major attacks by which most of internet users are being fooled by the phisher. The best way to protect from phishing is to recognize a phish. Phishing emails usually appear to come from well-known organization and ask your personal information such as credit card number, security number, account number or passwords. What actually attacker does? The attacker creates the no of replicas of authenticate sites, and users are forced to direct to that websites by attracting them with offers. As standard mentioned in W3C (World Wide Web Consortium), I am proposing a system which can easily recognize the difference between authenticate site and phishing site. There are certain standards which are given by W3C (World Wide Web Consortium), based on these standards I am choosing some features which can easily describe the difference between legit site and phish site.

To protect you from phishing, I am proposing a model to determine the fraud sites. To determine the phishing attack, URL features and HTML features of web page are considered. Clustering algorithm such as K-Means clustering is applied on the database and prediction techniques such as Naive Bayes Classifier is applied. By applying this, probability of the web site as valid Phish or Invalid Phish. To check the validity of URL,if still we are not able decide the validity of web page then Naïve Bayes Classifier is applied . also training model is applied for the extraction of HTML tag features of site and probability.

Index Terms

Anti Phishing Technique, Bayesian Approach, Data Mining, Database Clustering, and Phishing Attack.

1. INTRODUCTION

Phisher is the community of attacker which creates the replicas of the legitimate web sites to submitting user's personal information such as passwords, credit card number, and financial transaction information to illegitimate websites[1]. Since the last December 2012 to January 2013, there is rise in phishing attacks by 2% as described in survey of RSA fraud Surveyor [2]. The W3C has set some standards, specifications and recommendations that are followed by most of the authenticate 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 *http://www.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. My aim is to overcome such loop-holes in the anti-phishing systems. I have proposed an efficient method to detect the phishing sites. My model differentiates the phishing sites and authenticates sites. Model uses the URL features [3] and HTML features. To check the validity of the site, K-Means Clustering and Naive Bayes Classifier [4] used. The K-Means Clustering is applied on the URL features of the web site and the feature set is plotted in one of the two clusters of database. If the feature set is nearest to more suspicious then site is declared as Phishing, if site is nearest to less suspicious then it is a authenticate site but if the feature set not nearer to less suspicious cluster or more suspicious cluster it means it is nearer to cluster in between them.

If the site is in the nearer to the cluster between them then there is need of more feature extraction where I 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. APPROACH

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

System Architecture

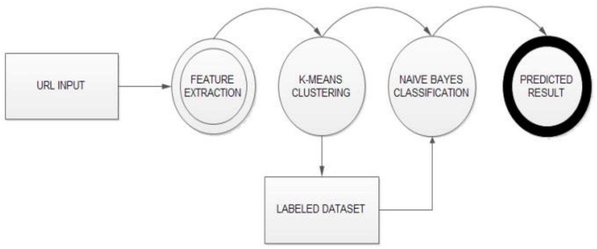

Fig. 1: System Architecture

The model I 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.

A Procedure

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

I. Given *X. is the URL from web browser*

II. Extract the URL features from *URL X such as Dots, Slashes etc.*

III. From source code extract the HTML tag features such as Null anchors, foreign anchors etc. and add this HTML features into X.

IV. Apply K-Means Clustering on dataset of *X* and predict the data point X to the cluster whose distance from the centroid is minimum. (0, 1).

// 0: Less suspicious, 1: More suspicious

V. Naive Bayes Classifier is used for classification of X and predicts the output 0 or 1.

3. URL FEATURES AND HTML FEATURES

A URL features:

For detection of phishing sites we need to consider the features of URL. I have studied the W3C standard, according to this following are the URL features:

IP Address: The domain name is one of the pieces inside the URL.It is an entire set of directions and it contains extremely detailed information.

It is also the most easily recognized part of the entire address. IP address needs to be registered. Authenticate site have their registered domain. This is an important features to recognize phishing sites.

Dots: You can create a domain name where there 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.

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.

