CSC 5825 Fall 2021 Term Project

Students enrolled in CSC 5825 are required to complete the term project (10% weight). The topic for this year is Credit Card Fraud Detection.

Project report due: **Dec. 20 midnight**

Problem Statement

It is important that credit card companies can recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

Dataset

The data set is published at: https://www.kaggle.com/mlg-ulb/creditcardfraud

The datasets contain transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly imbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … 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, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable, and it takes value 1 in case of fraud and 0 otherwise.

Method

Since the labels are imbalanced, so it is not right to apply binary classification in a straightforward manner, think about why for both discriminative and generative classifiers? Available methods include **supervised** imbalanced binary classification and **unsupervised** anomaly detection. **You are required to use at least one method from each category.** For method evaluation and comparison, **please report not only accuracy, but also F-1 score**. If the model tuning is needed, then Area Under Precision Recall Curve (AUPRC) is preferred.

Proposal

- Preprocess the data set and formulate problem.
- Split the data set into training set (70%), validation set (10%) and test set (20%). Think about how to make the data split for both supervised and unsupervised methods.
- You can use models form scikit-learn package.

Rubrics

- Data preprocessing and visualization. (10 points)
- Classification (and anomaly detection) methods. (50 points)
- Results and evaluation (accuracy and F- score). (30 points)
- Conclusion. (10 points)

Submission

For this term project, you must submit your source code together with a 2-page written project report following the specified format below, which includes your results, evaluation, and your conclusion. To earn full credit, I expect to see solid reasoning, analysis, and results to sustain your claims with a clear conclusion.

Format:

1. Please write your report either in MS WORD A4 format or LaTeX format. Please use the font size 11. Please submit your report in **PDF format**.

2. Page limit: 2 full pages (**excluding** bibliography and source code), no longer, no shorter. The font size should be **no larger than 12** point. Please include no more than 2 figures and 1 table to justify your analysis and report results.

3. Organize your report according to the following sections:

Section 1. Background/Introduction

Section 2. Methods

Section 3. Experiment, Results and Discussion

Section 4. Conclusion

Bibliography