# Phishing Detection using Content Based Associative Classification Data Mining

**Mitesh Dedakia**, Student, CSE, PIET, Vadodara, India
**Khushali Mistry**, Assistant Professor, CSE, PIET, Vadodara, India

## Abstract

*Phishing is a technique imitating an official websites of any company such as banks, institutes, etc. The purpose of phishing is that to theft a private and sensitive credentials of users such as password, username, PIN etc. Phishing detection is a technique to deal with this kind of malware activity. It is intended to prevent a phishing using data mining technique. MCAC Algorithm is gives higher efficiency towards to detect phishing activity. In MCAC algorithm does not consider a content based features of websites. It is intended to add content and page style features in that algorithm and change the system for better performance. This paper shows proposed method and flow chart. This paper also shows all the features of website which are considered during experimental analysis.*

**Keywords**: Phishing Detection, Data Mining, Associative Classification, Security, Multi Class

## Introduction

Phishing is a method to imitating a official websites or genuine websites of any organization such as banks, institutes social networking websites, etc. Mainly phishing is attempted to theft private credentials of users such as username, passwords, PIN number or any credit card details etc. Phishing is attempted by trained hackers or attackers.

Phishing is mostly attempted by phishy e-mails. This kind of Phishy e-mails may contains phishy or duplicate link of websites which is generated by attacker. By clicking these kinds of links, it is redirected on malicious website and it is easily to theft your personal credentials. Phishing Detection is a technique to detecting a phishing activity. there are various methods are given by so many researchers . Among them Data Mining techniques are one of the most promising technique to detect phishing activity. Data mining is a new solution to detecting phishing issue. So data mining is a new research trend towards the detecting and preventing phishing website. Associative Classification is a grooming research method in data mining. There for it is an interesting research topic that detecting phishing using associative classification.

In next section we are studied different algorithm related to data mining approach.

## Related Work

In this section , we studied and compare different data mining algorithm to detect phishing. In this section we also studied different characteristics and different feature of websites which might be infected by phisher.

### Phishing Characteristic

There are many characteristics and indicators that can be detect the phishing activity. These indicators are sufficient to classify a websites in terms legal and malicious.

There are so many various features categories are there to identify phishing activity.

a) URL & Domain Identity
b) Security & Encryption
c) Source Code & Java script
d) Page style and content
e) Web address and human factor

Data mining approaches can be apply for access to phishing websites. Two approaches of data mining techniques can be utilize for the detecting phishing activity i.e. association rule mining and classification. It is also possible to generate a new approach to detecting phishing combining above two approaches. This approach is known as Associative Classification.

## Different Data Mining Approach to detect Phishing

In table I it is shown that comparison of different method of data mining algorithm.

**Table 1. Comparison of Different Method**

| Methods | Advantage | Limitation |
|---|---|---|
| Legal Solution[6] | a) First time introduced law against phishing<br>b) Possible to catch Phisher<br>c) Possible to punish phisher | a) Do not sufficient to catch all phisher<br>b) Sometime lack to trace them<br>c) Qiuck disappearance of phisher |

| Methods | Advantage | Limitation |
|---|---|---|
| Education | d) Raise user awerness<br>e) Minimised risk<br>f) User lenrn new phishing techniques | d) Qiuck updation is hard task<br>e) Need to spend long time<br>f) Phishers are talented |
| White/Black List Approach[5] | g) Prevent from submitimg user information into fake page<br>h) Saved at client machine<br>i) It is sufficient near about 50-80% | g) Can not cover all websites specially newly introduced<br>h) Time consuming to update lists |
| Rule-Based Approach[7] | j) Gives rule based classification algorithm<br>k) Used larger feature set<br>l) Promising hiogher accuracy | i) Rule used based on human experience rather than intelligent data mining technique |
| CANATA Based Approach[8] | m) Content based approach<br>n) Moderate accuracy | j) Does not deal with hidden text |

We are contracting on MCAC algorithm and analysis complete phishing detection system developed with MCAC algorithm. MCAC algorithm is proposed by Neda Abdelhamid et. al. In this existing method, authors proposed Associative Classification method called Multi-label classifier Based Associative Classification (MCAC). Authors also identified different features set of legal websites. Proposed MCAC algorithm generate the hidden rule which could not be generated by any algorithm proposed previously.

