Random Forests for Big Data

Nathalie Villa-Vialaneix, Robin Genuer, Jean-Michel Poggi and Christine Tuleau-Malot

Based on decision trees combined with aggregation and bootstrap ideas, random forests were introduced by Breiman in 2001. They are a powerful nonparametric statistical method allowing to consider in a single and versatile framework regression problems, as well as two-class and multi-class classification problems. Focusing on classification problems, this paper reviews available proposals about random forests in parallel environments as well as about online random forests. Then, we formulate various remarks for random forests in the Big Data context.
Finally, we experiment three variants involving subsampling, Big Data-bootstrap and MapReduce respectively, on two massive datasets (15 and 120 millions of observations), a simulated one as well as real world data.