# An Intelligent Anti-phishing Correlation Based Ensemble Model for Phishing Website Detection

[1]Manisha Shevde, [2]Deepika Singh, [3]Manali Gaikwad

[1,2,3]UG Student, Smt. Indira Gandhi College Of Engineering, NaviMumbai, Maharashtra, India.

*[1]mailme.manishamn@gmail.com, [2]deepikaopsingh@gmail.com, [3]mgaikwad436@gmail.com.*

**Abstract - Phishing websites aim to cause direct harm to the users by using techniques such as email spoofing or messaging. These fake websites are generally a disguise of some popular social websites, banking sites and so on. As number of phishing incidents are growing day by day it is need of the hour to adopt technical security methods. We will propose to start the process of phishing website detection by extracting different features of the webpage such as keywords, strings, images etc. After this ten different classifiers will be built using Naïve Bayes Classifier Algorithm and Support Vector Machine Algorithm for 'string' related features. The values stored in these classifiers will be an approximate prediction whether site is phishing or not. The CBE method will finally combine all values from classifiers and give result. Hierarchical Clustering technique is also incorporated in the model to give category of the fake website.**

***Keywords — Phishing website, Clustering, CBE method, Naïve Bayes Classifier Algorithm, Support Vector Machine.***

## I. INTRODUCTION

Online Phishing is a criminal act of deception in obtaining the sensitive information such as username, passwords, credit card detail and etc. by masquerading as trustworthy entities in electronic communication[1]. It usually gained users credence by proclaiming they are from the legitimate party, such aspopular mail services providers (Gmail, Yahoo) or financial institution (Pay pal, Brandesco Bank), and then directing user to a fraudulent website to harvest users credentials. The word "phishing" comes from the analogy that Internet scammers are using fake email to steal for Passwords and personal financial data from the sea of Internet users .A typical phishing attack begins with an email to the victim, supposedly from a reputable institution, but actually from the phisher. The text of the message commonly warns the user that a problem exists with the user's account that must immediately be corrected. The victim is led to a spoofed website designed to resemble the institution's official website. We propose an intelligent anti-phishing strategy model for phishing website detection and categorization through learning and training samples from large and real daily phishing websites. We first analyze the webpage content and extract 10 different types of features such as title, keywords, description, alt and link text information to represent the webpage. Then we build heterogeneous classifiers according to the characteristics of different features. The CBE method is used to combine the prediction results of these heterogeneous classifiers for phishing detection, and a hierarchical clustering algorithm is employed for categorizing the phishing websites.

## II. RELATED WORK

Over the last few years, many research efforts have been conducted on developing intelligent techniques for phishing website detection and phishing prevention. But there are still some problems: Studies which use URL address, domain name information, website ranking, etc. as the features of the webpage always lead to lower recognition rates; Heuristics and machine learning methods[3] which use features that contain the text and the images of the webpage have been introduced to phishing detection, but most of them have high complexity.

CANTINA is a content-based phishing detection algorithm proposed by Zhang et al[9]. This method calculates term frequency-inverse document frequency (TF-IDF) of the content of a website and generates a lexical signature. The generated lexical signature will be used as the keyword to perform web search using Google search engine. The returned result will be used to classify the legitimacy of a website. However, CANTINA performance will be influenced by the language used in the website.

 

Recently, Huh and Kim propose a new heuristic phishing detection method[3] based on the search results returned from the popular search engines such as Google, Bing and Yahoo. The full URL of a website a user intended to access is used as the search query. The number of returned results and ranking of the website are used for the classification. Usually, searching legitimate websites will return large number of results and ranked top, whereas searching phishing websites will return no result or ranked low. Hara et al proposed an interesting phishing detection method using image similarity-based approach. This method is able to detect phishing websites even if the original websites are not registered first in their system. The authors also employed an application called Img Seek for the image similarity comparison. The comparison is not only involved phishing and targeted legitimate websites, but also included the comparison among the phishing websites which look similar to each other (two different phishing websites may look similar as they are targeting the same legitimate website). If the system discovers a different image displayed on web pages from the past, the system will registers it as unknown site and uses this unknown sites to detect the new phishing sites. By these two factors, this system does not need an initial database. Most of the existing systems suffer problems like incorrect detection in practical world and false alarm. Hence we will propose an intelligent CBE based detection model for phishing website detection.