This system developed with the consideration of 16 different features of websites from 4 different categories which are URL & Domain Identity, Security & Encryption, Source Code & Java script and Web address and human factor and all 16 features are listed as below.

a) IP address
b) Long URL
c) URL’s having @ symbol
d) Prefix and suffix
e) Sub-domain (dots)
f) Misuse/fake of HTTPs protocol
g) Request URL
h) Server form handler
i) URL of anchor
j) Abnormal URL
k) Using pop-up window
l) Redirect page
m) DNS record
n) Hiding the links
o) Website traffic
p) Age of doermain

This system has so many advantages which are Generates hidden rule, Detect phishy websites, Overcome of blacklist approach, Listed different features of website. This system does not consider Content and page style features of websites so this will be a research topic.

## Problem Statement

Multi label Class Associative Classification (MCAC) is a lasted technique to detecting phishing using various features of website. This method generate new hidden rule which cannot be generate by any other method. Its improve classifier predictive performance. There is one major limitation of this method is that; this method does not consider a content based feature to detecting phishing activity. Problem statement of this research work could be limitation of MCAC

## Proposed Work

In existing MCAC algorithm it was consider 16 different features of website from URL and Domain identity, Security and Encapsulation, source code and JavaScript etc to detecting the behavior of website. They are not considering Page Style and Contents features of websites to detecting phishing activity. Now days it is possible to make phishing attack through changing content of website. In new proposed model can be a modified version of MCAC model which include a Content and Page Style.

Fig. 1 shows a new modified proposed work for this research work. In that flow firstly users click on the browser link. If user click than it is redirected respective page. After redirected on website, PHP script embedded within browser and extracts the features of websites and makes it suitable for test data. After that MCAC algorithm start working and it is makes association rule for extracted features. After that based on that rule it is classify the websites in terms of legitimate and phishing. This work is done in MCAC existing methods.

Now Proposed work flow also perform the above step and add 2 steps in that which are based PHP script we extract content based features of website. And applied modified algorithm or existing algorithm on this features and will make association rule and classify website after applied content based filtering algorithm.

### ProposedContent Based Features

In this section, it includes some proposed content and page style based features of website which will considered in research work.

a) Spelling Error
b) Copying Website
c) Using Forms with Submit Button
d) Disabling Right-Click
e) Using Pop-Ups Windows

