Web Phishing Detection System: Bayesian and Clustering Approach

Nilima Ramdas Narad
Department of computer engineering
Dr. D. Y. Patil College of Engineering,
Ambi, Talegaon

Sandeep U. Kadam
Department of computer engineering
Dr. D. Y. Patil College of Engineering,
Ambi, Talegaon.

ABSTRACT

Phishing is an online crime that aims to create genuine looking websites to attract users and let them releasing their sensitive information on that 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 Naive Bayes Classifier is applied . Also training model is applied for the extraction of HTML tag features of site and probability.

Keywords

Anti Phishing, Bayesian technique, Data Mining, Database Clustering, and Phishing Attack.

1. INTRODUCTION

Web site attacker creates the replicas of the authenticate web sites and forcing to submit 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]. For prevention and detection of attacks various preventive strategies are developed by most common anti-phishing service provider such as Google Toolbar, an antivirus provider [3]. What actually this service provider does? This service provider creates and maintains the database of sites which are blacklisted. There are some organizations like http*://www.phishtank.com/* which are anti-phishing organizations. These organizations keep the record of blacklisted sites or phishing sites.

There are various techniques are available for detection of phish, such as, plug-In-browser .This techniques maintains the online repositories of blacklisted sites. The phisher always creates the site at such a rate that in a particular time period that site is not reported as phish, in that case these techniques fails. As we have seen the major disadvantages of is like the normal user will not always take the precaution of phishing site. Due to the overall look of site like legitimate site and this may happen this site is not blocked by service provider.

I have proposed a system to overcome with web phishing attacks. In my proposed system I am using two algorithms one is K-Mean clustering algorithm and second is Naïve Bayes Classifier. By using this system user can differentiate between authenticate site and phishing site. First I have used K-Mean clustering algorithms on URL features. URL features includes no of dots, no of slashes, no of special characters and IS IP Address. I have created two clusters i.e. less suspicious and more suspicious. Based on these clusters user can detect which are phishing sites and which is authenticate sites. If URL features are not sufficient for determination, then I have to extract HTML features.HTML features Includes No of Foreign Anchors, No of Null anchors and IS HTTPS. I have created labeled dataset. Finally I have to calculate final probability.For all these URL and HTML I have created two clusters. I have used Naïve Bayes Classifier to calculate the final probability.i have calculated probability for both i.e. 0 and 1. After calculation of both the probabilities I have compared both these probabilities

In this paper I have explained the overall flow of system. As well I have given detail algorithm for K-Mean clustering and Naïve Bayes Classifier.

I have given experimental results for both K-Mean Clustering and Naïve Bayes classifier.

2. SYSTEM ARCHITECTURE

The system architecture of system is given below. This architecture gives my approach towards designing of this system.

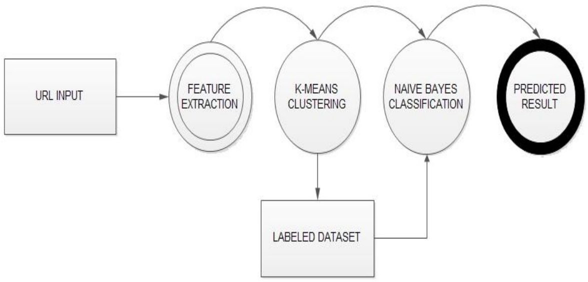

Fig. 1: System Architecture

Using architecture of pipes and filters I m proposing this model with this major modules.

A. *Procedure*

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

Step 1: Enter any URL From Web Browser To Determine The Results.

Step 2: Extract 4 URL Features Of Site-Is IP address,dots,slashes, special character

Step 3: Extract 3 html features of sites- null anchors, foreign null anchors, foreign anchors, Is https.

Step 4: Apply k-means clustering to label database on features-dots, slashes, special characters, null anchors, foreign anchors.

Step 5: Provide this complete database as the training Database to the Naive Bayes classification.

Step 6: Take all features of the site.

Step 7: Determine individual probabilities of each Feature with respect to true and false results.

Step 8: Calculate the final probabilities

Step 9: Declare the result true or false based on maximum

Determined probabilities.

3. FLOW OF K-MEANS ALGORITHM

Step 1: Take Feature *N*

Step 2: For Each Record *X* Repeat 3 To 8 until Previous Centroids! = Current Centroids.

Step 3: Determine Distance Of That Record *X* And Feature With Respect To Low And High Centroid

Step 4: If Low_Distance < High_Distance

Step 5: Label *X* As 0 (Low Suspicious)

Step 6: If Low_Distance > High_Distance

Step 7: Label *X* As 1 (High Suspicious)

Step 8: Recalculate Centroids

4. FLOW OF NB CLASSIFIER ALGORITHM

Step 1: For True Value

Step 2: Repeat 3 To 6 For All Features

Step 3: Take Record X

Step 4: Calculate *N* = Total Number Of Matches

Step 5: Calculate *Nc* = Total Number Of True Matches

Step 6: Calculate Probability Of Feature

Step 7: Calculate Final True Probability

Step 8: For False Value

Step 9: Repeat 10 To 13 For All Features

Step 10: Take Record *X*

Step 11: Calculate *N* = Total Number Of Matches

Step 12: Calculate *Nc* = Total Number Of False Matches

Step 13: Calculate Probability Of Feature

Step 14: Calculate Final False Probability

Step 15: Compare True And False Probability

Step 16: Declare Highest Probability As The Result

5. ESTIMATED RESULTS

K-Means clustering results:

Based upon the study of records from online repositories two initial clusters are created. The total no of URL features and HTML features are calculated. Two initial clusters are created based on this dataset.

