Gradient boosting for model classification in computational biology

Research project supervised by
Michael Blum (michael.blum@univ-grenoble-alpes.fr, 04 56 52 00 65)
Olivier François (olivier.francois@univ-grenoble-alpes.fr, 04 56 52 00 25)

Laboratory: TIMC-IMAG, Faculté de medicine, 38706 La Tronche

Relevant Track in MSIAM: Data Science

Short description
Numerical simulations in biology provide testable hypotheses about the mechanisms that can explain generated data. A statistical method called “Approximate Bayesian Computation” has been developed to measure the relative statistical support of different biological mechanisms based on numerical simulations (Csillery et al. 2010). Among others, it has been used to compare models of human evolution (Vernot and Akey 2014) or to fit stochastic models to data in systems biology (Liepe et al. 2014).

In the lab, we have developed the R package abc to perform Approximate Bayesian Computation (Csillery et al. 2012). Since the release of the R package, substantial statistical improvements have been proposed for model classification in ABC including a random forest approach (Pudlo et al. 2016). The objective of the Msc thesis is to implement a gradient boosting approach in the R package and to test if it is competitive with respect to other methods of model selection. Gradient boosting as implemented in XGBoost is an emerging method of data science, and it is one of the most frequently used package to win machine-learning data challenges (Chen and Guestrin 2016).

References

Chen, T. and Guestrin, C., 2016. Xgboost: A scalable tree boosting system. arXiv:1603.02754.

Csilléry, K., Blum, M.G., Gaggiotti, O.E. and François, O., 2010. Approximate Bayesian computation (ABC) in practice. Trends in ecology & evolution, 25(7), pp.410-418.

Csilléry, K., François, O. and Blum, M.G., 2012. abc: an R package for approximate Bayesian computation (ABC). *Methods in ecology and evolution*,*3*(3), pp.475-479.

Liepe, J., Barnes, C., Cule, E., Erguler, K., Kirk, P., Toni, T. and Stumpf, M.P., 2010. ABC-SysBio—approximate Bayesian computation in Python with GPU support. Bioinformatics, 26(14), pp.1797-1799.

Pudlo, P., Marin, J.M., Estoup, A., Cornuet, J.M., Gautier, M. and Robert, C.P., 2016. Reliable ABC model choice via random forests. Bioinformatics, 32(6), pp.859-866.

Vernot, B. and Akey, J.M., 2014. Resurrecting surviving Neandertal lineages from modern human genomes. Science, 343(6174), pp.1017-1021.