Data imputation & Data correction in Intensive Care Physiological Data

(Master Thesis)

ALEXANDER KRUSCHEWSKY

Motivation

The SMITH project is an effort to provide interdisciplinary communication and research to sustainably improve patient care in Germany. Part of this project is the algorithmic surveillance of intensive care patients (*ASIC*), especially patients, which developed an Acute Respiratory Distress Syndrome (*ARDS*). However, since some important parameter like paO2 are only present infrequently inside of the datasets and these may contain sensor- or operation-related errors, training such surveillance algorithms yields to be difficult. SMITH also aims to provide secondary data sets to be used in research like ASIC, which requires them to maintain a high quality-standard for the data.

All this motivates the development and implementation of methods that can detect and correct errors inside of the data to yield higher training-data quality and impute data for higher temporal resolution.

State of the Art

There already exist various research regarding data imputation (like MissForest or GAIN), including methods specialized on medical data. However, most of those publications focus on imputing Missing-At-Random Values. This differs from this case, where missing data distribution would be even, due to wanting to increase temporal resolution. Additionally, there is only little research present in regard to using prior medical knowledge like the correlation between paO2 and SaO2 to restrict and possibly improve imputation results.

Finally, regarding error-detection not much public research can be found so far, with most only covering basic ideas and error types. A previous master thesis addressed this problem implementing a data analysis software using a novelty detection algorithm, which will serve as a starting point for this topic.

Objective

The overall objective of this master thesis is to improve the usability and quality of the secondary datasets, which are to be used as training data for future algorithm and model development.

This will be achieved by the development of error-detection and imputation methods as well as their evaluation and integration into the previously mentioned data analysis software.

Procedure

After an initial literature research, the project is to commence with data exploration to get an overview over the data set. Following this, various imputation techniques found in literature like MICE, MissForest, GAIN etc. will be implemented, tuned and compared. Subsequently, some additional techniques with a-priori medical knowledge will be developed and juxtaposed. Finally, using the priorly evaluated imputation techniques, possible improvements to the novelty detection algorithm will be assessed and error-handling techniques implemented and evaluated.

Betreuer: Simon Fonck
11.06.2021

Informatik 11
Embedded Software

RWTH AACHEN UNIVERSITY