Astronomy in the XXIst century: combining big data and machine learning to detect supernovae

Dr Julien Peloton, IJCLab
peloton@lal.in2p3.fr

Internship Summer 2020 - Master level

1 Description

Our research group is investigating how to leverage the big data ecosystem tools to analyse current and future data sets in astronomy [1]. Among the future large experiments, the Large Synoptic Survey Telescope (LSST, [2]) will start soon collecting terabytes of data per observation night, and the efficient processing and analysis of both real-time and historical data remains a major challenge.

Each observation night, telescopes all around the world issue alerts based on what they observe on the sky. These alerts are typically streamed to other places, where the stream is analysed and the relevance of each alert is asserted in order to take a decision on the next steps to perform. Such decisions include for example classification of objects based on machine learning algorithms. Given the unprecedented precision of next generation of telescopes, the stream of alerts will be made of millions of alerts per night, reaching the TB per night, and decisions and actions must be taken extremely fast.

This internship focuses more specifically on the distributed computing framework Apache Spark [3]. Apache Spark is an open-source framework for data analysis, based on the so-called MapReduce cluster computing paradigm, popularized by the Hadoop framework using implicit data parallelism and fault tolerance. In addition Spark optimizes data transfer, memory usage and communications between processes based on a graph analysis (DAG) of the tasks to perform, largely relying on the functional programming.

In this internship, we propose to tackle the challenge of combining natively Apache Spark streaming capabilities and machine learning to produce a fast and reliable classification of supernovae [4] in real time.

2 Proposed tasks

We propose first to study a popular ensemble learning method for classification known as Random Forests [5]. Our group developed a prototype version for Apache Spark by interfacing scikit-learn based tools [6], and the student will make a true native implementation using Apache Spark MLlib. A special attention will be put in comparing performances of the two approaches and performing reproducible benchmarks on existing datasets. Second, if time allows, the student will have the opportunity to focus on other methods for classification based on neural networks.

3 Requirements

- At least one of the following programming language: Scala or Python.

- Knowledge of parallel and distributed computing.
- Being aware of the big data challenges and issues.
- Working with or willing to learn Apache Spark.
- Working with or willing to learn machine learning.

[1] https://github.com/astrolabsoftware
[2] https://en.wikipedia.org/wiki/Vera_C._Rubin_Observatory
[3] http://spark.apache.org/
[4] https://en.wikipedia.org/wiki/Supernova
[5] https://en.wikipedia.org/wiki/Random_forest
[6] https://github.com/astrolabsoftware/fink-broker