Annotated PHM dataset

For GitHub see: https://github.com/Carolus94/Annotated_PHM_data

Background

Labelled industry datasets are the most valuable asset in prognostics and health management (PHM) research. However, creating labelled industry datasets is both difficult and expensive, making publicly available industry datasets rare at best. While labels are generally unavailable, many industry datasets contain annotations, maintenance work orders, or logbooks, with free-form text containing technical language descriptions of component properties, valuable information for any PHM model. Alas, publicly available annotated industry datasets are also scarce, in particular ones with associated signals available. Therefore, we release data from an annotated process industry dataset, consisting of 21090 pairs of signals and annotations from one year of Kraftliner production.

As data scale is crucial for artificial intelligence (AI)-based PHM, successful usage of annotated industry data is key to inspiring more companies to store and share annotated signals. We therefore encourage researchers to use this data, showcase results to companies, and urge them to collaborate in sharing more annotated data, to facilitate faster and more secure PHM for all involved industries.

Annotated signals dataset

The annotations are written, in Swedish, by on-site Swedish experts, and the signals consist of accelerometer vibration measurements from two large (80x10x10m) paper machines. The data is cleaned and structured so that each annotation is associated with ten days of signal measurements leading up to the annotation date, where one signal measurement consists of 8192 samples over 6.4 seconds, which becomes 3200 samples stretching over 500 Hz in the frequency domain. The associated annotations are attached to each signal sample, so that the list of annotations is as long as the list of signals. In total, there are 43 unique annotations, though most are associated with multiple signals from different machines due to commonalities in fault descriptions. The language data is pre-processed so that all letters are lower case, numbers are removed, and names are replaced with the Swedish word "egennamn", meaning "name of a person" in English.

Also included are pre-computed embeddings, which facilitates faster and easier testing for researchers wanting to easily investigate training signal encoders supervised through technical language supervision. The data presented here was used in the article "Technical Language Supervision for Intelligent Fault Diagnosis in Process Industry" (https://papers.phmsociety.org/index.php/ijphm/article/view/3137).

Using the annotated signals dataset

Accessing the data is simple; all you need to do to load spectra and annotation pairs is:

```python
import pandas as pd
spectra_note_df = pd.read_pickle("TL_spectra_note_df_big.pkl")
all_spectra = TL_spectra_note_df['Spectra']
all_annotations = TL_spectra_note_df['Notes']
```

Pre-computed embeddings can be accessed through:

```python
all_embeddings = TL_spectra_note_df['Embeddings']
```