Anomaly detection using machine learning techniques for beam injections from the SPS to the LHC at CERN

Progress Report

Marc Ferriggi

Supervisor: Dr. Gianluca Valentino

Faculty of Science
University of Malta
December 2018

Submitted in partial fulfillment of the requirements for the degree of B.Sc. (Hons.) Computing Science AND Statistics and Operations Research

Contents

Abstract

The CERN particle accelerator complex generates around 2 TB of data per week from almost 1 million signals. In this dissertation, unsupervised machine learning techniques for applications such as anomaly detection shall be used to analyse past LHC and SPS data in order to visualize correlations, determine data driven models and identify opportunities for improving the LHC machine availability and performance reach.

Introduction & Motivation

The LHC (Large Hadron Collider) is filled to flat top intensity by injecting each beam with kicker waves 12 times [1]. This is a challenging task given the high energy of the beam, the very small apertures and the delivery precision's tight tolerances, thus multiple sensors are installed around the CERN particle accelerator complex [2] which gather readings and data that can be used to check the quality of the injected beam. For this dissertation, the sensors around the injection from the SPS (Super Proton Synchrotron) to the LHC will be of particular interest. This data is stored using CERN's LS (Logging Service). The LS is heavily used and in 2013, it was noted that close to 1000 users relied on it [3]. While many studies have been made using this logged data and lots of statistical tests have been done with regards to injection quality checks for the LHC (such as [4] and [1]), no studies can be found on the CERN Document Server [5] where researchers used unsupervised machine learning methods to analyse this data.

The purpose of this FYP is to go through the entire process in a data science project and learn to use unsupervised machine learning techniques for anomaly detection. These techniques will then be used to analyse past LHC data in different machine configurations to visualise correlations, determine data-driven models and identify opportunities for improving the LHC machine availability and performance reach in terms of beam lifetime, beam stability and luminosity.

Why is the Problem non-Trivial

The LS currently generates around 2 TB of data per week. This data is generated from multiple machines each measuring different features of the waves at different points in the accelerator complex. Thus the data must be thoroughly analysed and normalised in order to be able to apply the machine learning analysis techniques properly. Identifying opportunities to improve the beam lifetime, beam stability and luminosity is also a non-trivial problem to tackle.

Background Research and Literature Review

Unsupervised Anomaly Detection Techniques

Unsupervised machine learning algorithms refer to the class of machine learning algorithms where the observations only are available to the learner as there is no access to a training set or the aim of the algorithm is simply to observe patterns in these observations. In fact, A. Hyvärinen states in [6] that for unsupervised learning "(w)e don't have separate "inputs" and "outputs", just a lot of observations of one variable or vector". Hyvärinen continues to state some goals of unsupervised learning which include data visualisation, noise reduction, feature extraction and finding interesting components, which are all of particular interest for this study.

The following points are a summary of the research made on some of the unsupervised anomaly detection algorithms that will be used in this study:

- *DBSCAN*: This algorithm was created from the necessity of having a clustering algorithm with the following requirements:

    1. "Minimal requirements of domain knowledge to determine the input parameters,"
    2. "Discovery of clusters with arbitrary shape," and
    3. "Good efficiency on large databases" [7]

    DBSCAN manages to attain the above requirements by viewing clusters as "areas of high density separated by areas of low density" [8]. This algorithm also introduces the concept of *core samples* which was then used in the designing of other machine learning algorithms such as LOF (Local Outlier Factor).

- *Local Outlier Factor*: The LOF refers to a "degree of outlier-ness" that this algorithm considers for each point in the data rather than using the concept that "being and outlier is binary" [9]. This algorithm uses a clustering technique which takes concepts from DBSCAN to measure the LOF of each point where an LOF value > 1 implies that the point has lower density than its neighbours and is thus probably an outlier.

- *PCA (Principal Component Analysis)*: This technique uses statistical and mathematical techniques to reduce the dimension of a given set of data by exploiting the correlation between the different columns. The resultant transformed data points are called principal components. It's used commonly as a first step in analysing large data sets as well as in applications such as data compression and de-noising signals [10].

- *Other Algorithms and Techniques:* Some other techniques and algorithms that will be used in this study include Recursive Feature Selection and MDA (Multilinear Discriminant Analysis).

