Machine Learning Engineer Nanodegree

Capstone Proposal

Frank Albrecht

March 13, 2018

Proposal

Domain Background

Every day millions of dollars are lost due to credit card fraud. To reduce these losses, there is a desire to develop effective detection methodologies. This is where machine learning can help and provide a solution to help detect fraudulent transactions as they occur.

The design of fraud detection algorithms is not without challenges however. Developers are confronted with datasets that are:

- of highly unbalanced nature (ratio between non fraudulent transactions and fraudulent transactions is very high),
- very dynamic (customer behavior and fraud patterns change all the time),
- anonymized, thus not allowing for proper feature analyses and reduction,
- limited by low feedback that characterizes the transactions.

These challenges need to be addressed and solved for an algorithm to be effective and adaptable.

With this project I want to pursue the use of Auto-Encoders (3-6) for credit card fraud detection. Credit card fraud presents a problem set that can apply to many datasets in different industries. As such, I do have a personal interest in investigating this solution.

The main academic research (1) on this subject, I came across and prompted my choice and selection of problem, is the Phd. Thesis by

Andrea Dal Pozzolo on “Adaptive Machine Learning for Credit Card Fraud Detection “.

Problem Statement

Every day banks are confronted with millions of credit card transactions. It is up to them to determine if these transactions are legitimate or fraudulent, and they need to do so with very high accuracy.

Essentially we are dealing with a 2-class classification problem: is the transaction genuine or fraud.

How successful this detection process works is reflected in the positive detection of fraudulent transactions, thus the reduction of fraud cases and the losses incurred.

Datasets and Inputs

The dataset (2) used for analyses and the development of the proposed solution is an anonymized set of credit card transactions labelled as fraudulent or genuine. It contains transactions made on 2 days in September 2013 by European cardholders and consists of 492 fraud cases out of 284807 transactions. The dataset is highly unbalanced, the positive class (fraud) account for only 0.172% of all transactions.

The dataset has undergone a PCA Transformation and contains only numerical variables. This was done for confidentiality reasons, thus the original features and background information are unavailable. The data consists of 31 features:

- Features V1, V2, ... V28 are the principal components obtained with PCA.
- Feature 'Time': contains the seconds elapsed between each transaction and the first transaction in the dataset.
- Feature 'Amount' is the transaction Amount.
- Feature 'Class' is the response variable and takes value 1 in case of fraud and 0 otherwise.

The dataset has been collected and analyzed 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. More details on current and past projects on related topics are available on:

- http://mlg.ulb.ac.be/BruFence and
- http://mlg.ulb.ac.be/ARTML

Solution Statement

To classify credit card transactions as Fraud or genuine, I will develop a prediction model based on Auto-Encoder neural network. This model will be compared to the benchmark model via the defined evaluation metric Area under the Precision-Recall Curve (AUCPR).

The interesting aspect of Auto-Encoders is that they try to predict the input given the same input. Auto-Encoders try to learn to approximate the following identity function:

$$f_{W,b}(x) \approx x$$

Due to the anonymization of the features provided in the dataset (dataset has already undergone a PCA), no further reductions are foreseen during this project.

Benchmark Model

As a benchmark model I'll be using a "One Class SVM" classifier. This classifier should lend itself to identifying anomalies and will provide a baseline performance to compare against the chosen Auto-Encoder solution.

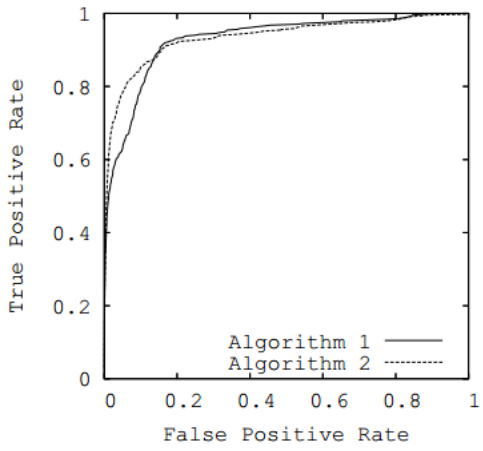
Accuracy, confusion matrix and the Area under the Precision-Recall Curve (AUPRC) can be used as measures for the success of the model. The main metric, AUPRC (see below in "Evaluation Metric"), is chosen as it is a more appropriate measure to evaluate the model considering the imbalanced nature of the dataset, while accuracy is given for reference.

