MACHINE LEARNING

WEB PAGE CLASSIFICATION

November 10, 2018

NAME : HIMANSHU
ROLL NO: 16MA20020
Assignment Number 8

Contents

0.1 PROBLEM STATEMENT

The goal of this assignment is to build a classifier to detect phishing/malicious web pages. Phishing is the fraudulent attempt to obtain sensitive information such as usernames, passwords and credit card details by disguising as a trustworthy entity in an electronic communication. The phishing problem is considered a vital issue in industry especially e-banking and e-commerce for the online transactions involving payments.

0.2 METHODOLOGY

Our aim to classify different URLs into three classes whether they are legitimate, suspicious or phisy website. First we look at data we are using and its features. Then we apply few machine learning algorithms for classification. We compare them using available evaluations measures.

0.2.1 Data

A group of researchers have identified different features related to legitimate and phishy websites and collected 1353 different websites from difference sources. Phishing websites were collected from Phishtank data archive (www.phishtank.com), which is a free community site where users can submit, verify, track and share phishing data. The legitimate websites were collected from Yahoo and starting point directories using a web script developed in PHP. The PHP script was plugged with a browser and they collected 548 legitimate websites out of 1353 websites. There is 702 phishing URLs, and 103 suspicious URLs. Following are the features used in the data and there number of categories. Result feature is the label with three classes legitimate, suspicious and phisy represented respectively by 1,0 and -1.

- SFH {1,-1,0}
- popUpWidnow {-1,0,1}
- SSLfinal_State {1,-1,0}
- Request_URL {-1,0,1}
- URL_of_Anchor {-1,0,1}
- web_traffic {1,0,-1}
- URL_Length {1,-1,0}
- age_of_domain {1,-1}
- having_IP_Address {0,1}
- Result {0,1,-1}

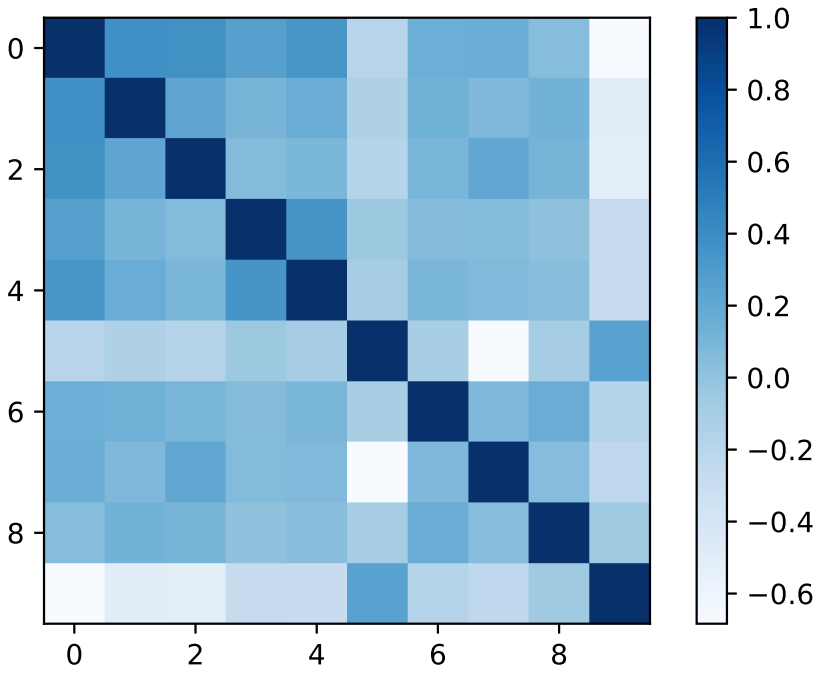


Figure 1: Comparison using correlation coefficients of features and label.

0.2.2 Preprocessing

The data has been pre-processed at source and all the features are categorical. We have used correlation coefficient to determine how our independent variable is connected to each other and to the dependent variable. We observe using plot above that first 6 features of all the listed features above have good relation to our label. We have used k-fold(10 way) validation using all the features and these 6 features separately on all the models(see table1).

