FCC WEEK 2017

KICKER PULSE GENERATOR ANOMALY DETECTION FOR RELIABILITY IMPROVEMENTS THROUGH ADVANCED MACHINE LEARNING

N. Wéry†, W. Meert†, P. Zuidberg,† P. Van Trappen*, M. Barnes*

† KU Leuven - Dept. Computer Science
Celestijnenlaan 200A, 3001 Heverlee,Belgium

DTAI is supported by Research Foundation Flanders, SBO-HYMOP.

***CERN - European Laboratory for Particle Physics**
CH-1211, Geneva 23, Switzerland

Introduction

Reliability, availability and maintainability are parameters that determine if a large-scale accelerator system can be operated in a sustainable and cost effective manner. Beam transfer equipment such as kicker systems are critical components with potential significant impact on global performance of the entire machine complex. Identifying the root cause of a malfunction can be a challenging and tedious task due to the increasing complexity of such systems. Manual extraction and analysis of this information is excluded for a future collider with more systems and an even higher complexity.

The use of Artificial Intelligence (AI) models can assist in this task leveraging existing frameworks and libraries for machine learning. A collaboration between CERN and the University of Leuven (KU Leuven) was founded to conduct such a research for an existing data set. A subset of historical data, from the LHC logging database, is used and data of the LHC injection kicker magnet pulse generators has been chosen as a first case study. The goal is to apply supervised and unsupervised learning techniques from open-source libraries such as the Scikit -learn Python library to extract useful features to create a model that detects anomalies without human interaction, both on historical data and live data from the equipment. The status and outlook of this research are presented.

Research data subset

The kicker pulse generators and magnets are complex machines; sensor and measurement data is stored in the CERN logging database for operation and diagnostics. For this research the LHC Injection Kickers (MKI) were chosen as there is a lot of data and several problems happened and were understood during the selected 6 month period (April - September 2016).

Problems that can be explained are important so the CERN equipment expert can confirm anomalies as false or true while the model is being trained. The data used can be summarised as follows:

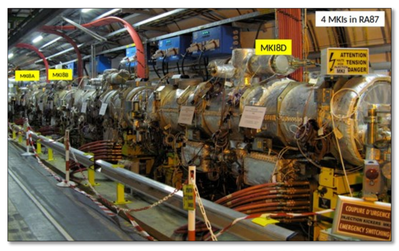

Fig. 1: MKI magnets in the LHC tunnel

- Fixed frequency sampled variables such as temperature, vacuum chamber pressure, beam intensity, slow control state, etc.
- Acquisition triggered sampling of IPOC data (Internal Post-Operational Check) such as magnet current waveforms and trigger delays.

Approach

Data extraction and sharing

After some trial and error both institutes have identified MongoDB as the most suitable NoSQL database for sharing the logged data - based on scalability, performance and memory needs. Python was used for all scripting, gratefully using free libraries such as PyTimber and Scikit-learn.

Feature Extraction

The raw values were analysed and several techniques were applied to build a set of useful features for the machine learning algorithm.

- Resampling
  All points are first resampled to one second. By doing this, every sensor has a value for every timestamp.
- Sliding Windows (fixed frequency)
  Features are generated as sliding windows. This restores temporal information between the data points.
- Trigger-based resampling
  From the fixed frequency data, only timestamps that also have a triggered acquisition measurement are kept. This discards data from when the equipment was not operating.
- Fast-Fourier Transform (FFT)
  FFT was applied to the data but it was shown that there was little useful frequency-domain patterns recognised.

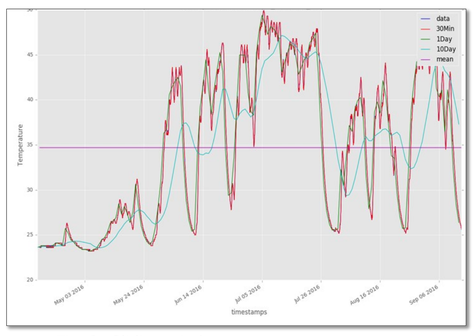

Fig. 2: Sliding windows on magnet temperature

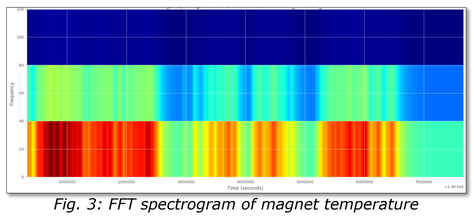

Fig. 3: FFT spectrogram of magnet temperature

Model

Two machine learning algorithms for unsupervised learning have been used to create a model that detects anomalies.

- Gaussian Mixture Model (GMM)
  A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Fitting a GMM on a multi-dimensional feature-set returns a probability density function. This is how normal behaviour and thus anomalies can be detected. Points with a high probability correspond with normal behaviour since they fit the model well, while anomalies have low probabilities.
- Principal Component Analysis (PCA)
  PCA is used to reduce the dimensionality of features by decomposing a multivariate dataset in a set of successive orthogonal components that explain a maximum amount of variance. This is done by only selecting the most important eigenvectors of the features. Most of the variance in the features can then be explained in a lower dimension and more easily visualised.

