Method for the Artificial Generation of Error Annotations on Time Series Data

(Master Thesis)

LINA GÖNNHEIMER

Motivation

The examination of intensive care data is a crucial aspect of medical research, yet the manual process can be laborious. A prior master's thesis focused on isolating anomalies within the dataset—data points that deviate from established patterns. These anomalies, often arising from sensor issues, may also indicate underlying medical concerns, underscoring the importance of their classification. Anomalies that are caused by sensor issues or anything that is not of physiological origin are defined as errors and need to be filtered out since they provide no information on the patient's condition and can be misleading if wrongly interpreted. An algorithm proficient in this task can help generate high quality labeled data that can be used by researchers.

State of the Art

Time series classification presents a challenge for which an array of algorithms has been developed. These encompass diverse methodologies such as distance-based approaches measuring overall similarity, dictionary approaches classifying based on predefined patterns, interval-based approaches extracting features from key sub-sequences, and notably, deep learning methodologies, among others. Deep learning methods for time series classification can be further divided into two categories: generative and discriminative models. Generative models create a model of the data before applying a classifier while discriminative models directly output a probability distribution over the class labels by learning a mapping of the raw input data.

In previous theses at the chair deep learning algorithms for anomaly detection were already evaluated and an active learning framework to optimize data annotation was developed.

Objective

The objective of this thesis is to develop a process to label errors in time series data, building on a previous thesis that dealt with anomaly detection. A statistical comparison will be made between real data and the annotated data generated by this process. Also, an anomaly detection algorithm will be trained on the generated data and its performance will be evaluated.

Planned Procedure

The first step will be a literature review on time series classification methods to find out which work best on intensive care data. Subsequently the most suitable method will be implemented and used to generate artificially annotated data. The artificial data will be statistically compared to real data and the performance of an error detection algorithm that is trained on the artificially annotated data will be analyzed in comparison to being trained on real annotated data.

Betreuer: Alexander Kruschewsky
12.02.2024

Informatik 11
Embedded Software

RWTH AACHEN UNIVERSITY