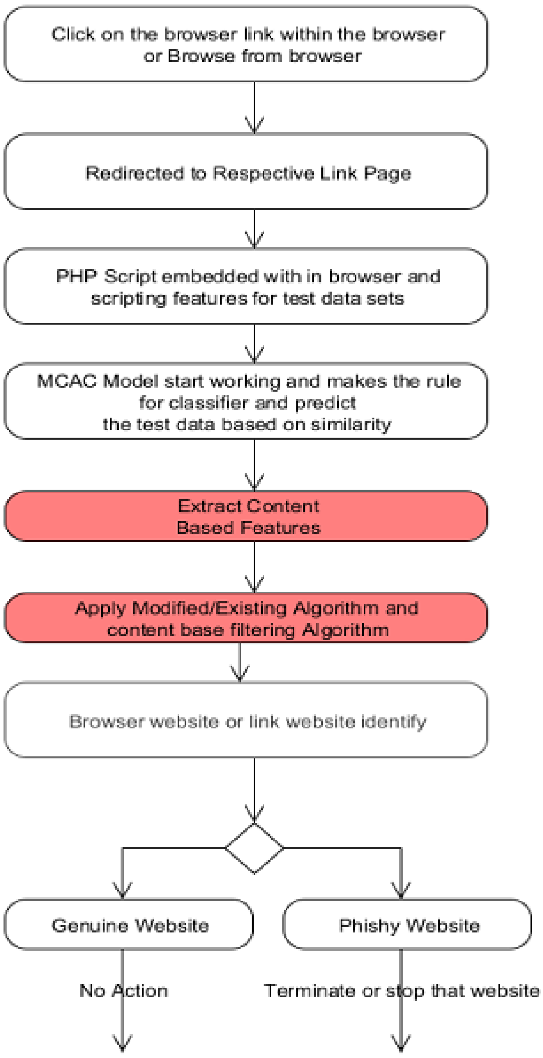

Click on the browser link within the browser or Browse from browser
Redirected to Respective Link Page
PHP Script embedded with in browser and scripting features for test data sets
MCAC Model start working and makes the rule for classifier and predict the test data based on similarity
Extract Content Based Features
Apply Modified/Existing Algorithm and content base filtering Algorithm
Browser website or link website identify
Genuine Website
Phishy Website
No Action
Terminate or stop that website

Fig. 1 Propose Method Flow chart

### Proposed Content Based Features Classification Rule Table

In next section we are study rule learning and prediction techniques and generates the following rules for proposed content based features from training datasets.

**Table 2. Rule Table**

| Features | Phishy | Genuine | Suspicious |
|---|---|---|---|
| Spelling Error | YES | NO | --- |
| Copying Website | Path Does not match | Path Match | ---- |
| Using Forms with Submit Button | ≥1 | =1 | ---- |
| Disabling Right-Click | Yes | -- | No |
| Using Pop-Ups Windows | ≥2 && ≤4 | =1 | >4 |

### Example of Rule Learning and Prection

It is shown how to generate rules for the proposed method. Rule can be generate from the training dataset as shown in below. Suppose attribute value A1 is lies into two class C1 and C2 with the frequency 4 and 3. Associative Classification algorithm will generate one rule i.e. A1 → C1. because we consider minimum support and minimum confidence 20% and 40% respectively.

**Table 3. Example of Rule Learning And Prediction**

| Serial No. | Attribute 1 | Attribute 2 | Class |
|---|---|---|---|
| 1 | A1 | B1 | C2 |
| 2 | A1 | B2 | C2 |
| 3 | A2 | B1 | C1 |
| 4 | A3 | B1 | C1 |

### Algorithm

In this section algorithm is given. This algorithm is used MCAC algorithm and this is a algorithm steps of the phishing detection system.

*Input: Training Dataset (D), Features Criteria (as minimum support minSup and minimum confidence minConf*

*Output: Classifier (3 Different Class)*
*Pre-processing: Change URL pattern format(if required)*

**Step 1**

a) Input dataset value
b) Extract following Features of websites
c) URL & Domain Identity
d) Security & Encryption
e) Source Code & Java script
f) web address

**Step 2**

a) Extract the Content Base features of websites

**Step 3**

a) Generate association rule for frequent value

b) Convert frequent value that passes minConf and minSup to single label rule

c) Merge any two single label rule to derive multi label rules

**Step 4**

a) Sort the rule classify the test data

## Experimental Analysis

*Databse Description*

In this section, we are discuss detail of database. We are using two publicly freely available datasets are used to make feasible our experimental environment of our proposed method.

We used PhishTank dataset. PhishTank is a free community site where anyone can submit, verify, track and share phishing data. This dataset is in the form of .csv file format.

## Experimental Result

Database is 1d column database. This database contain mass URLs. In this database all types of URL are found i.e. Phishy, Legitimate and suspicious

We are taking consider 16 features for existing work and 21 features for proposed in this 16 previous features consider and 5 proposed features consider . All the considered features have categorize into 3 different values. This three values can be labeled as Legitimate, Suspicious, and Phishy. This Three different category values are replace with 1,0,-1 respectively.

