Contribution ID: **60** Type: **Parallel Talk**

Classifying extremely imbalanced data sets

Tuesday, February 23, 2010 2:25 PM (25 minutes)

Imbalanced data sets containing much more background than signal instances are very common in particle physics, and will also be characteristic for the upcoming analyses of LHC data. Following up the work presented at ACAT 2008, we use the multivariate technique presented there (a rule growing algorithm with the meta-methods bagging and instance weighting) on much more imbalanced data sets, especially a selection of D0 decays without the use of particle identification. It turns out that the quality of the result strongly depends on the number of background instances used for training. We discuss methods to exploit this in order to improve the results significantly, and how to handle and reduce the size of large training sets without loss of result quality in general. We will also comment on how to take into account statistical fluctuation in receiver operation curves (ROC) for comparing classifier methods.

Primary author: BRITSCH, Markward (Max-Planck-Institut fuer Kernphysik (MPI)-Unknown-Unknown)

Co-authors: SCHMELLING, Michael (Max-Planck-Institut fuer Kernphysik (MPI)); GAGUNASHVILI, Nikolai (University of Akureyri)

Presenter: BRITSCH, Markward (Max-Planck-Institut fuer Kernphysik (MPI)-Unknown-Unknown)

Session Classification: Tuesday, 23 February - Data Analysis - Algorithms and Tools

Track Classification: Data Analysis - Algorithms and Tools