Contribution ID: **82**   Type: **Poster**

Comparison of the CPU and memory performance of StatPatternRecognition (SPR) and Toolkit for MultiVariate Analysis (TMVA)

High Energy Physics data sets are often characterized by a huge number of events. Therefore, it is extremely important to use statistical packages able to efficiently analyze these unprecedented amounts of data. We compare the performance of the statistical packages StatPatternRecognition (SPR) and Toolkit for MultiVariate Analysis (TMVA). We focus on how CPU time and memory usage of the learning process scale versus data set size. As classifiers, we consider Random Forests, Boosted Decision Trees and Neural Networks. For our tests, we employ a data set widely used in the machine learning community, "Threenorm" data set, as well as data tailored for testing various edge cases. For each data set, we constantly increase its size and check CPU time and memory needed to build the classifiers implemented in SPR and TMVA.

We show that SPR is often significantly faster and consumes significantly less memory. For example, the SPR implementation of Random Forest is by an order of magnitude faster and consumes an order of magnitude less memory than TMVA on Threenorm data.

Primary author: PALOMBO, Giulio (California Institute of Technology)

Presenter: PALOMBO, Giulio (California Institute of Technology)

Track Classification: Track 2 : Data Analysis - Algorithms and Tools