Evaluation Metrics

The metric proposed (7,8) to evaluate the models (benchmark and solution) is the Area under the Precision-Recall Curve as opposed to the Area under the ROC Curve.

The main reason for this is due to the imbalance in the dataset, the ratio between True Positives (TP) and True Negatives (TN): out of 284807 total samples only 492 are classified as TN. In this situation using the ROC curves (TPR vs. FPR) can be misleading when comparing models.

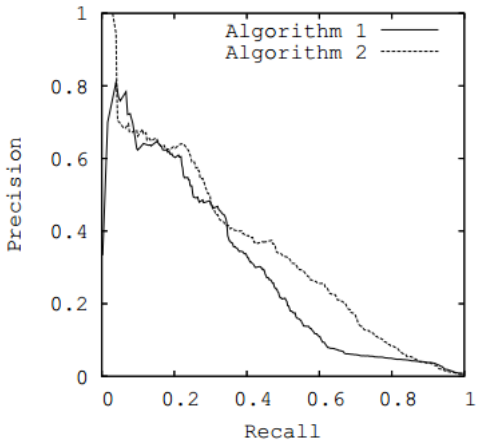
ROC Curve: Plots the True Positive Rate (TPR) against the False Positive Rate (FPR)



$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

PR Curve: Plots Precision against Recall:



$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

From the above formulas we can see that for the ROC curve, when we have a high imbalance and for example a high number of TN, FPR becomes very small. When evaluating 2 models the delta between 2 very small FPRs is also a very small number. Therefore differences between 2 models become less discernible. This characteristic is not present when using the PR curve. The PR Curve therefore lends itself better to datasets that are highly imbalanced like in this case.

When comparing the models, the area under the PR curves is calculated for each and compared: the larger the area under the curve the better the model.

Project Design

To Gain a basic understanding of the dataset the application of descriptive statistics is going to be used. After these basic calculation I will proceed with some data exploration to understand how each feature may contribute to the classification.

Since a PCA has already been applied to the data and the nature of each feature is not known (data is anonymized), data exploration may lead to some insights, but further feature reduction may not be applied due to concerns that the wrong feature(s) may be eliminated, as the actual nature and name of the features is unknown.

Next, outlier detection will be performed and standardization applied to bring all features on the same scale.

Since the dataset is highly imbalanced and it is very hard to learn from, a balancing technique will be applied using the under-sampling. This technique randomly removes samples from the majority class.

Next, the benchmark model and the solution will be developed and applied to the dataset. These models are then compared for performance using the evaluation metric.

The expectation is that the proposed solution using auto-encoder will yield the best result and be the most effective in predicting fraudulent trnasactions.

References:

1. “Adaptive Machine Learning for Credit Card Fraud Detection “, Andrea Dal Pozzolo
2. https://www.kaggle.com/mlg-ulb/creditcardfraud
3. https://medium.com/@curiousily/credit-card-fraud-detection-using-autoencoders-in-keras-tensorflow-for-hackers-part-vii-20e0c85301bd
4. http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/

5. https://github.com/llSourcell/anomaly_detection_for_CERN/blob/master/Credit%20Card%20Fraud%20Detection.ipynb
6. https://github.com/curiousily/Credit-Card-Fraud-Detection-using-Autoencoders-in-Keras/blob/master/fraud_detection.ipynb
7. http://pages.cs.wisc.edu/~boyd/aucpr_final.pdf
8. http://www.chioka.in/differences-between-roc-auc-and-pr-auc/