Generally there are two class which are Legitimate and Phishy and we are also consider third class that is Suspicious. This value will replace with 1,-1 and 0 respectively. Result analysis shown for existing and proposed is shown in below table V and VI.

**Table 4. Input Database**

| Experimental No. | Total URLs | Genuine URLs | Phishing URLs | Suspicious URLs |
|---|---|---|---|---|
| E1 | 35 | 10 | 20 | 5 |
| E2 | 70 | 20 | 35 | 15 |
| E3 | 110 | 20 | 55 | 35 |
| E4 | 40 | 0 | 0 | 40 |

We experiment on 4 different dataset which is shown in table IV . Based on that dataset we got the following result which is shown in table VII.

At the end we compare the existing method with our proposed method and we found that accuracy of our proposed method is increased. Our proposed solution have another advantages that we are consider 21 features of websites so security is high compare to existing system. We calculate accuracy of existing work and proposed work for every sample of dataset. We got more accuracy every time compare two existing work. Comparison accuracy graph is shown in fig. 2.

We also calculate time complexity and efficiency of proposed solution and we found that time complexity and efficiency of proposed solution did not affected compare to existing work. Comparison of existing work and Proposed work shown in fig. 3.

**Table 5. Sample of Phishing Data Result Analysis**

| nchor of URL | Requested URL | Server from Handler | Lenght of URL | Having @ Symbol | Prefix sufix | IP | Sub-Domain | Web Traffic | DNS Record | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -1 | 0 | -1 | 1 | 1 | 1 | -1 | 0 | 1 | 1 |
| 0 | 0 | 1 | -1 | 1 | 1 | 1 | 1 | -1 | 1 | 1 |
| 1 | 1 | 1 | 0 | -1 | 1 | 1 | 0 | 1 | 1 | -1 |
| 1 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | 1 | -1 |

**TABLE 6. Sample of Phishing Data Result Analysis for Proposed Work**

| pellling Error | Coping Website | Using Forms with Submit Button | Disabling Right-Click | Using Pop-Ups Windows | Class |
|---|---|---|---|---|---|
| 1 | -1 | 1 | 0 | 1 | 1 |
| -1 | 1 | 1 | 1 | 1 | -1 |
| 1 | 1 | -1 | --1 | -1 | -1 |
| 1 | 1 | 0 | 1 | 0 | 1 |

**TABLE 7. Result Analysis of Input Data**

| Experimental No. | Total URLs | Genuine URLs | Phishing URLs | Suspicious URLs | Incorrectly Classified | Accuracy (%) |
|---|---|---|---|---|---|---|
| E1 | 35 | 9 | 25 | 1 | 2 | 94.29 |
| E2 | 70 | 17 | 27 | 6 | 5 | 92.85 |
| E3 | 110 | 21 | 77 | 12 | 7 | 93.63 |
| E4 | 40 | 2 | 26 | 12 | -- | -- |