0.2.3 Learning Algorithms

Here we have used few algorithms as covered in this Machine Learning course. We have used the scikit-learn library for implementation. Results and comparisons have been provided in Table 1 of observation section.

0.2.4 Evaluation Strategy

In each model we have calculated Confusion matrix, ROC curve and f-score.

0.3 EXPERIMENTAL RESULTS

We have calculated averaged k-fold(10-fold) validation accuracy over all the folds. We have implemented all the models using all the features and only 6-best features. Results of all the experiments have been listed below. We see that Decision trees and Random Forest give good results. We notice in SVM that accuracy obtained using only 6 best features is equal to that obtained using all the features.

Table 1: Test accuracy using respective algrithms using all features and only 6 best features

Algorithm used	Accuracy using all features	Accuracy using 6 best features
SVM	0.8581	0.8588
Naive Bayes	0.8130	0.8093
Decision Trees	**0.8795**	0.8603
Random Forest	**0.8832**	0.8684
MLP	0.8537	0.8493

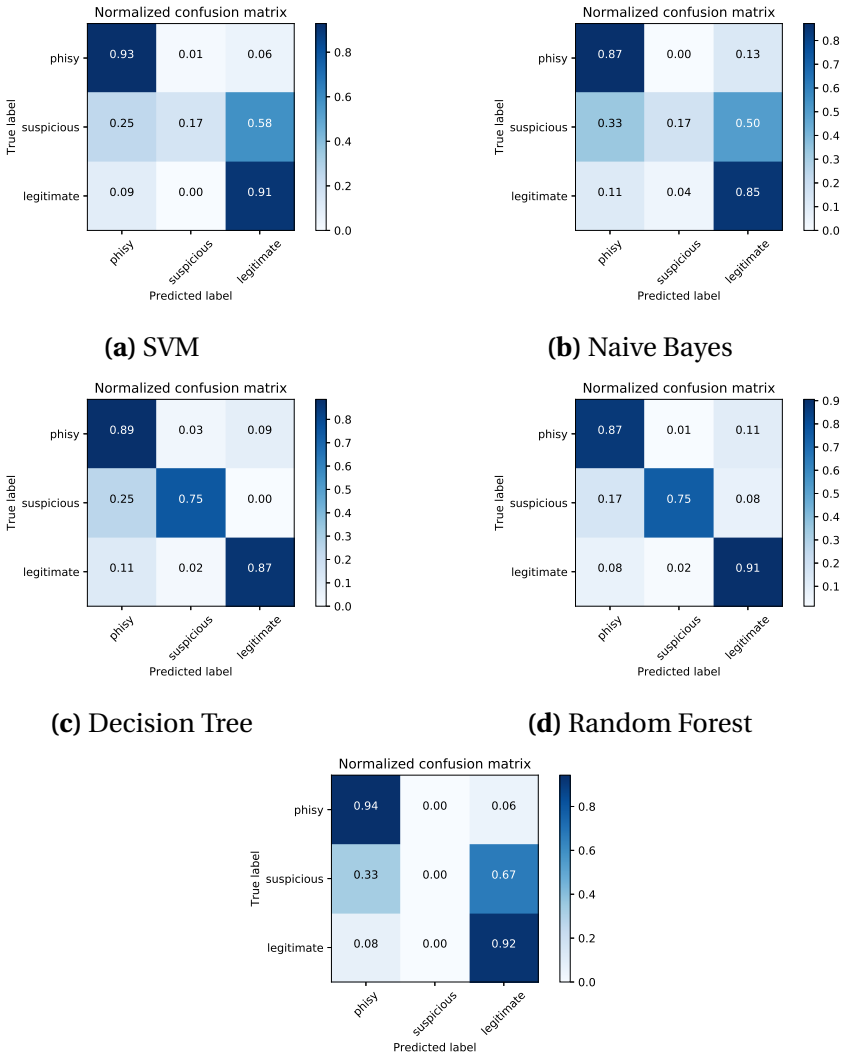
In the Decision tree implementation we look closely at the trees formed. We found out that first feature(i.e URL Anchor) is the most important feature in our tree at each fold and it is significantly more important than rest of the features. We have provided few examples of feature importance of each features during different folds.