Initial dataset is modified by labeling the features. Based on this new labeling dataset Cluster values are modified. Following tables shows the estimated results of the clustering algorithm.

Below table1 shows the dataset of features. Two clusters are created By applying K-Means clustering as shown in table2.

In table 3 initial dataset is labeled. For every feature separate labels are created. And by using this labeled dataset a cluster values ar e modified.

Table 4 shows the expected result of K-Means clustering algorithm.

Table 1: Dataset Before Labeling

SITE	DOTS	SLASHES	SP. CHARS	NULLA	FOREIGN A
A	2	0	1	10	11
B	2	0	3	1	2
C	3	0	2	0	0
D	6	9	8	0	0
E	1	1	0	1	9
F	2	3	3	0	1
G	10	2	2	0	0
H	8	11	3	0	3

I	2	4	6	0	0
J	6	8	1	7	1

Table 2: Initial Clusters With Respect To Records

CLUSTER	DOTS	SLASHES	SP. CHARS	NULL A	FOREIGN A
LESS (0)	1	0	0	0	0
MORE(1)	10	11	8	10	11

Table 3: Dataset After Clustering And Labelling

SITE	LABEL (DOTS)	LABEL (SLASHES)	LABEL (SP.CHARS)	LABEL (NULL A)	LABEL (FOREIGN A)
A	0	0	0	1	1
B	0	0	0	0	0
C	0	0	0	0	0
D	1	1	1	0	0
E	0	0	0	0	1
F	0	0	0	0	0
G	1	0	0	0	0
H	1	1	0	0	0
I	0	0	1	0	0
J	1	1	0	1	1

Table 4: After Clustering Records

CLUSTER	DOTS	SLASHES	SP. CHARS	NULL A	FOREIGN A
LESS (0)	2	1.5	1.85	0.25	0.825
MORE (1)	7.5	7.25	7	8.5	10

Naive Bayes Classifier results:

For NB classifier training dataset is required. Below table 4 gives the training labeled dataset (label either 0 or 1).

In below table 6, there is new sites X for which we have to calculate probability. In table 7 probabilities is calculated for 0 and in table 8 probability is calculated for 1.

Table 5: Training Model Of Naive Bayes Classifie

SITE	IS_IP	DOTS	SLASHES	SP_CHAR	N_AN	F_ANCH	IS_HTTPS	RESULT
A	0	0	0	0	1	1	1	1
B	0	0	0	0	0	0	0	0
C	1	0	0	0	0	0	0	0
D	0	1	1	1	0	0	1	1
E	0	0	0	0	0	1	0	0
F	0	0	0	0	0	0	1	0
G	0	1	0	0	0	0	1	1
H	0	1	1	0	0	0	0	1
I	0	0	0	1	0	0	1	0
J	1	1	1	0	1	0	0	1

Table 6: Unknown Site To Determine The Probability

SITE	IS_IP	DOTS	DOTS_LBL	SLASHES	SLASHES_LBL	SP. CHAR	SP_CHAR_LBL	N_ANCH	N_AN_LBL	F_ANCH	F_ANCH_LBL	IS_HTTPS	RESULT	
X	0	4	0	1	0		11	1	0	0	5	1	0	🟥

Table 8: Probability Of Result = 1

	N	NC	M	P	PROBABILITY
IS_IP	8	4	2	0.5	0.50
DOTS_LBL	6	1	2	0.5	0.25
SLASHES_LBL	7	3	2	0.5	0.44
SP_CHAR_LBL	2	1	2	0.5	0.50
N_AN_LBL	8	3	2	0.5	0.40
F_ANCH_LBL	8	4	2	0.5	0.50
IS_HTTPS	5	2	2	0.5	0.43
				FINAL PROBABILITY	0.002380952

Table 7: Probability Of Result = 0

	N	NC	M	P	PROBABILITY
IS_IP	8	4	2	0.5	0.50
DOTS_LBL	6	5	2	0.5	0.75
SLASHES_LBL	7	4	2	0.5	0.56

SP_CHAR_LBL	2	1	2	0.5	0.50
N_AN_LBL	8	5	2	0.5	0.60
F_ANCH_LBL	8	4	2	0.5	0.50
IS_HTTPS	5	3	2	0.5	0.57
			FINAL PROBABILITY		0.017857143

Probability for 0 is greater than probability for 1 (0.017857143> 0.002380952). Hence given site X is a phishing site.

6. CONCLUSION

I have proposed Web phishing detection system In this paper, I have explained the overall process of my proposed system with experimental results. I have used K-Mean Clustering and Naïve Bayes classifier. Feature extraction is applied on both URL features and HTML features. To calculate final probability I have used naive bayes classifier. Architecture of my system is given/

K-Means clustering algorithm provides faster output with accurate results. By using Bayesian classification also we can create more accurate results. Experimental results for both K-Mean and Naïve Bayes classifier are given.

7. 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

[10] A hybrid model for detection of phishing sites using

[11] clustering and Bayesian approach,6th April 2014.