Final Project Proposal

Anshupriya Srivastava

11/5/2019

Overview

This project will use a predictive model to identify whether or not a fraudulent transaction has been made. The data is obtained from European credit cardholders and contains transactions made in September 2013. The goal is to learn various techniques that are used to deal with highly unbalanced datasets. Also, the right kind of evaluation metric needs to be selected in this case so that the model can provide valuable information and answer the research question with the highest possible accuracy.

Research questions

This project aims to identify fraudulent credit card transactions so that customers are not charged for items they did not purchase. Much work has been performed on this dataset. The approach will be first to study all the previous work. Then, the idea is to segment the dataset into buckets based on some parameters. These buckets will determine the true nature of the transactions.

Data

The dataset contains transactions made by European credit cardholders in September 2013. This dataset shows the transactions that took place in two days, where out of 284,807 transactions 492 are fraudulent transactions. The data set is unbalanced, with 0.172 % of all transactions accounted for by the positive classification (frauds).

The dataset only comprises of only numerical input variables resulting from a PCA transformation. However, the original features and more background information about the data are unavailable due to privacy concerns. Characteristics V1, V2,... V28 are the main components obtained with PCA, with 'Time' and 'Amount' being the only features not transformed with PCA. The 'Time' feature contains the seconds between the transaction and the dataset's first transaction. The feature 'Amount' is the transaction amount and is used for cost-sensitive learning, for example. Function 'Class' is the response parameter, and in case of fraud, it takes value 1 and 0 otherwise.

Thie dataset has been sourced from kaggle: https://www.kaggle.com/mlg-ulb/creditcardfraud

Project plan

The models that are to be explored are - logistic regression, clustering methods, and support vector classifiers. It is also planned to explore various methods of dealing with unbalanced datasets. The course of action is to update the work regularly on Github so that it is easier to track and also have a back up for the entire code. The target is to study the previous work by the 12th of this month and then proceed.