Automated Feature Engineering for Active Learning

Background

Many industries today have been collecting data for several decades. This data has often been haphazardly collected with no concrete plan of what to do with it and is often unlabelled, prohibiting the use of powerful supervised learning techniques. Because of this many industrial companies are data rich, but information poor.

At Viking Analytics AB, we strive to enable industrial engineers/experts to self-serve their data analytics needs, but for many companies this is currently unfeasible because of the lack of annotated data. For most applications labels are acquired from human experts, which is a very time consuming and monotonous process and can become very expensive when a task requires expertise from highly educated professionals. A field in machine learning called Active Learning has emerged to make the labelling process faster and more efficient. This is achieved by providing an expert with the most informative instances to annotate, proven to decrease the number of labels required to achieve equal performance manifold[1].

Some of the most abundant industrial data comes in the form of vibration and other types of time series data. This data is unfortunately not a perfect fit for many Active Learning methods without first extracting higher level features, often referred to as feature engineering. The process is often application specific and requires knowledge of the machine learning algorithm to be used.

Goals

The goal of this thesis is to research methods of preparing unlabelled timeseries data for active learning through automated data pre-processing, to enable domain experts to utilise their data without knowledge of data science methods.

In this thesis, we will be looking into the following two approaches:

- AutoML (e.g deep feature synthesis[2] or distributed and parallel time series feature extraction[3])
- Time series imaging[4]

You will be exploring these two approaches in the first phase of the thesis. We will then together select few methods to deep dive and compare using some real world data from industry.

Requirements

- Curious, motivated, and driven!
- Good knowledge and strong interest in machine learning.
- Good programming skills specially with Python.
- Strong mathematical background is a plus.

Contact: arash.toyser@vikinganalytics.se

[1] https://pubs.acs.org/doi/10.1021/ci049810a
[2] https://dai.lids.mit.edu/wp-content/uploads/2017/10/DSAA_DSM_2015.pdf
[3] https://ui.adsabs.harvard.edu/abs/2016arXiv161007717C/abstract
[4] https://arxiv.org/pdf/1506.00327.pdf