Detection of Website Phishing Using MCAC Technique Implementation

Prof.T.Bhaskar*, Aher Sonali[1], Bawake Nikita [2], Gosavi Akshada[3] ,Gunjal Swati[4]

*Asst .Prof(Computer Engineering)

[1] Student, Computer Engineering, Sanjivani Collage of Engineering,Kopargaon, Maharashtra, India
[2] Student, Computer Engineering, Sanjivani Collage of Engineering,Kopargaon, Maharashtra, India
[3] Student, Computer Engineering, Sanjivani Collage of Engineering,Kopargaon, Maharashtra, India
[4] Student, Computer Engineering, Sanjivani Collage of Engineering,Kopargaon, Maharashtra, India

ABSTRACT

One of the important security challenges of Website phishing for the online people because of the large numbers of online transactions performed on a daily basis. Phishing is a kind of malicious attack where cyber criminals create a phishy website meant to look like a popular online resource (an online games, online banking services or social network) and use different social engineering methods to attempt to incentive users to the website .Examples to minimize the threat of this problems are White List, Black List and the utilization of search methods. The Black List one of the popular and widely used technique into browsers, but they are not much more effective and unsure. Associative Classification (AC) is one of the techniques based on data mining used to find phishing websites with high purity. By using If-Then rules AC extracts classifiers with a large degree of guessing accuracy.AC method developed Multi-label Classifier based Associative Classification (MCAC) for the problem of website phishing and to find features that differentiate phishing websites from legitimate ones. In this paper, MCAC identify phishing websites with higher purity and MCAC originate new hidden rules that other algorithms are not able to find and this has improved its classifiers predictive performance.

Keyword: - *Classification, Data mining, websites, Phishing, Internet security.*

1. INTRODUCTION

Phishing is a method to mimicking an official websites or genuine websites of any organization such as banks, institutes social networking websites, etc. Mainly phishing is tackled to fraud private documentation of users such as username, passwords, PIN number or any credit card details etc.

Phishing is tackled by trained hackers or attackers. Phishing is mostly attempted by fake e-mails. This kind of fake e-mails may include duplicate link of websites that is originated by attacker. By clicking these kinds of links, it is redirected on malicious website and it is easily to cheat your personal information. Phishing Detection is a method to identify a phishing activity. There are different methods given by number of researchers. Among them Data Mining techniques is one of the most likely used techniques to detect phishing activity. Data mining is a new solution to identifying phishing is problem. So data mining is a new research direction towards the identifying and avoiding phishing website. Associative Classification is a grooming research method in data mining. There for it is an attractive research topic that identifying phishing using associative classification [8].

2. LITERATURE SURVEY

Phishing website is a current problem, but because of its large impact on the commercial and on-line marketing sectors and since preventing such attacks is an important step towards protecting against website phishing attacks, there are various promising approaches to this problem and a comprehensive collection of related works. In this section, we briefly survey existing anti-phishing solutions and list of the related works [9].

Neda Abdelhamid et. al proposed the Multi-label Classifier based Associative Classification (MCAC) data mining approach for detecting phishing website. The associative classification effectively detects phishing websites with high accuracy. MCAC creates new hidden knowledge (rules) that other algorithms are unable to find. The MCAC increases classifiers predictive performance. The Associative Classification merges association rule and classification technique of data mining. The AC algorithm works in three phases. In first phase, it looks for hidden correlations among the attribute values and the class attribute in the training data set and creates class association rule. In second phase, it begins the ranking and pruning procedure. The ranking procedure sorts rules according to certain thresholds like confidence and support. Pruning, duplicate rules are rejected. The third phase measures accuracy or error-rate of the classifier [2].

Phishing is the act of attempting to acquire information such as usernames, passwords, and credit card details (and sometimes, indirectly, money) by masquerading as a trustworthy entity in an electronic communication. Communications purporting to be from popular social web sites, auction sites, online payment processors or IT administrators are commonly used to lure the unsuspecting public. Phishing emails may contain links to websites that are infected with malware. Phishing is typically carried out by e-mail spoofing or instant messaging and it often directs users to enter details at a fake website whose look and feel are almost identical to the legitimate one. Phishing is an example of social engineering techniques used to deceive users, and exploits the poor usability of current web security technologies. Attempts to deal with the growing number of reported phishing incidents include legislation, user training, public awareness, and technical security measures. Website phishing is considered one of the crucial security challenges for the online community due to the massive numbers of online transactions performed on a daily basis. Website phishing can be described as mimicking a trusted website to obtain sensitive information from online users such as usernames and passwords. Black lists, white lists and the utilization of search methods are examples of solutions to minimize the risk of this problem. One intelligent approach based on data mining called Associative Classification (AC) seems a potential solution that may effectively detect phishing websites with high accuracy [4].

