Detecting Malicious URLs Using Lexical Analysis

Feature selection:

The motive of evaluation and search strategy is to find the features which are

significant and contribute most in the analysis.

In this paper, they used CFSSubsetEval and Infogain as feature selection algorithms.

CfsSubsetEval evaluates the worth of a subset of features by considering the

individual predictive ability of each feature along with the degree of redundancy.

Infogain searches the space of feature subsets by greedy hill-climbing strategy

augmented with a backtracking facility. Later a ranker ranks features by their

individual evaluations.

Lexical analysis:

In the paper, they select Random Forest classifier for further investigation

Confidence interval estimates the prediction interval based on member decision tree scores

Figure 2 depicts the data points overlaid with error

bars. The error bars corresponding to a soft prediction is represented by a Standard

Deviation (SD) of uncertainty for a certain class

* Yellow-filter contains data points whose SD is greater than 0 but less than or equal to .1

– green-filter contains data points whose SD is greater than .1 but less than or

equal to .2

– Orchid-filter subset contains data points whose SD is greater than .2 but less

than or equal to .3

– Blue-filter contains data points whose SD is greater than .3 but less than or

equal to .4

– Red-filter contains data points whose SD is greater than .4

Considering these facts, we can answer how much risky a URL is? If the soft prediction is closer to 1 with a small threshold value of SD score (e.g. up to Orchid filter), the URL is risky. In the opposite way, the closer the URL is to 0 with a small SD score, the more secure it is.

Random forest computes various measure of variable importance that can

be very informative to understand how much each variable contributing to the accuracy of the model. Figure 3 exhibits two different Variable importance graph

for all four binary-class and a multi-class: MeanDecreaseAccuracy that is the

mean decrease in accuracy and MeanDecreaseGini that is the Gini index or

Mean decrease in Node Impurity.

Mean decrease in accuracy is usually described as "the decrease in model accuracy from permuting the values in each feature".

The higher the value of mean decrease accuracy or mean decrease gini score, the higher the importance of the variable to our model.