Zusammenfassung

Die korrekte Vorhersage von betrügerischen Kreditkartentransaktionen oder ausfallenden Krediten verringert die Verluste der Banken durch Kreditkartenbetrug oder nicht korrekt bezahlte Kredite. Aktuell erfreut sich Deep Learning großer Beliebtheit, welches zwar häufig eine gute Performance besitzt, aber keine Erklärung des Zustandekommens der Ergebnisse liefert. In einigen Bereichen wie im Finanzsektor ist die Erklärbarkeit der Modelle eine Voraussetzung für deren Einsatz.

In dieser Masterthesis wird ein Prozess vorgestellt, wie mithilfe der topologischen Datenanalyse in Form des Mapper Graphs die Performance der Klassifikation durch Entscheidungsbäume und die logistische Regression verbessert werden kann. Ein Ziel der Masterthesis ist das Design und die Implementierung des vorgestellten Prozesses.

Der vorgestellte Prozess wird auf den Kreditkartentransaktionsdatensatz von Kaggle, auf den German Credit Datensatz und auf einen Datensatz der Peer-to-Peer Kreditplattform Lending Club angewandt. Auf dem Kreditkartentransaktionsdatensatz ist die Performance der logistischen Regression um rund 6 Prozentpunkte angestiegen, die Performance des Entscheidungsbaumklassifikators erhöht sich um rund 5 Prozentpunkte. Der vorgestellte Prozess führt auf dem German Credit Datensatz zu einem Anstieg des AUPRC des Entscheidungsbaumklassifikators um etwa 14 Prozentpunkte. Auf dem Lending Club Datensatz ist ein Anstieg der Performance des Entscheidungsbaumklassifikators um 2 Prozentpunkte zu verzeichnen. Die Durchführung des Prozesses verursacht nicht in jedem Fall eine Zunahme des Performanceindexes AUPRC.

Insgesamt haben die Experimente gezeigt, dass der vorgestellte Prozess eine Möglichkeit darstellt, die Performance von einfacheren Klassifikatoren anzuheben. Daher lohnt es sich, die Einsatzmöglichkeiten der topologischen Datenanalyse auf dem eigenen Datenbestand zu evaluieren.

Abstract

The accurate prediction of credit card fraud or loan default helps banks to reduce the losses incurred by fraudulent transactions or loan default. Currently deep learning models are applied frequently due to their high accuracy. A major drawback of these models is that the outcome of these models are not explainable and interpretable. For some applications, including the finance industry, the explainability of models is key.

This master thesis presents a process based on the generation of the mapper graph to increase the performance of decision tree models and logistic regression. The mapper graph is an approach of topological data analysis and gives a summary of the shape of the data. The design and implementation of this process is also comprised in this master thesis.

The presented process is applied to a credit card fraud data set, the German credit data set and a dataset of all loans of Lending Club, which is an American peer to peer lending platform. On the credit card transactions dataset, the performance of the logistic regression increases by approximately 6 percentage points, the performance of the decision tree classifier is increased by approximately 5 percentage points. On the German credit data set, the presented process leads to an improvement of the AUPRC performance measure of the decision tree classifier by 14 percentage points. On the Lending Club dataset, the performance is increased by 2 percentage points on the decision tree classifier. But not all experiments show an increase of the performance by using the process presented in this thesis.

In total, the experiments are showing that the process, which is presented in this thesis, is able to increase classification performance on the data sets used. Therefore, it is worthwhile to check if topological data analysis leads to an improvement of the classifier on the data used.