3. DESIGN ISSUES

For detecting the website phishy we use the following techniques.

1. Feature Extraction.
2. MCAC rule learning and example.
3. MCAC Algorithm.

3.1 Feature Extraction:

In this section we extract the website features for detection of phishy website. These features are:

1) IP Address.
2) Lon URL.
3) URL having @ symbol.
4) Adding prefix and suffix.
5) Sub-domains.
6) Fake HTTPs protocol.

7) Request URL.
8) Anchor of URL.
9) Server form handler.
10) Abnormal URL.
11) Using pop-up window.
12) Redirect page.
13) DNS Record.
14) Hiding links.
15) Website traffic.
16) Age of Domain.

3.2 MCAC rule learning and example

In this section we show an example for how MCAC generates and produces the rules. For that assume *minsupp and minconf* is 20% and 40 % respectively.

Table 1 Displays an initial training dataset.

Instance number	Att1	Att2	Att3
1	a1	b1	c 2
2	a1	b1	c 2
3	a2	b1	c1
4	a1	b 2	c1
5	a3	b1	c1
6	a1	b1	c2
7	a4	b2	c1
8	a1	b2	c1
9	a1	b3	c1
10	a1	b2	c2

The candidate rules extracted are depicted in Table 3a.While the algorithm is generating the rules, it checks whether there exists a candidate rule that is already extracted with a similar body of the current rule. If this condition is true, the algorithm appends the current rule with the already extracted rule to form a new multi-label rule. For example, in Table 3a, the attribute value ha1i is connected with two class labels, i.e. (c2, c1) with frequencies 4 and 3, respectively. Current AC algorithms will produce only one rule for this attribute value, i.e. a1->c2 and simply discards class (c1) because (c1) has more number of occurrences in the training data set with attribute value <a1>. However, MCAC produces a multi-label rule for <a1> as a1->c2 $\vee$ c1.

Table 3a Candidate rules produced from the data in Table 1.

Rule items		Support count (%)	Confidence (%)
Attribute	Class		
a1	c2	40	57
a1	c1	30	42
b1	c2	30	60
b1	c1	20	40
b2	c1	30	75
a1 ∧b1	c2	30	100
a1∧b2	c1	20	66

Table 3b Candidate multi-label rules produced from the data in Table 1 via MCAC.

Rule items		Support (%)	Confidence (%)
Attribute	Class		
a1	c2,c1	35	50.00
b1	c2,c1	25	50.00

Table 3c The classifier of MCAC algorithm from the data in Table 1.

Rule items		Support count (%)	Confidence (%)
Attribute	Class		
a1 ∧ b1	c2	30	100
b2	c1	30	75
a1	c2, c1	35	50.00
b1	c2, c1	25	50.00

The candidate multi-label rules must pass the *minsupp* and *minconf* in order to be considered while making the classifier and their actual support (frequency) and confidence values are updated when they are formed as displayed in Table 3b. The candidate rules in bold within Table 3a represents the possible candidate multi-label rules shown in Table 3b. Once the rule extraction is finished MCAC sorts all possible candidate rules according to confidence, support, and rule's length. The candidate rules are ready for evaluation against the training data set in order to choose the best ones that can make the classifier. MCAC selects rules that have at least one training data coverage. Table 3c shows the classifier devised by our algorithm that consists of four rules, two of which are multi-label ones, i.e. a1->c2 ∨ c1 and b1->c2 ∨ c1.

3.3 MCAC Algorithm

The phishing detection process using our model from the user prospective can be explained in the following steps:

(1) The end-user clicks on a link within an email or browses the internet.

(2) He will be directed to a website that could be legitimate or phishy. This website is basically the test data.

(3) A script written in PHP that is embedded within the browser starts processing to extract the features of the test data (current website) and saves them in a data structure.

(4) Now, the intelligent model will be active within the browser to guess the type of the website based on rules learnt from historical websites (previous data collected). The rules of the classifier are utilized to predict the type of the test database on features similarity.

(5) When the browsed website is identified as legitimate no action will be taken. On the other hand, when the website turned to be phishy, the user will be warned by the intelligent method that he is under risk.

We have implemented steps (1)–(4) in the above model where we utilized MCAC learning strategy to generate the rules. MCAC comprises of three main steps: Rules discovery, classifier building and class assignment. In the first step, MCAC iterates over the training data set (historical websites features) in which rules are found and extracted. In this step, the algorithm also merges any of the resulting rules that have the same antecedent (left hand side) and are linked with different classes to produce the multi-label rules. In addition, redundant rules that have no training data coverage are discarded. The outcome of the second step is the classifier which contains single and multi-label rules. The last step involves testing the classifier on test data set to measure its performance. In predicting a website the rule in the classifier that matches the test data features often fired to guess its type (class).

