Neural networks and ensemble methods for imputation of inconsistent intensive care data

(Master thesis)

XUAN THANH NGUYEN

Motivation

Data-driven research and technologies have the potential to improve medical care significantly. Therefore, more than 300 researchers and scientists collaborate to enhance the quality and availability of medical data as part of the SMITH consortium. Within the clinical use case ASIC (Algorithmic Surveillance in Intensive Care), technologies that help detecting and treating Acute Respiratory Distress Syndrome (ARDS) are developed. These technologies heavily rely on good data quality and density.

However, since medical data is prone to be incomplete and noisy, algorithms that improve data quality and density are employed. A common method is data imputation. Especially more sophisticated imputation methods that use neural networks, or ensemble learning can boost accuracy of data reliant programs.

State of the Art

A lot of imputation algorithms have been developed already. The most common and basic approaches are Complete-Case-Analysis, mean imputation, interpolation and the K-nearest neighbors algorithm. Moreover, Multiple Imputation with Chained Equation (MICE) generally has set the gold standard for imputation. However, MICE and the basic approaches perform worse when dealing with medical data due to their temporal property. In order to deal with temporal data, researchers make use of Recurrent Neural Networks and Graph Neural Networks. Other researchers proposed Generative Adversarial Networks and Ensemble Learning which combine various Machine Learning algorithms.

Objective

The objective of the thesis is to reimplement and adjust state of the art imputation methods and integrate them into the analysis tool developed at the chair. The goal is ultimately to increase the accuracy of detection and treatment algorithms developed within the use case ASIC. The focus specifically lies on methods that use neural networks and ensemble learning. Furthermore, adjustments to the existing methods should be designed to improve handling of noisy data.

Procedure

Before suitable imputation algorithms for our use case can be chosen, it is necessary to get an overview of the data. After that follows further literature research on imputation methods, evaluation methods and potential databases for training. Then several machine learning algorithms will be implemented and tuned. They will be compared to each other and to baseline methods. The best method coming out of the comparison will be further tweaked and lastly, integrated into the analysis tool.

Betreuer: Alexander Kruschewsky
23.06.2022

Informatik 11
Embedded Software
RWTH AACHEN UNIVERSITY