Peer Community In
Evolutionary Biology

Machine learning methods are useful for Approximate Bayesian Computation in evolution and ecology

Michael Blum

Laboratoire TIMC-IMAG, Univ. Grenoble Alpes -- Grenoble, France
michael.blum@univ-grenoble-alpes.fr
doi: 10.24072/pci.evolbiol.100036

Open Access

Cite as: Blum M. 2017. Machine learning methods are useful for Approximate Bayesian Computation in evolution and ecology. *Peer Community in Evolutionary Biology*, 100036. doi: 10.24072/pci.evolbiol.100036

A recommendation – based on reviews by Dennis Prangle and Michael Blum – of

Raynal L, Marin J-M, Pudlo P, Ribatet M, Robert CP, Estoup A. 2017. **ABC random forests for Bayesian parameter inference**. arXiv 1605.05537v4, https://arxiv.org/pdf/1605.05537

Published: 17th November 2017

It is my pleasure to recommend the paper by Raynal *et al.* [1] about using random forest for parameter inference. There are two reviews about the paper, one review written by Dennis Prangle and another review written by myself. Both reviews were positive and included comments that have been addressed in the current version of the preprint.

The paper nicely shows that modern machine learning approaches are useful for Approximate Bayesian Computation (ABC) and more generally for simulation-driven parameter inference in ecology and evolution.

The authors propose to consider the random forest approach, proposed by Meinshausen [2] to perform quantile regression. The numerical implementation of ABC with random forest, available in the abcrf package, is based on the RANGER R package that provides a fast implementation of random forest for high-dimensional data.

According to my reading of the manuscript, there are 3 main advantages when using random forest (RF) for parameter inference with ABC. The first advantage

is that RF can handle many summary statistics and that dimension reduction is not needed when using RF.

The second advantage is very nicely displayed in Figure 5, which shows the main result of the paper. If correct, 95% posterior credibility intervals (C.I.) should contain 95% of the parameter values used in simulations. Figure 5 shows that posterior C.I. obtained with rejection are too large compared to other methods. By contrast, C.I. obtained with regression methods have been shrunken. However, the shrinkage can be excessive for the smallest tolerance rates, with coverage values that can be equal to 85% instead of the expected 95% value. The attractive property of RF is that C.I. have been shrunken but the coverage is of 100% resulting in a conservative decision about parameter values.

The last advantage is that no hyperparameter should be chosen. It is a parameter free approach, which is desirable because of the potential difficulty of choosing an appropriate acceptance rate.

The main drawback of the proposed approach concerns joint parameter inference. There are many settings where the joint parameter distribution is of interest and the proposed RF approach cannot handle that. In population genetics for example, estimation of the severity and of the duration of the bottleneck should be estimated jointly because of identifiability issues. The challenge of performing joint parameter inference with RF might constitute a useful research perspective.

References

[1] Raynal L, Marin J-M, Pudlo P, Ribatet M, Robert CP, Estoup A. 2017. ABC random forests for Bayesian parameter inference. arXiv 1605.05537v4, https://arxiv.org/pdf/1605.05537

[2] Meinshausen N. 2006. Quantile regression forests. Journal of Machine Learning Research 7: 983-999. http://www.jmlr.org/papers/v7/meinshausen06a.html

Appendix

Reviews by Dennis Prangle and Michael Blum: http://dx.doi.org/10.24072/pci.evolbiol.100036