Li, DING
8 May 2018

A Machine Learning Model for Credit Card Fraud Detection

Domain Background

In recently years, credit card are widely used by customers in many countries. The credit card issuer, banks, issued the card as the customers promise, borrow money to them, and expected the money been returned in a given period. With the market growing, some customers intended to cheat bank in some way. The situation is getting worse when more and more customers to apply credit card and the cheater use sophisticated and diverse ways to gain advantages. This calls for a solid way to detected fraud actions and against malicious activities even before it happen.

I choose this domain for my capstone project, because it is a really large market, and I am quite interested about it.

Problem Statement

Banks have valuable information on how customer use credit card, and those information can be analysis to predict the fraud.

While traditional machine learning classification technicals have proven to be efficient at balanced data, they may not very good at the outlier detection, which positive samples are greatly outnumbers the negative samples.

Dataset and input

This project will use Kaggle Credit Card DataSet(https://www.kaggle.com/mlg-ulb/creditcardfraud/data) to build models and detect anomalies. Here are some extracts from the data introduction:

The datasets contains 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 unbalanced, 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-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

Solution Statement

The goal of this project is to train a model that can works well under unbalanced data.

Most machine learning classification algorithms work well on balanced data, but may be proved work useless on unbalanced data. One potential solution would be increased the small number data size.

Benchmark Model

The model trained on original data will be taken as benchmark model, compare with the data be balanced.

Evaluation Metrics

Accuracy and False Alarm Rate (FAR) are used here for evaluation metrics.

Accuracy = (TP + TN)/(TP + TN +FP + FN)

FAR = (False Positive Rate(FPR) + False Negative Rate(FNR)) /2

FPR = FP / (FP + TN)

FNR = FN / (FN + TP)

Project Design

The project follows the typical machine learning flow:

1. Explore the dataset
2. Preprocess the dataset
    - 2.1 Identify features and apply transformation with log function
    - 2,2 Increase the negative samples to balance the data
    - 2.3 Perform the minimal scaling and one hot encoding on dummy variables
    - 2.4 Clean outlier data
    - 2.5 Feature selection or PCA
3. Train different models with:
    - 3.1 Neural Network
    - 3.2 Several ML classification models
4. Compare the result with benchmarks
5. Fine-tune the parameters to improve the accuracy.
6. Analysis the confusion matrix and suggest the further improvement.

Citation

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