Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt WS 2022/23

CAST: Classifying Time Series Anomalies

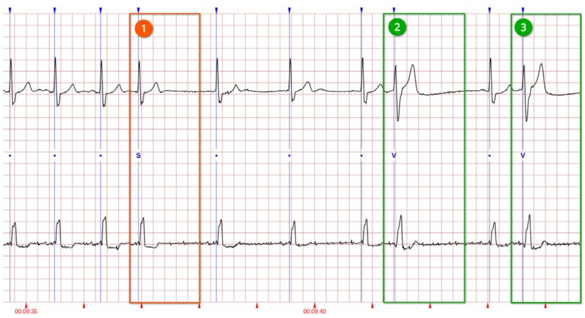
Keywords: multivariate time series, classification, anomaly detection, ML

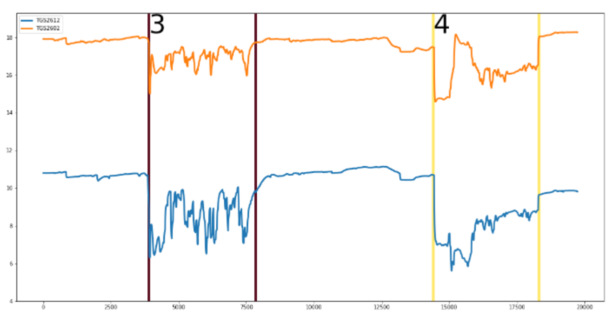
Time Series Anomaly Detection and Classification

Many real-world time series are generated sensor measurements of air craft turbines, car motors, industrial production processes, or health tracking devices. Aircraft turbines and car motors have multiple sensors that record temperature, acceleration, and pressure. Health trackers provide information about heart rate, blood pressure, or oxygen saturation. Depending on the domain of a time series, important events, such as heart failures in health tracking [1], structural defects in aircraft turbines [2], or ecosystem disturbances in earth sciences [3], are represented as anomalous subsequences in the time series. Detecting such anomalous subsequences and assigning them to known failure classes, such as supraventricular or premature heart beats in health tracking or broken sensors, bearing failures, or bending moments in aircraft turbines, is of central interest of the data owners to deduce actionable insights.

While time series anomaly detection (TSAD) [4] and time series classification (TSC) [5] algorithms are well studied research areas, TSAD algorithms can only flag anomalous subsequences within a time series, but they cannot distinguish between different anomaly classes or types; and TSC algorithms assign entire time series to anomaly classes, but cannot tell the position of a detected anomaly within the time series. For this reason, we need a combined approach that detects anomalous subsequences within a certain time series and can assign the detected anomalies to known failure classes. We call this task ***C**lassifying **A**nomalous **S**ubsequences of **T**ime-Series data* (CAST).

For example, in ECG data a single patient can experience various heart conditions at the same time (left image), while in gas sensor data different objects emit different gas concentrations (right image). A TSAD algorithm can find the changes in the signal, while a TSC algorithm could assign a general class to each time series. However, we are looking for a CAST algorithm that not only detects anomalous subsequences, but it also labels each occasion with its category, e.g. S-beat (1) vs. V-beat (2) in ECG data or wine (3) vs. banana (4) in the gas sensor data:





Challenges

An algorithm that can detect and label anomalous subsequences within a time series must overcome the following challenges:

1. **Imbalanced**: Target classes for the classification task are imbalanced, because anomalies within the time series are rare compared to the normal patterns. This requires special treatment when using a classifier [6].

2. **Large-scale**: In order to find enough anomalous subsequences, many large time series have to be processed. A CAST algorithm has to efficiently deal with these large datasets to produce reliable compact outputs.
3. **Conflicting objectives**: A CAST algorithm must balance between multiple conflicting goals: low false negative rate to catch all anomalies vs. low false positive rate to reduce human effort, and high quality results vs. low resource consumption. The conflicting goals make choosing a good evaluation metric hard.
4. **Variable classification input**: Most classification algorithms require fixed-length, non-temporal data structures as input. The right way to transform raw time series data into the required classification input format is unknown.

Project Goal

In this master project, we will follow an *entire research cycle* from problem inception and literature research to algorithm development and, finally, to evaluation. Together, we will prepare a research article that is ready for submission to an international conference.

We start the project with a literature search and data exploration phase. We provide multiple labeled datasets from different domains, but are open for additional recommendations. Afterwards, we decide on and implement baseline approaches to classify subsequence anomalies. One way to build a naive baseline is sliding a fixed-size window over the time series to extract subsequences and feeding them into a selected TSC algorithm to find their class memberships. Based on existing TSAD and TSC approaches, we will design and develop a novel CAST algorithm, which we will evaluate against the baseline approaches. We will look at both quality and efficiency of the approaches and deduct additional experiments showing strengths and weaknesses of the developed algorithm.

Prerequisites

- Interest in working with time series data and data analytics
- No fear of a bit of math and statistics ;)

Experience in the following is beneficial:

- Time series analytics (such as anomaly detection, forecasting, etc.)
- Working with Python and its data science ecosystem

pandas

Contact Details

This project will be supervised by Sebastian Schmidl, Phillip Wenig, and Prof. Dr. Felix Naumann at the Information Systems chair. If you have any questions, please do not hesitate to contact us.