

Using machine learning algorithms for Quality Assurance in ALICE experiment

Tuesday 20 March 2018 16:00 (15 minutes)

Data Quality Assurance (QA) is an important aspect of every High-Energy Physics experiment, especially in the case of the ALICE experiment at the Large Hadron Collider (LHC) whose detectors are extremely sophisticated and complex devices. To avoid processing low quality or redundant data, and to classify it for analysis, human experts are currently involved in an offline assessment of the quality of the data itself and of the detectors' health during the collisions' recording. Since this assessment process is cumbersome and time-consuming, it typically takes experts days or months to assess the quality of the past data taking periods. Furthermore, since the ALICE Experiment is going to undergo a major upgrade in the coming years and it is planned to record much more data at higher frequency, manual data quality and detector health checks will simply not be feasible.

This is exactly the environment where machine learning can be utilized to its full extent. Based on the recent advancement in the field of machine learning and pattern recognition, we conducted several experiments that aim at automating QA for the ALICE experiment. More specifically, we collected a multi-dimensional dataset of attributes recorded during over 1'000 data taking periods by the Time Projection Chamber (TPC) and the corresponding quality labels. We normalized the data to disregard temporal dependencies as well as to minimize the noise. We cast our problem as a classification task whose goal is to assess the quality of the data collected by TPC detector. Since the space of assigned quality labels is very sparse, we simplified the multi-class problem to a binary classification task, by considering all bad and unlabeled data points as 'suspicious', while the remaining data portion was labeled as good. This normalization was recommended by detectors' experts, since the lack of labels is typically caused by unprecedented characteristics of the detector data.

The resulting binary classification task can be solved by several state-of-the-art machine learning algorithms. In our experiments, we have used both traditional shallow classification architectures, such as random trees and SVM classifiers, as well as modern neural network architectures. To ensure the generalization of our results and the robustness of the evaluated methods, we followed k-fold cross-validation procedure with $k=10$. The obtained results in terms of False Positive rate of less than 2% indicate that machine learning algorithms can be directly used to automatically detect the suspicious runs and hence reduce the human burden related to this task.

Our future research includes extending the analysis to other detectors' data, focusing first at those where the quality assessment procedure is more time-consuming. We then plan to investigate application of unsupervised machine learning methods to detect anomalies in detectors' data in real-time.

Author: Dr TRZCINSKI FOR THE ALICE COLLABORATION, Tomasz (Warsaw University of Technology)

Presenter: Dr TRZCINSKI FOR THE ALICE COLLABORATION, Tomasz (Warsaw University of Technology)

Session Classification: Poster

Track Classification: 4: Performance evaluation