Machine learning for data mining and performance optimization at the CERN Large Hadron Collider

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 clustering and anomaly detection shall be used to analyse past LHC 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 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. 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 thesis is to go through the entire process in a data science project and learn to use unsupervised machine learning techniques for applications such as clustering and 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 Machine Learning 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 machine learning algorithms that will be used in this study:

- *K-Means Clustering*: K-Means is considered to be one of the most popular clustering methods [7]. The idea behind the algorithm is to split the data into k clusters where a data point forms part of the cluster with the closest centroid. The time complexity for this algorithm is $O(nki)$ and the space complexity is $O(n(d+k))$ where n is the number of points, k the number of centres, d the number of dimensions and i the number of iterations required to converge [7].

- *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" [8]

  DBSCAN manages to attain the above requirements by viewing clusters as "areas of high density separated by areas of low density" [9]. 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" [10]. 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.

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 in the next subsection.

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], "(t)he loss of even a very small fraction of this beam may ... cause physical damage to machine components." Furthermore as stated in [14], a fast loss of 3×10^{-9} and a constant loss of 3×10^{-12} times the flat top beam intensity can quench a dipole magnet and a one turn loss of 3×10^{-6} times the nominal intensity can even damage a magnet. As of yet, readings from the TCDIs (Transfer Line Collimators) 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 readings which corresponds to the time of injection from the SPS to the LHC. These

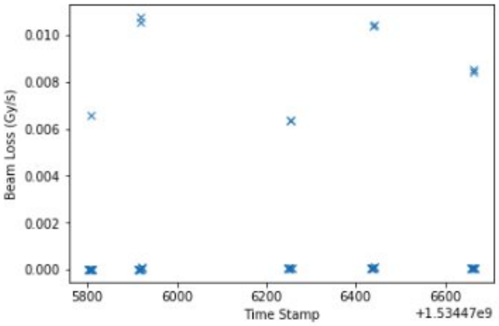

Figure 1: BLM losses at injection

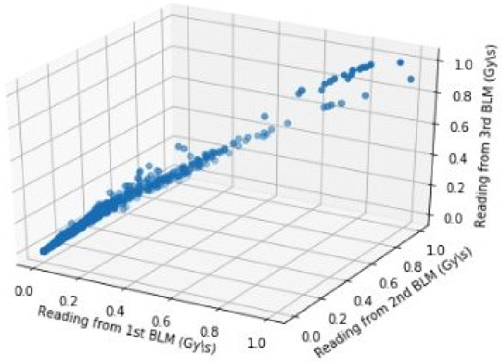

Figure 2: Normalised losses of 3 BLMs

readings were then filtered and a scatter plot as seen in Figure 2 was derived to visualise the correlation between the readings of the three BLMs.

The Evaluation Strategy and Technique being Proposed

Deliverables

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 14$^{\text{th}}$ 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] J. Xin and H. Jiawei. “K-means clustering.” *Encyclopaedia of Machine Learning*, pp. 563-564, 2011.

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

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

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

[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.