Case Study Progress Report, IE7275

Group 04 – Ronit Mankad, Shashank Nukala

Classifying Phishing Websites

1. Introduction

Phishing is a form of cyber-attack that utilizes counterfeit websites to steal sensitive user information such as account login credentials, credit card numbers, etc. Throughout the world, phishing attacks continue to evolve and gain momentum. In June 2018, the Anti-Phishing Working Group (APWG) reported as many as 51,401 unique phishing websites.

Identifying phishing websites across different platforms still proves to be a major challenge in the industry. Websites can be classified by monitoring a variety of different indicators. Example: does the website uses https or not, does the website uses an external favicon etc. As the number of indicators increase, they introduce more complexity to the classification process.

We propose a solution which uses machine learning techniques like KNN, CART, Naive Bayes, Logistic regression, Neural Nets, Latent Discriminant Analysis and Support Vector Machine to model and classify the data. A well trained and generalized model will be able to classify websites with a reasonable accuracy. Furthermore, we will also implement dimension reduction techniques like Principal Component Analysis to streamline the dataset. The models will be tested for their accuracy and robustness using evaluation metrics like Lift Chart and ROC curve.

2. Solution Design

- ***Data Visualization***
  Data Visualization is the graphical representation of information and data. By using visual elements like charts, graphs and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

  We will use base-r and packages like *ggplot2* to visualize the data.
- ***Data Preprocessing***
  Data Preprocessing is the process of converting the into a format which is suitable for Data Mining algorithms. It involves steps like handling missing data, encoding categorical data, feature scaling and normalization.

  We will run all data mining algorithms on the normalized dataset.
- ***Data Reduction and Transformation***
  The method of data reduction achieves a condensed description of the original data which is much smaller in quantity but keeps the quality of the original data.

  To eliminate extraneous noise from our data whilst retaining as much information as possible, we will use Principal Component Analysis (PCA) to reduce the dimensionality of our data.
- ***Data Mining***
  The method of data reduction achieves a condensed description of the original data which is much smaller in quantity but keeps the quality of the original data.

  To eliminate extraneous noise from our data while retaining as much information as possible, we will use Principal Component Analysis (PCA) to reduce the dimensionality of our data.
- ***Model Evaluation***
  Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

  We will use model evaluation methods like lift charts and ROC curves to compare different models and choose the model which performs best on the validation data.

3. Algorithm Selection

Our data has two classes which makes this a binary classification task. We propose a solution which uses data mining techniques like KNN, CART, Naive Bayes, Logistic regression, Neural Nets and Support Vector Machine to model and classify the data. A well trained and generalized model will be able to classify websites with a reasonable accuracy.

4. Implementation

Our dataset has 48 attributes all in different scales. We first normalize our entire dataset of 10,000 rows using *Min-Max Normalization*.

We then use *Principal Component Analysis* to reduce the dimensions of our data from 48 attributes to just 13 principal components which capture *61%* of the dataset's variance.

After performing *PCA*, we split the data into training and validations sets with a 60/40 split. The training set contains *6000* rows and validation set contains *4000* rows.

Important: Assume the data used in the following algorithms to be the normalized dataset with PCA performed on it (with 13 predictor variables) unless otherwise mentioned.

We then use the following algorithms on for the classification –

1. ***KNN (K Nearest Neighbors)***

   We applied KNN on the dataset with our CLASS_LABEL attribute as the response variable and the other 13 PCA components as the predictor variables. We used a $k = \sqrt{6000} \div 2 + 1 = 39$.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1820  196
         1  157 1827

               Accuracy : 0.9117
                 95% CI : (0.9025, 0.9204)
    No Information Rate : 0.5058
    P-Value [Acc > NIR] : < 2e-16
```

   This k value is a good balance between preserving the noise in the dataset while also preventing overfitting. Using k = 39, we get a classification accuracy *91.17%.*

2. ***Binary Logistic Regression***

   We applied *Logistic Regression* on the dataset with our CLASS_LABEL attribute as the response variable and the 13 PCA components as the predictor variables. We set the cutoff probability as 0.5 achieving an accuracy of *90.375%.*

   However, *Logistic Regression* on the original data performs better than on the PCA dataset with an accuracy of *94.5%.*

3. ***Naive Bayes Classification***

   The Naïve Bayes classifier performs poorly on the dataset with an accuracy of *71.125%.*

5. Future Work

In the upcoming weeks, we want to apply more sophisticated algorithms like Classification Trees, Neural Networks and Support Vector Machines. We will also compare all the models using *lift charts* and *ROC curves*. Moreover, we plan on creating more visualizations which will help us figure out the most influential attributes.