Project Proposal

Credit Card Fraud Detection Using Machine Learning

Team Members:

Zheng Li

Yash Shingadiya

Nan Xu

Objective:

The objective of the project is to detect all the fraudulent transactions while minimizing incorrect fraud classification using various machine-learning techniques.

Definition:

- This project includes modeling past credit card transactions with the knowledge of the ones that turned out to be fraud and then use it to identify whether a new transaction is fraudulent or not.
- We have decided to use the credit card fraud detection dataset available on Kaggle. The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, consisting of 492 frauds out of 284,807 transactions.
- The dataset contains only numerical input variables which are the result of a PCA transformation. Due to confidentiality issues, the original features and more background information about the data are not provided. Features V1, V2, through V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Preliminary plan:

Creating a training set:

- As the dataset is highly unbalanced, first we need to create a training set with balanced class distribution that will allow the algorithms to detect fraudulent transactions.

Outlier detection and removal:

- We need to detect and remove those anomalies which can impact the predictions and can skew the results, by identifying the correlation between the features.

Dimensionality reduction:

- For visualization, we aim to reduce high dimensional data to a low dimensional data space using t-SNE technique.

Implementing various algorithms and analyzing the results:

- We intend to implement three algorithms to solve this problem namely Logistic Regression, Support Vector Machine and Random Forest and finally, compare their results.

Evaluating the success:

- We will be using F1 score and ROC-AUC as performance metrics.

Creating a final project report:

- Finally, we intend to analyze the results obtained using different algorithms and create a final project report out of it. This will be our final step towards the completion of the project.

Distribution of work:

- Zheng is going to work on the creation of training set and other members will help him with that, if needed. Apart from that, he is mainly going to focus on Logistic Regression algorithm.
- Yash is mainly going to focus on Random Forest algorithm, t-SNE technique and ROC-AUC evaluation technique.
- Nan is mainly going to work on Support Vector Machine algorithm, outlier detection and removal and F1 score evaluation technique.
- Once we learn all these concepts we all will sit together, discuss the concepts, and implement them and will help each other whenever needed.

Papers that we will go through:

John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis, IEEE, 2017

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015

L.J.P. van der Maaten and G.E. Hinton, Visualizing High-Dimensional Data Using t-SNE (2014), Journal of Machine Learning Research