Contribution ID: **188** Type: **Parallel Talk**

A Numeric Comparison of Feature Selection Algorithms for Supervised Learning

Wednesday, 5 November 2008 14:00 (25 minutes)

Datasets in modern High Energy Physics (HEP) experiments are often described by dozens or even hundreds of input variables (features). Reducing a full feature set to a subset that most completely represents information about data is therefore an important task in analysis of HEP data. We compare various feature selection algorithms for supervised learning using several datasets such as, for instance, imaging gamma-ray Cherenkov telescope (MAGIC) data found at the UCI repository.

We use classifiers and feature selection methods implemented in the statistical package StatPatternRecognition (SPR), a free open-source C++ package developed in the HEP community (http://sourceforge.net/projects/statpatrec/). For each dataset, we select a powerful classifier and estimate its learning accuracy on feature subsets obtained by various feature selection algorithms. When possible, we also estimate the CPU time needed for the feature subset selection. The results of this analysis are compared with those published previously for these datasets using other statistical packages such as R and Weka. We show that the most accurate, yet slowest, method is a wrapper algorithm known as generalized sequential forward selection ("Add N Remove R") implemented in SPR.

Primary author: Dr PALOMBO, Giulio (University of Milan - Bicocca)

Co-author: Dr NARSKY, Ilya (California Institute of Technology)

Presenter: Dr PALOMBO, Giulio (University of Milan - Bicocca)

Session Classification: Data Analysis - Algorithms and Tools

Track Classification: 2. Data Analysis