Carl von Ossietzky Universität Oldenburg

Fraunhofer ITWM

Master Thesis „Instance ranking prediction and aggregation in fraud detection “

Instance Ranking Prediction

Abstract

When dealing with fraud detection, one often faces the challenge not only to detect fraudulent data but also to prioritize those cases for auditing. For example, a company might want to arrange potentially fraudulent cases in increasing order concerning expected claimable refund. To address this problem, we make use of instance ranking prediction. Let $D^{train} = (X, Y)$ be a training set containing regressors $X_i \in \mathrm{X} \subset \mathbb{R}^p$ with potentially large p, i.e., we may face high-dimensional data, columns $X_{.,j} \in \mathbb{R}^n$ and $Y \in \mathbb{R}^{n\,x\,K}$ for $K \geq 1$.

Scenario a)
Let $K = 1$. We apply an algorithm that performs model selection tailored to the ranking loss function

$$L_n^{hard}(\theta) = \frac{1}{n(n-1)} \sum \sum_{j \neq i} I\left((Y_i - Y_j)\left(\hat{Y}_i - \hat{Y}_j\right) < 0\right), \hat{Y}_i = f_\theta(X_i),$$

e.g., the gradient-free gradient boosting algorithm from the R-package gfboost [Gradient-Free Gradient Boosting [1]], in order to get a sparse model whose predictions induce an ordering of the instances.

Scenario b)
Let $\hat{\theta}$ be the estimated sparse coefficient. Since this coefficient has been estimated on one single data set, we would like to stabilize it. We do so by drawing B subsamples from our training data set. The estimated coefficients

[1] T. Werner. Gradient-Free Gradient Boosting. PhD thesis, Carl von Ossietzky Universität, Oldenburg, 2019.

10.06.2021

Carl von Ossietzky Universität Oldenburg

Fraunhofer ITWM

$\hat{\theta}^b, b = 1, \ldots, B$, of those subsamples are then aggregated using suitable techniques as described in [Model-based Boosting in R [2]] or [1].

These basis scenarios should be extended to the case $K > 1$.

This thesis will be a corporation of your supervising university, Fraunhofer ITWM and the Institute for Mathematics of Carl von Ossietzky University Oldenburg.

Demands on Applicants

Programming skills in R

Interest in learning new data science and machine learning methods

Familiarity with basic mathematical-statistical concepts

Contact

Dr. Stefanie Grimm
Methodcoordination Data Science
Department Financial Mathematics/Fraunhofer ITWM
0631 31600 4040
stefanie.grimm@itwm.fraunhofer.de

[2] B. Hofner, A. Mayr, N. Robinzonov, and M. Schmid. Model-based Boosting in R: A Handson; Tutorial Using the R Package mboost. Computational Statistics, 29(1-2):3–35, 2014.

10.06.2021