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.

I would like to build a classification machine learning model that learned from history fraud data and then predict the possible fraud in the future.

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 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-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

Machine learning models work well on almost all data in real world, except when the data is imbalanced. **So the goal of this project is to train an effective model that works well on credit card data**, which is highly unbalanced (492 frauds out of 284,807 transactions, that is the positive class account for 0.172% of all transactions), **and compare the efficiency on different models.**

The possible solution are outlier detection and imbalanced learn.

Outlier detection assume the same distribution as existing observations (it is an inlier), and considered as different if not(it is an outlier). It has four models in sk-learn, which are elliptic envelope, Isolation Forest, Local Outlier Factor and One-class SVM.

Imbalanced learn is a way to under/over sampling data, and feed into normal machine learning models.

Benchmark Model

The benchmark model is the random selected model.

Evaluation Metrics

Area Under the Precision-Recall Curve (AUPRC) is used here for evaluation metrics.

Precision-Recall Curve are precision-recall pairs for different probability thresholds.

The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives.

The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives.

Area Under the Precision-Recall Curve is the area under precision recall curve.

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 Perform the minimal scaling and one hot encoding on dummy variables
    - 2.3 Feature selection
3. Train different models with:
    - 3.1 Neural Network on balanced data
    - 3.2 Several traditional machine learning classification models on balanced data
    - 3.3 Models from outlier detection
4. Compare the result with benchmarks
5. Fine-tune the parameters to improve the accuracy.
6. Analysis the result and suggest 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