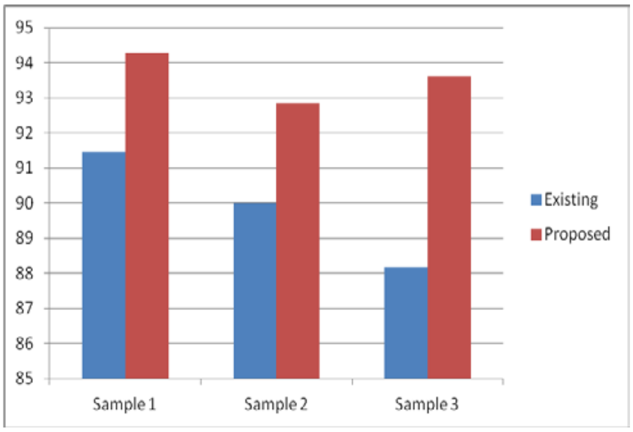
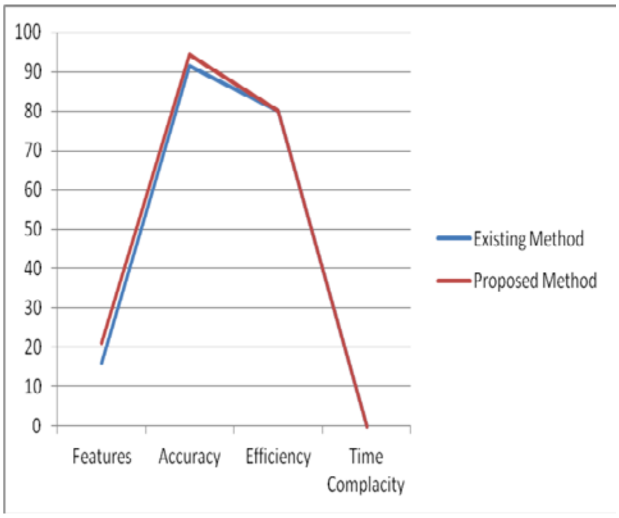


Fig: Accuracy Graph(Based on Sample)



Comparision of Existing and Proposed Work

## Conclusion

This research work consists comparative study of all existing methods. Existing MCAC has limitation that is MCAC is not considering content based features of website to detecting phishing activity. In this research work, develop a modified system that consider content based features of website such as spelling error, coping website, using forms with submit button, disabling right click, using pop- up windows. In this research work we are experiment existing method and proposed work for single URL and also on dataset and on live dataset. In this research work, it is also listed comparison of existing and proposed work. In near future it is possible to add more content base features to get more accuracy . It is also possible to use content base algorithm instead of MCAC algorithm to design more accurate and get higher security.In this research work we are increase accuracy of phishing detection system.

## References

[1] Maher Aburrous, M.A. Hossain, Keshav Dahal, Fadi Thabtah, " Associative Classification Techniques for predcting e-Banking Phishing Websites", *MCIT,IEEE*, 2010.

[2] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, " Associative Classification Mining for Website Phishing Classification", *Classification Proccedings of the International conference on AI, LV, USA*, 2013, pp 687-695.

[3] Suzan Wedyan, Fadi Wedyan, " An Associative Classification Data Mining Approach for Detecting Phishing Websites, *Jouranal of Emerging Trends in Computing and Information Sciences*, Vol. 4, No. , 12 December 2013, ISSN 2079-8407, pp. 888-899.

[4] Moh'd Iqbal Al Ajlouni, Wa'el Hadi, Jaber Alwedyan, " Detecting Phishing Websites Using Associative Classification", *Journal of Onformation Engineering and Application*,Vol. 3, No. 7, 2013, ISSN 2224-5782, pp. 6-10.

[5] Steve Sheng, Brad Wardman, Gary Warner, " An Empirial Analysis of Phishing Blacklists", *Sixth Conference on Email and Anti-Spam July 16-17*,2009.

[6] Michael Kunz, Patrick Wilson, "Computer Crime and Computer Fraud", *University of Maryland*, 2004.

[7] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, " Phishing detection based Associative Classification data Mining", *ScienceDirect*, 2014, pp. 5948-5959.

[8] Guage X. Jason O., Carolyn P. R., Lorrie, CANITA+: A feature-rich machine learning framework for detecting phishing websites", *ACM*,2011,pp.1-28

[9] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah,Samad Ahmadi, Wael Hadi, " MAC: A multiclass Associative Classification Algorithm", *Journal of Information & Knowledge Management*, Vol. 11 , No. 2, 2012, pp.125011-1-1250011-10.

[10] T. Fadi, C.Peter , Y.Peng, "MCAR: Multi Class Classification Based on Association Rule", *IEEE In-*

*ternational Conference on Computer Systems and Application*,2005, pp. 127-133.
[11]