Credit Card Fraud Detection

Final project

Qinxiao Shi & Sen Yang

13 October 2018

1 Overview

https://www.kaggle.com/mlg-ulb/creditcardfraud It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

2 Dataset Description

1. This dataset presents transactions that occurred in two days, where the record has 492 frauds out of 284,807 transactions.
2. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
3. The original features and more background information about the data are not provided due to confidentiality issues.
4. Features V1, V2, ... V28 are the principal components obtained with PCA transactions, the only features which have not been transformed with PCA are 'Time' and 'Amount'.
    1) Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset.
    2) The feature 'Amount' is the transaction amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

3 Goal

Identify fraudulent credit card transactions.

4 Method

Measure the accuracy using the Area Under the Precision-Recall Curve (AUPRC), and get the point where AUPRC is the largest. The method for determining largest AUPRC is EM algorithm.

5 Reference

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