## III. NEW ANTI-PHISHING MODEL

This section presents the idea how feature extraction of webpage will be done and use of ensemble algorithm to combine the prediction results. The hierarchical clustering algorithm will be employed for automatic phishing categorization.

### *A. Model Description*

Figure 1 shows the architecture of our IACBEM(Intelligent Anti-Phishing Correlation Based Ensemble Model) for Phishing website detection and we briefly describe the main components below.

1) **Feature Extractor:** It is used to extract different features of webpage like title, keyword, strings etc.
2) **Classifier Training Module:** NBC Algorithm and SVM algorithm are employed for building 10 heterogeneous classifiers from different webpage features.
3) **Ensemble Classification Module:** It will combine all prediction results from classifiers and give final detection result.
4) **Cluster Training Module:** Hierarchical clustering algorithm is used for categorization of phishing website.

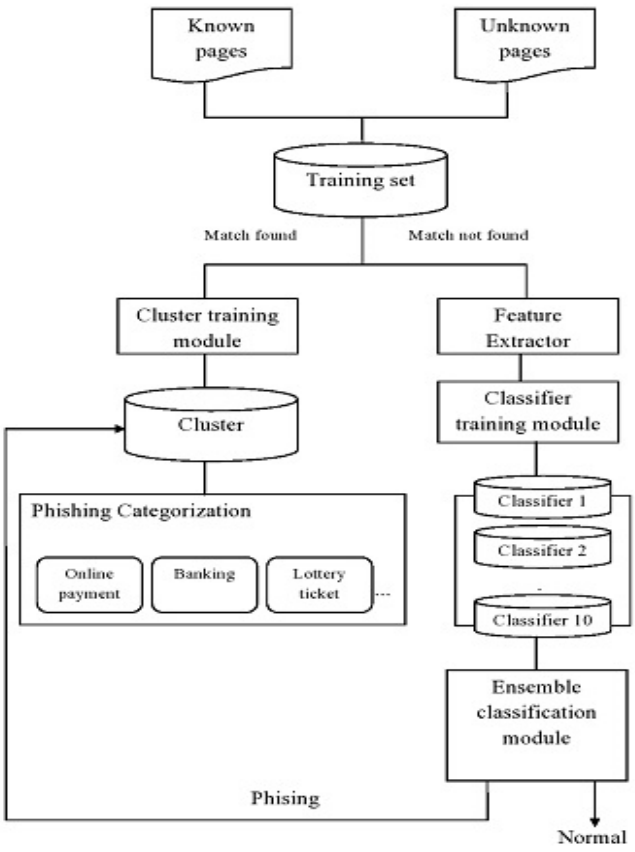


**Figure 1. Intelligent Anti-Phishing CBE Model**

### *B. Feature Extraction*

*i)Naive Bayes Classification Algorithm*

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

It calculates explicit probabilities for hypothesis and it is robust for noise in input data.. It can solve diagnostic and predictive problems.

1. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.
2. Naïve Bayes really easy to implement and often works well.

***Algorithm Working :***

1. Dictionary Generation:- Count occurrence of all word in our whole data set and make a dictionary of some most frequent words.

 

2. Feature set Generation

All documents are represented as a feature vector over the space of dictionary words. For each document, keep track of dictionary words along with their number of occurrence in that document. Calculate Probability of occurrence of each label .Here label is negative and positive.

3. Training

In this phase we have to generate training data (words with probability of occurrence in positive/negative train data files).Calculate for each label. Calculate for each dictionary words and store the result (Here: label will be negative and positive).Now we have word and corresponding probability for each of the defined label.

To build classifiers according to characteristics of different features we will use NBC

Formula:- $P(c|x)=P(x|c)P(c)/P(x)$

***ii) SVM Algorithm***

- Support vector machines focus only on the points that are the most difficult to tell apart, whereas other classifiers pay attention to all of the points.
- The intuition behind the support vector machine approach is that if a classifier is good at the most challenging comparisons (the points in B and A that are closest to each other in Figure 2), then the classifier will be even better at the easy comparisons (comparing points in B and A that are far away from each other).
- Unlike other classifiers, the support vector machine is explicitly told to find the best separating line. How? The support vector machine searches for the closest points which it calls the "support vectors" (the name "support vector machine" is due to the fact that points are like vectors and that the best line "depends on" or is "supported by" the closest points).