The general description of the MCAC algorithm is,

Input: Training data *D*, minimum confidence (*MinConf*) and minimum support (*MinSupp*) thresholds

Output: A classifier Pre-processing: Discretize continuous attributes if any

Step One:

- Scan the training data set *T* to discover the complete set of frequent attribute values.
- Convert any frequent attribute value that passes *MinConf to* a single label rule.
- Merge any two or more single label rules that have identical body and different class to derive the multi-label rules.

Step Two:

- Sort the rule set according to confidence, support and rule's length.
- Build the classifier by testing rules on the training data and keeping those in Cm that have data coverage.

Step Three:

Classify test data using rules in CM

4. RESULT AND ANALYSIS

For checking the URL is phishy or legitimate, we apply the MCAC algorithm. Input for MCAC algorithm is m*insup*, *minconf* and training dataset.

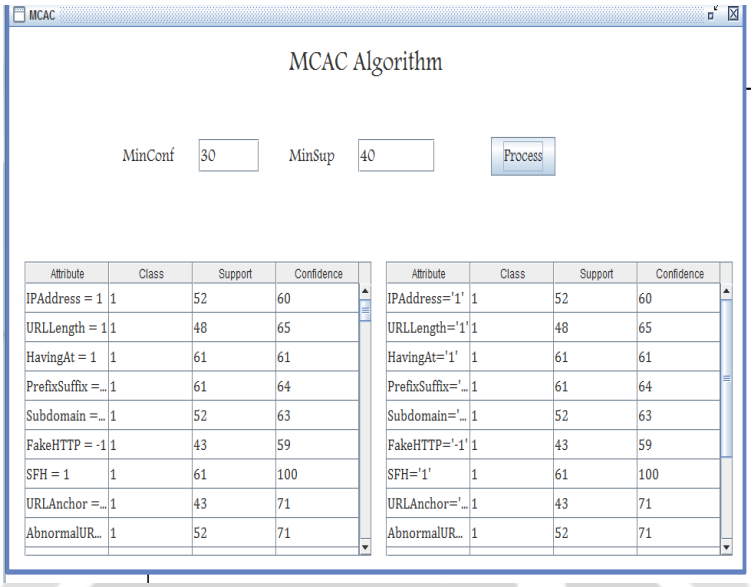

Fig 1:Screenshots of MCAC Algorithm.

User gives input as URL and click on check button. If URL is legitimate then it will display message as "website is legitimate" and also display extracted features of URL.

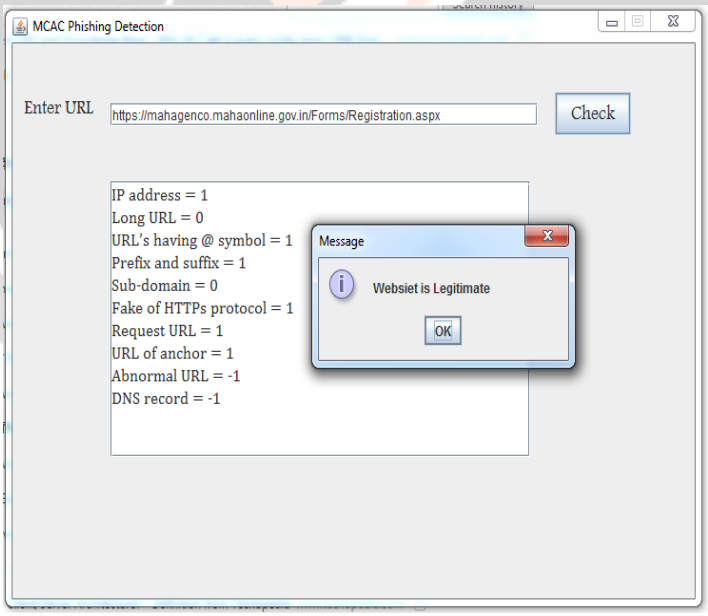

Fig 2:Screenshot of legitimate website.

User give input as URL and click on check button.If URL is phishy then it will display message as "website is phishy" and also display extracted features of URL.

