ProjectName = OSG-CSC00001

Sponsor = OSG-XSEDE

Last_name = Rudolph

First_name = George

Email = rudolphg1@citadel.edu

Organization = The Citadel

Department = Department of Mathematics and Computer Science

Join_date = July 29, 2013

Field-of-Science= Computer Science

Abstract-of-Work:

Develop a metric that measures the real similarities and differences between machine learning algorithms (in this case classifiers) based on output behavior. My previous study included 17 representative algorithms and used 30 datasets from the UCI Machine Learning Repository.
The main goal of my current effort is to extend the metric using semi-supervised learning techniques. I would also like, if possible, to experiment with more recent datasets beyond what is traditional from the UCI Repository; and to add more algorithms to the study.

The method involves training 10 (arbitrarily chosen number) copies of each algorithm on each dataset, then comparing the outputs of each copy that was trained on the same data against a training set and counting the differences. Each difference value is divided by the size of the respective test set in order to obtain a difference measure between 0…1. The differences between pairs of algorithms are then placed in a dissimilarity matrix. The matrix is then used to generate a visualization of the differences in 2 or 3 dimensions. Generating data for the visualization involves dimension reduction, for which I have used multi-dimensional scaling. However, I would like to consider alternative techniques, because I believe MDS breaks down when the number of variables is around 20.

Each copy of an algorithm can be trained and tested separately in parallel. Computing the average difference between each pair of algorithms is a gather/reduction operation, and could be so coded. Generating the visualization has not taken much compute time, and does not necessarily need to be parallelized, depending on the technique used.