  - Once it has found the closest points, the SVM draws a line connecting them.
  - Support vector machines focus only on the points that are the most difficult to tell apart, whereas other classifiers pay attention to all of the points.
  - In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.
  - Classification is done by finding a hyperplane
  - It draws this connecting line by doing vector subtraction (point A - point B). The support vector machine then declares the best separating line to be the line that bisects and is perpendicular to the connecting line.
  - The support vector machine is better because when you get a new sample (new points), you will have already made a line that keeps B and A as far away from each other as possible.

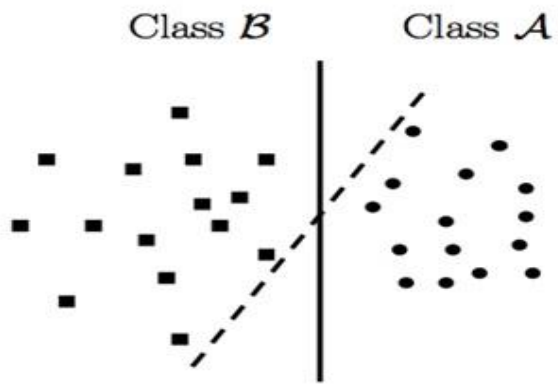


**Figure 2. Classification Of hyperplane**

***C. Correlation Based Ensemble Method for Detection***

Ensemble methods[10] are preferred as they represent good significance over specific predictor regarding accuracy and confidence in classification. The ensemble method with multiple independent feature subsets in order to classify high-dimensional data. First, the method selects the feature subsets using Correlation-based feature Selection with Stratified Sampling. It minimizes the redundancy in the features. After generating the feature subsets, each feature subset is trained using base classifier and then these results are combined using Correlation base ensemble method.

Ensemble learning method is then used to combine all the prediction results from above heterogeneous feature classifiers. The ensemble classifier has better detection performance than any individual classifier. CBE (Correlation Base Ensemble ) which will automatic weighted the predict result from each classifier according to the relationship between classifiers in the history detection results.

1. CBE will automatic weight the predicted results from each classifier
2. Given a webpage x, Ci(x) and Cj(x) are predicted results from classifier i and j respectively.
3. Then the possibility that Ci(x) is final result is defined as follows:-

$$f(C_i(x),C_j(x),Class=C_i(x))=\frac{Count(C_i(x),C_j(x),Class=C_i(x))}{Count(C_i(x),C_j(x))}$$

**Figure.3 Probabiltiy that Ci(x) is final result**

4. While Count(Ci(x),Cj(x)) denote the number of training samples that were predict as Ci(x) by classifier i and as Cj(x) by classifier j

 

5. Then we generate the final ensemble result from all classifier as:-

$$Score(x) = \sum_{C_i(x)\neq 0} C_i(x) \times \frac{\sum f(C_i(x), C_j(x), Class = C_i(x))_{C_j(x)\neq 0}}{\sum_{C_j(x)\neq 0} |C_j(x)|}$$

**Figure. 4 Final Ensemble Result**

### *D. Hierarchical Clustering Algorithm*

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. A cluster is a collection of phishing websites that share common traits between them and are "dissimilar" to the phishing website belonging to other clusters.

Given pair wise dissimilarities dij (dissimilar Xi, Xj) between data points, hierarchical clustering produces a consistent result, without the need to choose initial starting positions( number of clusters).Given the linkage, hierarchical clustering produces a sequence of clustering assignments. At one end, all points are in their own cluster, at the other end, all points are in one cluster. In this CBE Method the legitimate websites are clustered as entertainment, financial, sports etc. on the basis of different words extracted from the webpages. Phishing websites are categorized in 'others' category.

Hierarchical algorithms can be categorized as 'agglomerative' and 'divisive' [11]

**Agglomerative** algorithms are simple and due to its lower computation cost we use agglomerative clustering.
- Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. The classic example of this is species taxonomy. Gene expression data might also exhibit this hierarchical quality (e.g. neurotransmitter gene families). Agglomerative hierarchical clustering starts with every single object (gene or sample) in a single cluster. Then, in each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster.

- The hierarchy within the final cluster has the following properties:

1.Clusters generated in early stages are nested in those generated in later stages.