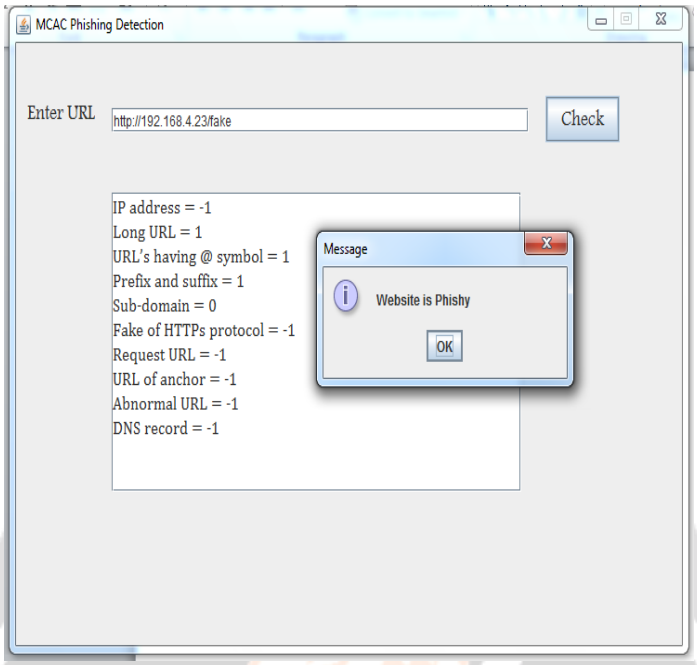

Fig 3: Screenshot of phishy website.

4. CONCLUSIONS

By using Phishing detection tool we can detect website is phishy or not. Phishier mimics a trusted website to obtain sensitive information from users such as usernames and passwords. Our system goal is to detect phishy website by using MCAC algorithm. The MCAC algorithm generate rules further that rules are sorted by using sorting algorithm. By using the Feature Extraction algorithm we can extract the features and store in training dataset. That features are used to find out the website is phishy or not. If the website is phishy then display warning message to user.

5. REFERENCES

1. Abdelhamid, N., Ayesh, A., & Thabtah, F. (2013) Associative classification mining for website phishing Classification. In Proceedings of the ICAI '2013 (pp.687–695),USA.
2. Extraction of Feature Set for Finding Fraud URL Using ANN Classification in Social Network Services. IPGCON-2015, SPPU, PUNE.
3. Pallavi D. Dudhe, Prof. P.L. Ramteke, (2015) Detection of Websites Based on Phishing Websites Characteristics, International Journal of Innovative Research in Computer and Communication Engineering, april 2015.
4. Aanchal Goel, Deepika Sharma.*Department of Computer Science and Engineering, Raj Kumar Goel Institute of Technology for Women,Ghaziabad, Uttar Pradesh, India, "*Prevention from hacking attacks: Phishing Detection Using Associative Classification Data Mining", November 2014, Volume 2 Issue 6, ISSN 2349-4476.
5. Vaibhav V. Satane, Arindam Dasgupta(2013) Survey Paper on Phishing Detection: Identification of Malicious URL Using Bayesian Classification on Social Network Sites, International Journal of Science and Research (IJSR) 2013.
6. Sonali Taware, Chaitrali Ghorpade, Payal Shah,Nilam Lonkar (2015) Phish Detect: Detection of Phishing Websites based on Associative Classification (AC), International Journal of Advanced Research in Computer Science Engineering and Information Technology, Volume: 4 Issue: 3 22-Mar-2015,ISSN_NO: 2321-3337.
7. Komatla. Sasikala, P. Anitha Rani(2012) " An Enhanced Anti Phishing Approach Based on Thresho ld Value Differentiation", International Journal of Science and Research (IJSR) 2012.

8. Mitesh Dedakia, Khushali Mistry, Phishing Detection using Content Based Associative Classification Data Mining *Journal of Engineering Computers & Applied Sciences(JECAS) ISSN No: 2319-5606 Volume 4,*
9. Moh'd Iqbal AL Ajlouni1*, Wa'el Hadi2, Jaber Alwedyan3 Dept of Business Administration, Al-Zaytoonah University, "Detecting Phishing Websites Using Associative Classification", ISSN 2224-5782 (print) ISSN 2225-0506, Vol.3, No.7, 2013

BIOGRAPHIES

	**T.Bhaskar** is currently working as Asst. Professor in Computer Engineering Department, Sanjivani College of Engineering, Kopargaon and Maharashtra India. His research interest includes data mining, network security
	**Sonali Aher** is pursuing B.E Computer Engg in SRESCOE, Kopargaon. Her areas of research interests include Information Security; Data Mining.
	**Nikita Bawake** is pursuing B.E Computer Engg in SRESCOE, Kopargaon. Her areas of research interests include Information Security; Data Mining.
	**Akshada Gosavi** is pursuing B.E Computer Engg in SRESCOE, Kopargaon. Her areas of research interests include Information Security; Data Mining.
	**Swati Gunjal** is pursuing B.E Computer Engg in SRESCOE, Kopargaon. Her areas of research interests include Information Security; Data Mining.