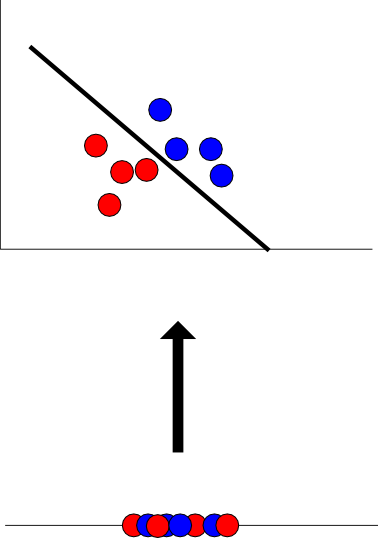
Classification of Phishing website

(Supervised learning)



impacts of Algorithms and features on classifying phishing website

Kishore Ramakrishnan, Andrew Janson | Machine Learning

## Problem:

Phishing is a chronic problem to obtain personal information like account number, password, Social Security Number, etc.… through internet. Initially phishing is done using AOL tool where the hackers tried to snoop America online user’s personal information. In the mid of 2016 the number of phishing websites around the world was approx. 100000 so, companies like pay-pal, G-Mail, etc. have worked on Anti-Phishing tools like secured servers, https, etc. to stop luring of personal information. This report contains the technique to find whether the website in the internet is phishing or non-phishing based on certain features. The major problem here is due to these kinds of phishing websites many thefts occur in day-to day life, to stop this we need to completely ban these phishing website. So, the first step in solving the problem is identifying the website based on features like URL length, Prefix suffix, SSL Service length, etc.

## Solution:

The result for this problem is declaring the website either phishing or non-phishing. This is can be achieved by supervised learning, i.e. classifying the website based on certain constraints. These constraints are developed using various algorithms like Naïve Bayes technique, Support vector machines, Decision tree, etc. These algorithms are built using MATLAB which creates a program to calculate how efficient is the algorithm based on the accuracy it provides and the same algorithm also predicts the result of the newly created website whether it is phishing or non-phishing. Addition to this Feature selection is also done i.e. finding how many number of features are required for the website to decide whether the website is a phishing or non-phishing. So, that to find the minimum number of features to predict the accuracy out of large number of features.

## Dataset:

Dataset is taken from the UCI Repository website for reference purpose, the link is given below.

<http://mlr.cs.umass.edu/ml/datasets/Phishing+Websites>

This dataset consists of totally 11055 websites 30 unique features represent each.

Out of this 11055

60% of dataset for Training the classifier (6633)

40% of dataset for Testing the classifier (4422)

Last column of the dataset is the Label of the instance i.e. “+1” or “-1” to say whether the instance dictating about the website is a phishing or non-phishing website.

## Software:

The main backbone for the creation of these algorithms is MATLAB

MATLAB is a computational software used to work with mathematically and statically registered data. It uses computational workspace to find the result of the command executed in the command prompt at each step. It helps us use the inbuilt library as well as use scratch for new function creation, there are even inbuilt function for Naïve Bayes, Decision tree, SVM, Etc. MATLAB version 2010 till 2016 can be used for this classification. Other than MATLAB software’s like SVM library, WEKA machine learning software’s can also be used. Preference for MATLAB is that it does fast computation even with higher order of matrix compared to other Machine Learning software’s.

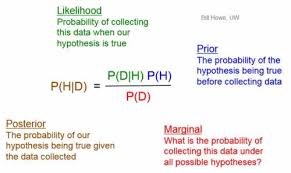
## Results:

## Discussion:

To find whether the given website is phishing or non-phishing we need to build a classifier based on certain predicted data and use the classifier for validation purpose to confirm whether the build classifier is efficient enough to predict the future driven website. This classifier is built using three algorithms Naïve Bayes, Decision tree and Support vector machines. Since the Dataset has only few instances therefore we have high bias and low variance, So Naïve Bayes can be used including with this to check the highest accuracy rate addition of two more algorithms are used to find which algorithm has the best accuracy. Secondarily to increase accuracy of the classifier we reduce the number of features i.e. feature selection technique to identify which set of features are reasonable enough to achieve best accuracy for the classifier. So, that the classifier with highest accuracy can predict the future developed or newly added website correctly. Thus, the classifier with highest accuracy acts the boundary for classifying the website this barrier contains the requirements of non-phishing website thus acts as a tester.

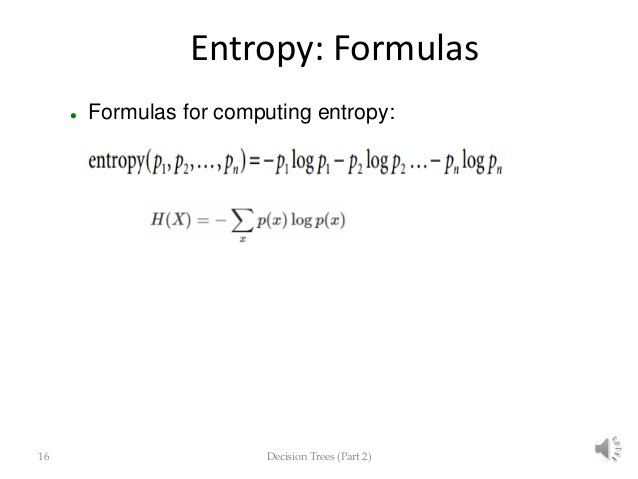
Why Different algorithms have different classification accuracy?

