28.04.2009

Barış Yüksel
1449826

CENG 492 WEEKLY REPORT

This week we have parallelized the SimpleKMeans class under weka.clusterers package. The code segment we have parallelized is the one which takes an instance and assigns it to a cluster. In our parallelization – which is data parallelism by the way - the instances are almost equally shared among the processors, and each processor is responsible for assigning the cluster to the instances belonging to them. Then, these cluster assignments are gathered for each processor at each iteration. Whether the system has been converged or not is also reduced (with and operation) for each processor at each iteration.

The results were good for datasets with large amount of instances. For the dataset we have tested with about 2500 instances and 2000 attributes, the parallelized version worked better than the normal one, in spite of the high communication cost of PJ. On the other hand, with smaller datasets, the parallelized version worked slightly worse.

Also for this week, we tried to build parallelization algorithms for matrix decompositions, however we couldn't manage it. For all decompositions, the matrix changes at each iterations, and we couldn't manage to do it without the data dependencies.

For next week, we will parallelize another algorithm under clusterers, associations or classifiers packages (most probably we will start NaiveBayesSimple algorithm). Since lower level parallelizations for Instances and Utils didn't work, we will now on try to parallelize the algorithms under these packages instead.