DIVIDE & RECOMBINE WITH DELTARHO FOR BIGGER DATA AND HIGHER COMPUTATIONAL COMPLEXITY

Abstract: Computational performance is challenging today. Datasets can be big, computational complexity of analytic methods can be high, and computer hardware power can be limited. Divide & Recombine (D&R) is a statistical approach to meet the challenges.

The analyst divides the data into subsets by a D&R division method. Each analytic method is applied to each subset, independently, without communication. The outputs of each analytic method are recombined by a D&R recombination method. Sometimes the goal is one result for all of the data, such as a logistic regression; D&R theory and methods seek division and recombination methods to optimize the statistical accuracy. Much more common in practice is a division based on the subject matter. The data are divided by conditioning on variables important to the analysis. In this case the outputs can be the final result, or further analysis is carried out, an analytic recombination.

D&R computation is mostly embarrassingly parallel, the simplest parallel computation. DeltaRho software is an open-source implementation of D&R. The front end is the R package datadr, which is a language for D&R. It makes programming D&R simple. At the back end, running on a cluster, is a distributed database and parallel compute engine such as
Hadoop, which spreads subsets and outputs across the cluster, and executes the analyst R and datadr code in parallel. The R package RHIPE provides communication between datadr and Hadoop. DeltaRho protects the analyst from management of parallel computation and database management.

D&R with DeltaRho can increase the dramatically the data size and analytic computational complexity that are feasible in practice, whether hardware power of an available cluster is small, medium, or large. The data can have a memory size that is larger than the physical cluster memory.