Evaluation

The probability density function from the GMM is used to return timestamps of the data-points with the lowest probability, i.e. the anomalies. These can then be cross-checked with the PCA model representation and manually analysed by CERN equipment experts for correctness.

Later a labelled dataset from manual reports (so-called e-logbook) was provided to allow for an automated evaluation. Segmentation was applied to the points with the lowest probabilities (i.e. anomalies) to tackle the fact that a single point is less likely to be an anomaly than multiple points with a low probability close together in time.

Objectives

The main objective is to develop a machine learning model to detect anomalies based on unsupervised learning. The model will then be able to detect anomalies, even if the failure mode is *new* to the model. It should be capable of handling vast amounts of data and detect unknown correlations and thus provide assistance to equipment experts when analysing problems.

Additionally by using tagged data from the CERN e-logbook, supervised learning algorithms will be used to create a model that's very accurate in recognising *known* failure modes that have been taught by the training set.

Results

GMM versus PCA

A problem with GMM is that the number of Gaussians need to be determined. The Bayesian Information Criteria (BIC) is used to find an optimum using a complexity penalty term not to overfit the data. Some absolute filtering was applied as well to discard false measurements, such as a minimum magnet current of 1kA. Significant amount of time was spent on selecting proper features from the large raw dataset provided.

A visual inspection of both models was done by plotting a certain percentile of the GMM anomalies on the dataset transformed by PCA. The pictures below show an analysis using the fixed-frequency sampled data only. In general extreme outliers were found to clearly match but discrepancies were found once analysing further. The need for a custom segmentation algorithm was identified to continue.

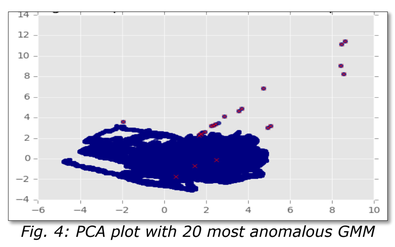

Fig. 4: PCA plot with 20 most anomalous GMM

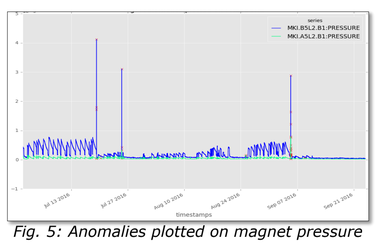

Fig. 5: Anomalies plotted on magnet pressure

Evaluation with segmentation

Segmentation is important to group anomalies together, based on the nature of the data and anomalies, and apply proper scoring to evaluate the models. Scoring is needed to fine-tune the feature extraction and GMM BIC parameters. Using the mean probability scoring can be applied.

The figures below show the result of the segmentation of the 200 points with the lowest probabilities for the two analysed MKI installations (MKI2 and MKI8). For visual reference the temperature and pressure is plotted for respectively MKI2 and MKI8. The red lines are the segment centres with its height reflecting the anomaly probability. Green are the labelled anomalies as provided by the equipment expert. The table below summarises the findings.

No. of anomalies	Type	MKI2 (beam 1)	MKI8 (beam 2)
200	Anomalies detected	4 / 9 (44 %)	3 / 7 (43 %)
	Others detected (info, fault, etc.)	9 / 52 (17 %)	14 / 52 (27 %)
	False detections	30 / 43 (70 %)	47 / 64 (73 %)
1000	Anomalies detected	5 / 9 (56 %)	5 / 7 (71 %)
	Others detected (info, fault, etc.)	17 / 52 (33 %)	21 / 52 (40 %)
	False detections	66 / 88 (75 %)	102 / 128 (80 %)

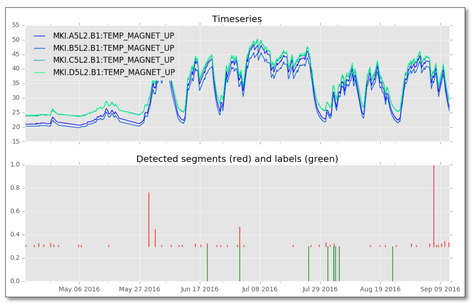

Fig. 6: MKI2 segments for 200 anomalies

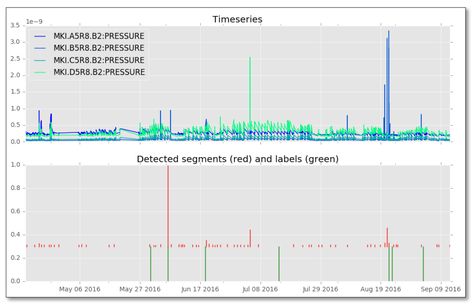

Fig. 7: MKI8 segments for 200 anomalies

Conclusion

Many anomalies are detected by the GMM model and significant effort has been done to provide proper segmentation. Although several of the labelled entries are detected, many more anomalies with a higher-or-equal probability are detected as well (i.e. false detections). The parameters should now be fine-tuned to provide optimal scoring. Because the GMM yields a single probability density function it is difficult to find which feature contributes most to the found anomaly.

Future effort will now focus on optimising these parameters and using different models and algorithms that can explain which feature contributes most to an anomaly (e.g. Density Estimation Trees). Furthermore the feasibility of supervised learning models (e.g. Random Forest Classifier) will be tested. An interesting direction for future work is to also look at semi-supervised models that combine unsupervised and supervised models.