**Naïve Bayes:** Naive Baye technique uses Bayes theorem as the base for building classifier model. Bayes theorem states that the prediction of new event can be obtained if the relevant or related feature knowledge is obtained and if the probability of the event seems to be greater than 0.5 then it seeks to be the correct prediction. Mathematically Bayes theorem is represented by,



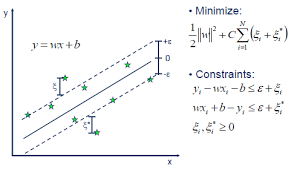
Naïve Bayes is the fastest technique to predict data but doesn’t work for large amount of training data. Addition to the above-mentioned limitation Naïve Bayes also assumes predicting features to be independent but in real time applications it is difficult to get those kinds of feature set.

**Decision Tree:** Decision tree builds classifier as tree structure the entropy formula determines every node and leaf of the tree



The feature with highest entropy is the root node. The feature with second highest entropy is considered as the child of the root node by considering the entropies of the other features the tree is formed. To shorten the tree length pruning is done i.e. removing the features that gives less contribution to the tree classifier. Entropy of a feature defines how efficient is the feature to classify the instance whether it is true or false in our case whether the website is phishing or non-phishing. The main con of this decision tree technique is that each time a new example is created the decision tree should be reformed again so that it might overfit.

**Support Vector Machines:** This technique builds a hyperplane or boundary based on discriminant function. This hyperplane categorizes the data into two groups one on one side of the boundary and the other on second side, the model represents the data in space as points. The boundary can be a linear boundary or non-linear boundary which is later created as linear boundary in difference space using kernel trick. Also, SVM supports unsupervised data’s also by support vector clustering which clusters similar data into same groups.



There are two major advantages of SVM they are:

1. High accuracy
2. Guarantees theoretically regarding underfitting and overfitting

**Feature Selection:** Since data set contains 31 features considering all the features for building a classifier might reduce the accuracy as well as time computation. Therefore, reducing the number of features to build a classifier increases both accuracy and time for building classifier. In this project forward feature selection technique is used i.e. finding a single feature with maximum accuracy and adding features to that single feature in order to increase the accuracy of the classifier, Thus at a single point of time the feature set starts reducing the accuracy that point defines the actual number of features required to build a classifier and what features required to build it to achieve maximum accuracy.

## Inference:

|  |  |
| --- | --- |
| **Total number of features** | **Feature Number** |
| having\_IP\_Address | 1 |
| URL\_Length | 2 |
| Shortining\_Service | 3 |
| having\_At\_Symbol | 4 |
| double\_slash\_redirecting | 5 |
| Prefix\_Suffix | 6 |
| having\_Sub\_Domain | 7 |
| SSLfinal\_State | 8 |
| Domain\_registeration\_length | 9 |
| Favicon | 10 |
| port | 11 |
| HTTPS\_token | 12 |
| Request\_URL | 13 |
| URL\_of\_Anchor | 14 |
| Links\_in\_tags | 15 |
| SFH | 16 |
| Submitting\_to\_email | 17 |
| Abnormal\_URL | 18 |
| Redirect | 19 |
| on\_mouseover | 20 |
| RightClick | 21 |
| popUpWidnow | 22 |
| Iframe | 23 |
| age\_of\_domain | 24 |
| DNSRecord | 25 |
| web\_traffic | 26 |
| Page\_Rank | 27 |
| Google\_Index | 28 |
| Links\_pointing\_to\_page | 29 |
| Statistical\_report | 30 |

Out of these 30 features the required set of features for achieving maximum accuracy is

Feature set= {8, 14, 6, 15, 2, 7, 9, 16, 13, 15, 28, 29, 18}

Target accuracy with all features=91.45% **(Naïve Bayes)**

Feature set Accuracy= 94.2108% (**Naïve Bayes)**

Target accuracy with all features=89.1904% (**Decision Tree)**

## division of labour:

**KISHORE:** Feature selection technique and finding maximum accuracy using Naïve Bayes and Decision tree algorithm.

**ANDREW JANSON:** Feature selection technique and finding maximum accuracy using Support Vector Machine algorithm.

## References:

Mohammad, Rami, McCluskey, T.L. and Thabtah, Fadi (2012) An Assessment of Features Related to Phishing Websites using an Automated Technique. In: International Conferece For Internet Technology And Secured Transactions. ICITST 2012 . IEEE, London, UK, pp. 492-497. ISBN 978-1-4673-5325-0  
  
Mohammad, Rami, Thabtah, Fadi Abdeljaber and McCluskey, T.L. (2014) Predicting phishing websites based on self-structuring neural network. Neural Computing and Applications, 25 (2). pp. 443-458. ISSN 0941-0643  
  
Mohammad, Rami, McCluskey, T.L. and Thabtah, Fadi Abdeljaber (2014) Intelligent Rule based Phishing Websites Classification. IET Information Security, 8 (3). pp. 153-160. ISSN 1751-8709

S. E. Schechter, R. Dhamija, A. Ozment and I. Fischer, “The Emperor's New Security Indicators,” in Proceedings of the 2007 IEEE Symposium on Security and Privacy, Washington, DC, USA, 2007.

“PhishTank,” October 2006. [Online]. Available: http://www.phishtank.com/. [Accessed 25 November 2011].

“WhoIS,” [Online]. Available: http://who.is/. [Accessed 13 March 2012].

“Alexa the Web Information Company,” [Online]. Available: http://www.alexa.com/. [Accessed 26 January 2012].

“More than 450 Phishing Attacks Used SSL in 2005,” [Online]. Available: http://news.netcraft.com/archives/2005/12/28/more\_than\_450\_phishing\_atta cks\_used\_ssl\_in\_2005.html. [Accessed 8 March 2012].