Prediction Of Patient-level Outcomes In The Renal Context: Acute Kidney Injury In Cardiac Surgery Patients

Frederic Schneider

Introduction

Cardiorenal syndrome (CRS) is a medical term used to describe the phenomenon of injury or stress on the heart or kidney causing chronic dysfunction or acute injury of the other organ. As a result heart patients have an increased risk of *acute kidney injury* (AKI), a condition associated with complications and poor patient outcomes.

Previous work on this topic is primarily focused on the detection of AKI in patients in the ICU. Identifying heart patients who are likely to present post-operative AKI before a prospective surgical intervention can allow physicians and medical care professionals to better assess the risk of the surgery and possibly amploy kidney-protective measures.

We describe two clinical prediction models (CPM) based on clinical care data of heart patients and compare their ability to asses the risk for AKI after surgery. Many state-of-the-art machine learning techniques offer no explanation for their results, potentially hindering the adoption of CPMs in medical domain. Therefor we also discuss the intelligibility of the employed models and employ local interpretable model-agnostic explanations (LIME) [5].

Methods

Data used for model training and validation has been obtained from the MIMIC-III critical care database and comprises approximately 6700 relevant heart patients [3]. The feature extraction pipeline encompasses feature selection, imputation, and normalization, and the featurization of the laboratory results' temporal context.

While both models are trained and validated on the same data, they differ with regard to the employed machine learning techniques. We train decision trees (DT) and gradient-boosted decision trees (GBDT) on labeled training data, optimizing hyperparameters for each model using grid search.

Prediction performance of the developed CPM's is evaluated using 5-fold cross-validation and calculating the area under receiver operatinv curve (AUROC) among other common classifier performance measures [1].

In contrast to results from the DT model, GBDT predictions are not easily interpretable. We therefor employ LIME on classification results for specific patients from the CPM. It provides explanations in the form of the the relevant features for the prediction result by approximating the CPM's behaviour locally for the patients feature vector.

While the initial exploration of the employed prediction techniques and feature extraction were conducted using RapidMiner, the data extraction and preprocessing pipeline, optimized prediction models, and result exeplanation using LIME were implemented as a proof-of-concept application using the Python programming language and the scikit-learn library [2] [4].

Preliminary Results

Compared by the AUROC score the GBDT prediction model outperforms the DT's predictions by a margin of almost 30 %. At 90 % accuracy both models achieve similar class recall for the pathological class - 38 % for DT and 41 % for GBDT - but Figure 1 demonstrates the significant difference between both models. While the DT has strictly binary class output, the GBDT model can be adjusted via scoring threshold. It can be adjusted to display a 80 % class recall for the pathological class when trading off precision and accepting a false positive rate (FPR) of 0.2.

However, understanding the reason behind the CPM's predictions is vital for the willingness by medical professionals to use CPM's in the clinical environment. We can show that, using LIME, we can regain some meta-information about the prediction results decisive features, which is traded off for a more accurate prediction by using a non-interpretable CPM.

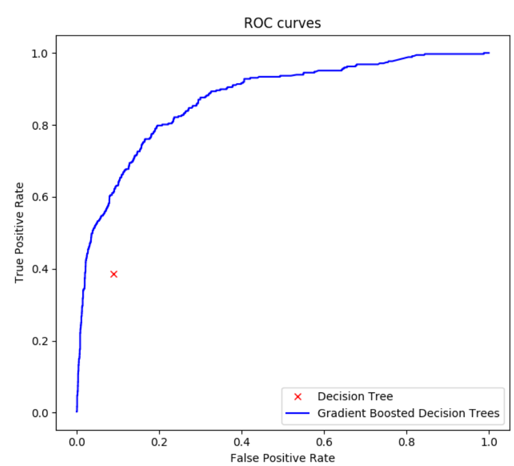

Figure 1: Receiver operating curve and point for GBDT and DT prediction model respectively

References

[1] Viv Bewick, Liz Cheek, and Jonathan Ball. Statistics review 13: Receiver operating characteristic curves. *Critical Care*, 8(6):508, Nov 2004.

[2] RapidMiner Inc. Rapidminer official website, 2017. Accessed: 2018-01-30.

[3] Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3, 2016.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.