Parallel Data Mining from Multicore to Cloudy Grids

Geoffrey Fox
Indiana University

Cetraro HPC2008

Abstract
We describe a suite of data mining tools that cover clustering, Gaussian modeling and dimensional reduction and embedding. These are applied to three class of applications; Geographical information systems, cheminformatics and bioinformatics. The data vary in dimension from low (2), high (thousands) to undefined (sequences with dissimilarities but not vectors defined). We use deterministic annealing to provide more robust algorithms that are relatively insensitive to local minima. We use embedding algorithms both to associate vectors with sequences and to map high dimensional data to low dimensions for visualization. We discuss the algorithm structure and their mapping to parallel architectures of different types and look at the performance of the algorithms on three classes of system; multicore, cluster and Grid using a MapReduce style algorithm. Each approach is suitable in different application scenarios.