Random Forests for Big Data

R. Genuer[a], J.-M. Poggi[b], C. Tuleau-Malot[c], N. Villa-Vialaneix[d]

[a]*ISPED, Univ. Bordeaux, France,* [b]*LMO, Univ. Paris-Sud Orsay & Univ. Paris Descartes, France,* [c]*Univ. Côte d'Azur, CNRS, LJAD, France* [d]*MIAT, Univ. de Toulouse, INRA, France*

Big Data are a major challenge of statistical science and has numerous algorithmic and theoretical consequences. Big Data always involve massive data and often include online data and data heterogeneity.

Recently statistical methods have been adapted to process Big Data, like linear regression models, clustering methods and bootstrapping schemes. Based on decision trees combined with aggregation and bootstrap ideas, random forests (RF) are a powerful nonparametric statistical method allowing to consider in a single and versatile framework regression problems, as well as classification ones.

Focusing on classification problems, this talk proposes a review of proposals that deal with scaling random forests to Big Data problems. These proposals rely on parallel environments or on online adaptations of RF. More precisely, one variant relies on subsampling while three others are related to parallel implementations of random forests and involve either various adaptations of bootstrap to Big Data or divide-and-conquer approaches. The fifth variant is related to online learning of random forests. We also describe how the out-of-bag error is addressed in these methods. Then, we formulate various remarks for RF in the Big Data context. Finally, we experiment five variants on two massive datasets, a simulated one and a real-world dataset. These numerical experiments lead to highlight the relative performance of the different variants, as well as some of their limitations.

This talk is related to the recent paper [1].

Keywords: Random forests, Big Data, Statistics

References

[1] R. Genuer, J-M. Poggi, C. Tuleau-Malot, N. Villa-Vialaneix (2017). Random Forests for Big Data. *Big Data Research*, **9**, 28–46.