Extended Abstract: Hierarchical Data Analysis

Wendy Di, Stefan Wild

March 27, 2017

Many scientific data analysis problems require discovering latent/intrinsic structure within the data, grouping the data into corresponding clusters, and then performing operations on the respective groups. To address the curse of dimensionality, we are developing hierarchical methods that exploit the nested structure embedded in many natural phenomena and their mathematical representations, such as those arising from spatial/temporal discretization, network structure, and parameter space characteristics. By exploiting these hierarchies, we intend to develop a more scalable approach to a broad range of analysis tasks.

A motivating example that we will target initially is the halo-center-finding problem arising in computational cosmology. The goal is to find a particle within a halo that provides the lowest potential, where the potential for a given particle requires a global calculation over all particles (e.g., proportional to the sum of inverse interparticle distances). Given n particles within a halo, the standard brute force algorithm will end up with $O(n^2)$ complexity. A typical halo has 20 million particles and thus such a global operation comes with considerable expense. Our approach is to design a hierarchical clustering scheme. Instead of calculating correlations between every pair of particles, we calculate the correlations of centers provided by every pair of sub clusters that can at least reduce the computation cost to $O(n \log n)$. We will test our algorithm on benchmark datasets provided by the HACC ECP application project. Our hypothesis is that in many cases, the hierarchical algorithm will perform significantly better than its worst-case complexity guarantee.

We will also benchmark the hierarchical approach on material science analysis and compression problems encountered by the EXAALT application project. The hierarchical nature of these algorithms means that we can often find surrogate models with less complexity, and approximate solutions (by operating at the coarsest levels of the hierarchy whenever possible) in far less time than single-level traditional approaches. As such, we will perform a full tradeoff study using emerging platforms to quantify the tradeoffs between accuracy and performance on these test applications, and provide guidance on its suitability for real-time online analysis.