Table 2: Importance of different features in Decision trees

0	1	2	3	4	5	6	7	8
0.506	0.091	0.097	0.080	0.086	0.043	0.059	0.023	0.010
0.492	0.095	0.094	0.081	0.093	0.039	0.064	0.023	0.015
0.510	0.096	0.096	0.086	0.089	0.041	0.048	0.019	0.010
0.505	0.080	0.103	0.066	0.111	0.036	0.058	0.025	0.011
0.499	0.081	0.099	0.092	0.083	0.048	0.061	0.019	0.013

0.4 Discussion

With above results we infer that we have got best result using decision tree classifier. Also first feature(i.e "URL Anchor") has played major role in the classification. Earlier we looked at the correlation coefficient of each feature among themselves and with dependent variable. There we too found that dependent variable was very much related to the label. So, we can infer that URL plays a very major role in determining whether site is phisy or not.



(a) SVM

(b) Naive Bayes

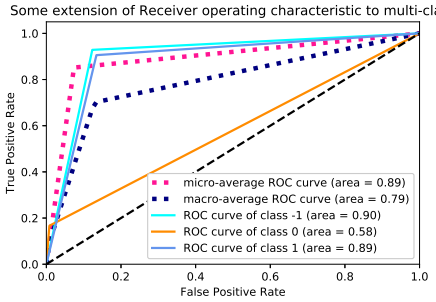
(c) Decision Tree

(d) Random Forest

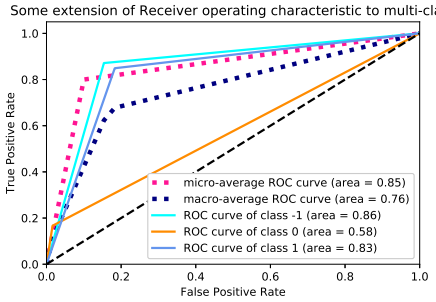
(e) MLP

Figure 2: Confusion matrix of each model

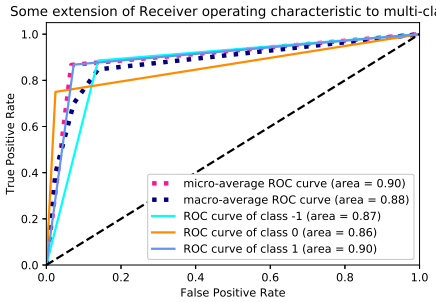
Now we plot the ROC curve of each model implemented upon the data. We can clearly observe that Random forest and Decision tree has more area under the curve and so works better at classification on this web data.



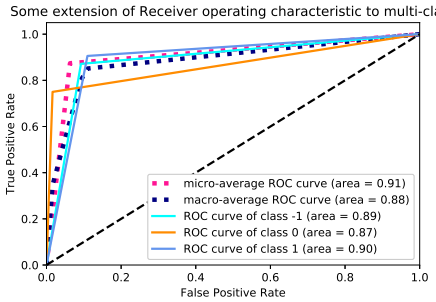
(a) SVM



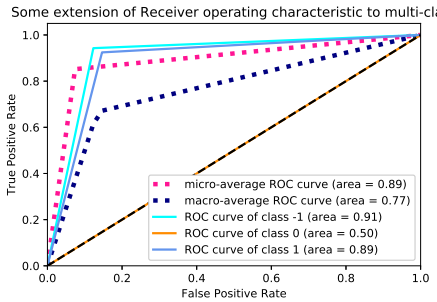
(b) Naive Bayes



(c) Decision Tree



(d) Random Forest



(e) MLP

Figure 3: ROC curve of each model