Slashes: The slashes in the URL shows existence of sub-folders in it. The sub-folders are added to hide the information in the web

B HTML features:

Only URL features cannot predict whether site is phishing or not. when its not possible to predict phishing sites then we need to extract more features of site called as HTML features.

HTML features are extracted from source code. For extraction of HTML features we require HTML DOM-tree parser [8].following are the HTML features:

NULL tags: A NULL tag doesn't return anything. After clicking on such link,it returns the same page. 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.

Foreign tags: 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.

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.

4. K-MEANS CLUSTERING TECHNIQUE

By referring the *http://www.phishtank.com/* [8], I have determined a set of template as follow:
The phish URL have the higher values of above features. The authenticate URL shows the lower values of above features.

The database can be divided into two clusters on the characteristics:

Less suspicious: If the values of the feature are considerably then they may not be treated as the phishing site.

More Suspicious: If the values of the feature are very much larger then it has more suspicious of being phishing site

Now that we know how the database is to be clustered in horizontal partitions, I employ K-Means Clustering algorithm. K-Means Clustering denotes the clustering of the database of *n* feature sets using *k* partitions, for my clustering *k=2.* 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:

$$\arg\min \sum_{i=1}^{2} \sum_{x_j \epsilon s_i} \left\| x_i - u_j \right\|^2 \qquad (1)$$

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

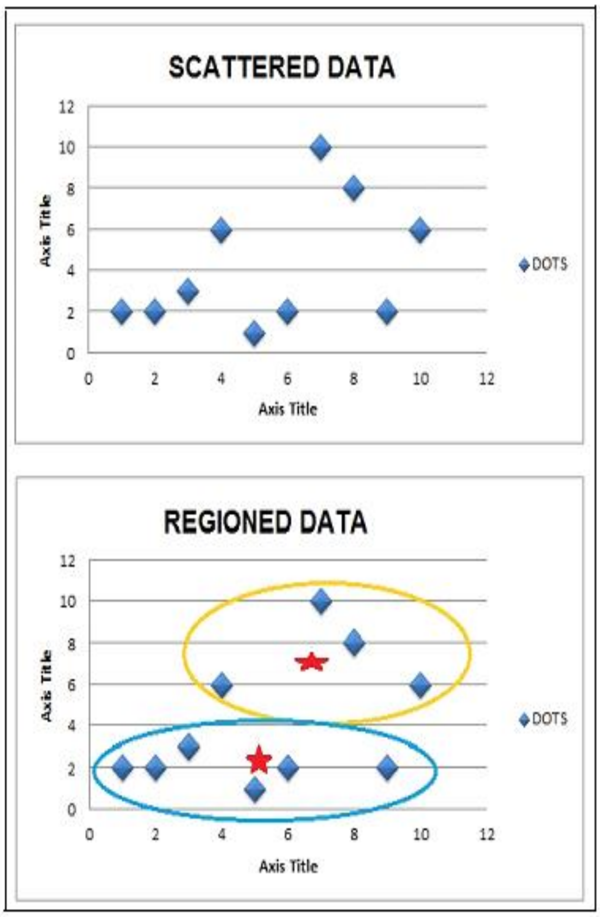

Fig. 2: Plotting of the feature set with respect to 2 clusters

5. NAIVE BAYES CLASSIFIER

Bayesian classifiers find the distribution of attribute values for each class in the training data. To find the probability $p(C_j|d)$ of the instance d being in class $C_{j,}$ Bayesian classifiers use Bayes theorem [7] which says:

$$p(C_j|d) = \frac{p(d|c_j)p(C_j)}{p(d)} \qquad (2)$$

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

$$V_{nb} = \arg\max_{fs_i \in V} P(fs_i) \, \Pi \, P(fs_i|C_j) \qquad (3)$$

I are using the Naive Bayes Classifier which estimate $p(f_s|C)$ using m-estimates as follows[9]:

$$p(fs_i|C_j) = \frac{n_c + mp}{n+m} \qquad (4)$$

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.

I need to provide the strong training data set in order prediction to be close to correct, I have studied 500 records in *www.phishtank.com* out of which 300 are valid phishing sites and 200 are legit site records. This training set will be useful according to my assumptions in order to predict the result using NB classifier.