Choosing a Python Package

Although performance of k-means and k-Nearest Neighbours is not as optimal as in other Python packages (such as '*PyMVPA*' [11] or '*shogun*' [12]), it was decided to use the '*scikit-learn*' machine learning package for this thesis due to its "state-of-the-art implementation" and "easy-to-use interface tightly integrated with the Python language" [13]. Furthermore, the algorithms implemented using this package can be "used as building blocks for approaches specific to a use case" [13] which will be useful if one would like to extend the scope of this thesis.

Methods and Techniques Used or Planned

The Parameters Studied

In this subsection, the data that has already been studied as part of this project shall be discussed in some detail. Furthermore, the readings to be included in the model that have not yet been studied will be listed and discussed.

Beam Loss Monitors: BLMs were installed around the CERN particle accelerator complex to detect losses in the beam intensities as they're circulating. These monitors are safety critical as a very large amount of energy is stored in these circulating beams. In fact, as Barbara Holzer *et. al.* mention in [14], "The loss of even a very small fraction of this beam may ... cause physical damage to machine components." As of yet, readings from the TDI BLMs for both beams have been analysed. Each collimator in the TDI has 3 BLMs each taking readings of the beam losses at the same time in order to ensure accuracy in the readings. As can be seen in Figure 1, it was noted that there is a spike in the

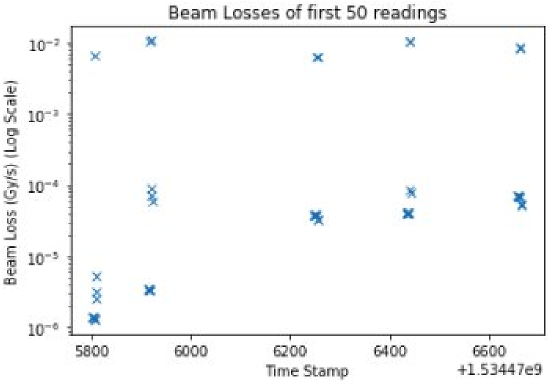

Figure 1: BLM losses at injection

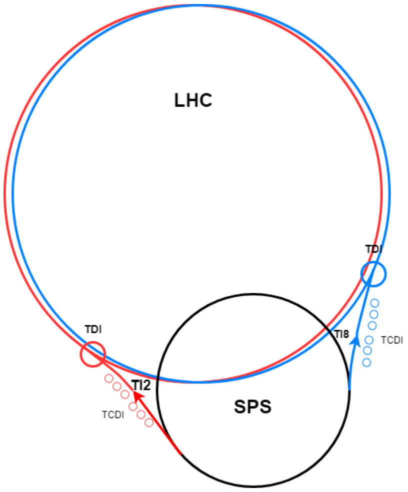

Figure 2: Diagram of the particular area of interest of the CERN particle accelerator complex for this dissertation

readings which corresponds to the time of injection into the LHC. As expected, since these readings are being taken for the same beam at almost the same position, they are highly correlated. Furthermore, some other readings (yet to be analysed) that will be included as parameters to the implemented model are the TCDI (Transfer Line Collimator) BLMs as well as the BLMs present at the point of extraction from the SPS. Figure 2 shows the positions of these monitors in the CERN accelerator complex.

Abort Gap: The abort gap refers to the number of buckets in a row which are kept empty in order to account for the time it takes for the beam abort kicker magnets to rise. In order to fill the LHC with the nominal 2808 bunches, there must be an abort gap of at least $3\mu s$ [15]. These readings have also been extracted, analysed and prepared to be included as a parameter for the anomaly detection algorithms.

Beam Intensity: The beam intensity readings were taken from two sources; at the moment of extraction from the SPS and at the moment injection into the LHC. These readings were also analysed and prepared to be included as parameters for the anomaly detection algorithms. Some points of interest at this stage can already be identified as anomalies, an example of such an anomaly is when the SPS would have a positive intensity reading at the point of extraction but the change in intensity at the LHC would actually be negative.

Other parameters of interest: Some other parameters which are yet to be extracted and analysed which are planned to be included in this project are the emittance, the BPMs

(Beam Position Monitors) and the Collimator Positions.

The Evaluation Strategy and Technique being Proposed

In order to evaluate the performance of the anomaly detection algorithms being proposed, the results produced will be checked with those of the current IQC (Injection Quality Check). However since the purpose of this FYP is to improve the performance of the IQC and propose a better technique for injection quality checks, the results will also be examined manually on a case-by-case basis.

