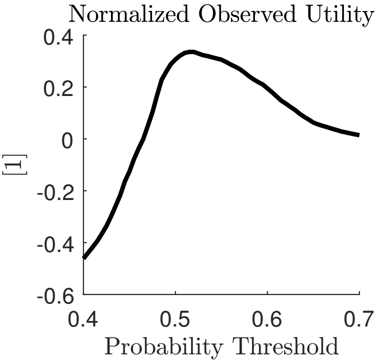
Early Prediction of Sepsis using Random Forest Classification for Imbalanced Clinical Data

Simon Lyra, Steffen Leonhardt, Christoph Hoog Antink

Medical Information Technology, Helmholtz-Institute for Biomedical Engineering, RWTH Aachen University, Aachen, Germany

This year, the objective of the CinC challenge was an early prediction of sepsis from ICU patient data sourced from three different hospital systems. The approach presented in this paper is based on a Random Forest Classifier. Several data analysis techniques to pre-process the data before training the classifier were explored. Since the data provided for the challenge was both fragmentary for many lab parameters and imbalanced regarding to the number of actual sepsis patients, interpolation algorithms and synthetic minority oversampling techniques (SMOTEs) were explored. While the interpolation of missing data was done immediately after loading the features, the SMOTE algorithm was applied after separation of the datasets in training and validation folds for proper cross validation. For implementation, Python and MATLAB were explored.

In the final version, the TreeBagger as implemented in MATLAB was used. While training the classifier, each time-step was treated individually, i.e. no information from past or future was used. To compensate for the imbalance, $N_0 = 20$ trees were trained using all available positive data and an equal amount of randomly selected negative instances in each step. The process was repeated 35 times until all negative samples were considered, resulting in a forest with $N = 700$ trees. After the prediction step, a median filter of width 5 was applied to the output of the classifier, in effect introducing information from 2 time steps of the past and 2 time steps from the future into the final prediction, leading to minimal improvement of the final result.



Using 10-fold cross validation on the public dataset, the following figures of merit were obtained for a probability threshold of 0.52: A Normalized Observed Utility of 0.335, an f-Measure of 0.125, an AUROC of 0.775, and an AUPRC of 0.078. On the hidden test-set, a Normalized Observed Utility of 0.339 was achieved.