6. LITERATURE SURVEY

Current Phishing Status:

Looking at the First fortnight report by Anti-Phishing Organization (www.antiphishing.org) and RSA Online Fraud Attacks Surveys few major points:

1. Phishing attacks has been increased by 2% since December 2012.
2. India is having 4% of global attacks by volume of attack.
3. India is being targeted 4% of global attacks by volume of brands attacked.

Taking the reference of phishing activity trends report, 2nd quarter (2014) produced by APWG (antiphishing work group) few major points are noticed:

1. Total 128,378 sites were observed as phishing sites.
2. Most targeted and more frequently industries are online payment and crypto-surrency.
3. Increase in PUPs (Potentially unwanted programs) and this leads to higher global infection
4. One of the top hosting phishing site country is United states.

Below fig.1 shows the most targeted industry with their total percentage. Few major points observed in diagram:

1. As like with the previous phishing report payment services is the most targeted industry with 39.80% of attacks.
2. 2nd targeted industry is financial industry with 20.20% of attacks.
3. 3rd targeted industry in Retail/services industry with percentage 16.53 of attacks.

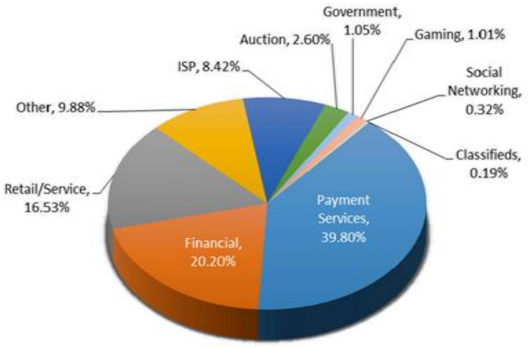

Fig.1: Most targeted industry sector

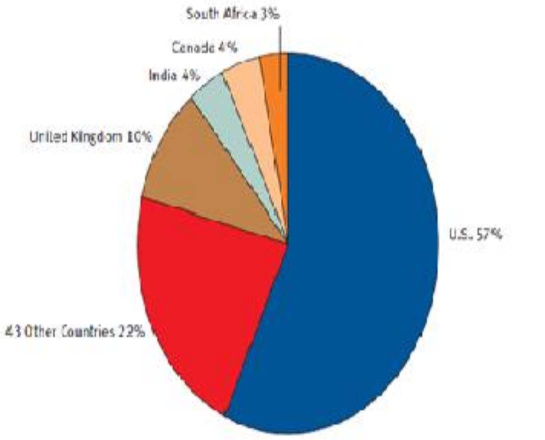

Fig.2: Major attacked countries by volume of attack

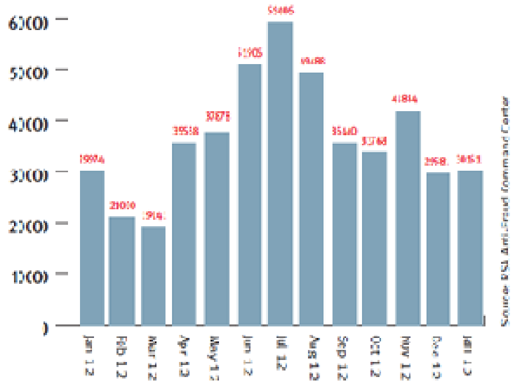

Fig.3: Total reported attacks per month for 1 year

7. CONCLUSIONS

Phishing attacks are increasingly rapidly.There is need to develop techniques for detection and prevention of phishing sites.In this survey paper , There is 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. Technique is based on system which is a trail of combination of these two mechanisms 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.

System has used a combination of these two algorithms resulting into a mechanism which is more efficient and reliable than these two separate techniques. 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.

8. 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. 900-903.

[8] PHISHTANK.COM- The Online Valid Phish Sites Repository, http://data.phishtank.com/data/online-valid.csv

[9] Eric Meisner, Naive Bayes Classifier Example, 22nd November 2003-*unpublished*