Deliverables

- A detailed literature review on understanding the problem domain as well as extensive details on the anomaly detection algorithms used.

- Details on the methodology used and techniques applied for analysis.

- The results obtained and comparisons with the IQC.

Current Progress

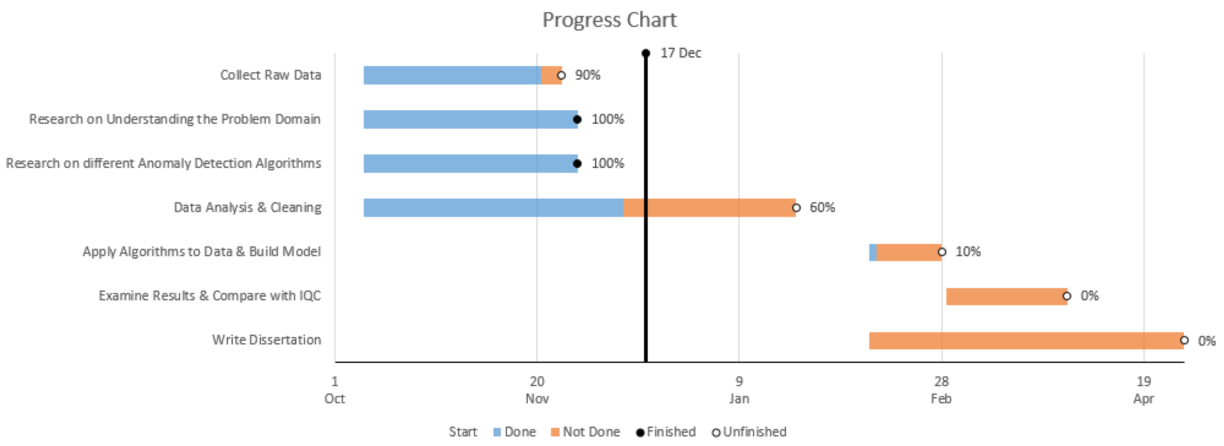

Figure 3: Gnatt Chart showing progress

Bibliography

[1] V. Kain *et. al* "Injection beam loss and beam quality checks for the lhc." in *Proc. IPAC*, 2010, pp. 1671-1673.

[2] C. Lefevre. "The cern accelerator complex." Technical report, 2008.

[3] C. Roderick, L. Burdzanowski and G. Kruk. "The cern accelerator logging service- 10 years in operation: a look at the past, present and future," presented at the 14th Int. Conf. Accelerator & Large Experimental Physics Control Systems, USA, 2013.

[4] L. N. Drosdal *et. al.* "Automatic injection quality checks for the lhc." in *Proc. ICALEPCS*, 2011, pp. 1077-1080.

[5] "Cern document server" Internet: `cds.cern.ch`, [Nov. 11, 2018].

[6] A. Hyvärinen. Lecture Notes, Topic: "Unsupervised machine learning." University of Helsinki, 2015.

[7] M. Ester *et. al.* "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Proc. KDD*, 1996, pp. 226-231.

[8] "Clustering." Internet: `scikit-learn.org/stable/modules/clustering.html`, [Nov. 27, 2018].

[9] M. Breunig *et. al.* "Lof: identifying density-based local outliers," in *Proc. ACM SIGMOD*, 2000, pp. 1-12.

[10] M. Richardson. Class Lecture, Topic: "Principal Component Analysis." May, 2009.

[11] PyMVPA Authors. "Pymvpa developer guidelines." Internet: `www.pymvpa.org`, Aug. 28, 2017 [Nov. 26, 2018].

[12] "The shogun machine learning toolbox." Internet: `pypi.org/project/shogun-ml/`, [Nov. 26, 2018].

[13] F. Pendregosa *et. al.*. "Scikit learn: machine learning in python." *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, Oct. 2011.

[14] E. Barbara Holzer *et. al.* "Beam loss monitoring system for the lhc." in *Proc. IEEE NSS*, 2005.

[15] M. Meddahi *et. al.* “Lhc abort gap monitoring and cleaning,” presented at the 1^{st} Int. Particle Accelerator Conf, Japan, 2010.