Parallel Random Forests – PARF

Contributed by Viktor Bojović
Thursday, 21 February 2008
Last Updated Friday, 09 May 2008

Data mining commonly relies heavily on Pattern recognition, which aims to classify data based on either a priori knowledge or on statistical information extracted from the patterns. A range of classification algorithms has been devised. The Random Forests algorithm is one of the best among the known classification algorithms, able to classify big quantities of data with great accuracy. For the Random Forests, in addition to a set of important statistical features, its loosely coupled structure allows the classifier training procedure to be readily parallelized. The RBI/CIC team reimplemented the original algorithm, and explored several parallelization strategies, ranging from SMP, through MPI, to Grid job schedulers. The creator of the algorithm, late Berkeley professor emeritus Leo Breiman, expressed a big interest in this idea in our correspondence. He has confirmed that no one was yet working on a parallel implementation of his algorithm, and promised his support and help. Leo Breiman is one
of the pioneers in the fields of machine learning and data mining, and a co-author of the first significant programs (CART – Classification and Regression Trees) in that field.
The present application is command line based, MPI-enabled, and if statically linked can be deployed to any system in the grid (or any system with MPICH support, if MPI execution is desired).