2. Clusters with different sizes in the tree can be valuable for discovery.

***Algorithm Working:***

- Assign each object to a separate cluster.
- Evaluate all pair-wise distances between clusters (distance metrics are described in Distance Metrics Overview).
- Construct a distance matrix using the distance values.
- Look for the pair of clusters with the shortest distance.
- Remove the pair from the matrix and merge them.
- Evaluate all distances from this new cluster to all other clusters, and update the matrix.
- Repeat until the distance matrix is reduced to a single element.

### *E .Comparison of different classification methods*

One of the different methods used for phishing website detection include Majority Vote Method and our proposed Intelligent CBE method.

| Ensemble Method | Precision | recall |
|---|---|---|
| Majority Vote | 95.45% | 93.51% |
| CBE Method | 98.12% | 98.73% |

The results obtained precision wise are tabulated as follows:

**TABLE I.PREDICTED RESULTS OF TWO METHODS**

From above table we confirm that our proposed CBE Method gives results in higher precision and recall. The CBE Method gives results with exact precision whereas Majority vote method just gives a count of the majority of the results.

### *F. Implication Of the project*

The CBE method has proved quite useful in detecting phishing websites as it covers each and every aspect of the website from extracting critical features of the webpage to categorization of the website. This CBE based system when installed on any compatible computer will efficiently detect phishing website even for a naive user with less complexities.

## IV. CONCLUSION

We know that phishing is an attack which aims at exploiting weaknesses found during electronic communications such as user leaking their passwords to any unknown random websites. Hence awareness and defence both are required against these sites. Our proposed system model will take the webpage through various levels of detection and user of this system will prove beneficial for detecting a phishing website.

 

We have proposed a framework for intelligent phishing website detection via an ensemble of the prediction results generate by different feature classifiers and hierarchical clustering algorithm for phishing categorization. Experimental results show that CBE predicts results with a higher precision than other commonly used phishing detection methods .Empirical studies on large and real daily data sets collected by Kingsoft Internet Security Lab will illustrate that our Intelligent Anti-phishing CBE Method gives good results than other methods in phishing website detection and categorization. CBE Method does not have much complexities in implementation and understanding and gives a good categorization to its users regarding the website.

## REFERNCES

[1] Gang Liu, Bite Qiu, Liu Wenyin. *Automatic Detection of Phishing Target from Phishing Webpage*[C].2010 20th International Conference on Pattern Recognition. Istanbul: IEEE Computer Society, 4153-4156, 2010.

[2] Weiwei Zhuang[1,2], Qingshan Jiang[2,3*], TengkeXiong[2].*An intelligent anti-phishing strategy model for phishing website detection*. 2012 32nd International Conference on Distributed Computing Systems Workshops.[1]Department of Cognitive Science, Xiamen University, Xiamen, 361005, P.R.China.

[3] Jin-Lee Lee, Dong-Hyun Kim, Chang-Hoon, Lee. *Heuristic-based Approach for Phishing Site Detection Using URL Features*. Proc. of the Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015.

[4] W. Zhuang, Y. Ye, T. Li,Q. Jiang. *Intelligent phishing website detection using classification ensemble*, Systems Engineering, Theory & Practice. in Chinese, Volume 31(10), P2008-2020. 2011.

[5] H. Peng, F. Long, and C. Ding. *Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy*. IEEE Trans. Pattern Analysis and Machine Intelligence, 27, 2005.

[6] Written I H ,Frank E . *Data Mining: Practical Machine learning Toolsand Techniques with Java Implementation*[M]. Seattle: Morgan Kaufmann Publishers,2000:265-314.

[7] Salton, G., Buckley, B. *Term-Weighting approaches in automatic text retrieval. Information Processing and Management*,1988,24(5):513~523.

[8]Anti-Phishing Working Group[EB/OL], http://www.antiphishing.org.

[9] Xiang Guang, Hong Jason, Rose Carolyn P. *CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites*[J]. ACM TRANSACTIONS ON INFORMATION AND SYSTEM SECURITY. Vol:14(2), SEP 2011.

[10] Thomas G. Dietterich. Ensemble learning. *In The Handbook of Brain Theory and Neural Networks*[M]. Cambridge: The MIT Press, 2002.

[11] C. Williams. *A mcmc approach to hierarchical mixture modeling*[J]. Advance in NIPS12, pages 680–686, 2000.

 