The Journal of Open Source Software

compboost: Modular Framework for Component-Wise Boosting

Daniel Schalk[1], **Janek Thomas**[1], **and Bernd Bischl**[1]

1 Department of Statistics, LMU Munich

DOI: 10.21105/joss.00967

Software

- Review
- Repository
- Archive

Submitted: 14 August 2018
Published: 12 October 2018

Summary

In high-dimensional prediction problems, especially in the $p \geq n$ situation, feature selection is an essential tool. A fundamental method for problems of this type is component-wise gradient boosting, which automatically selects from a pool of base learners – e.g. simple linear effects or component-wise smoothing splines (Schmid & Hothorn, 2008) – and produces a sparse additive statistical model. Boosting these kinds of models maintains interpretability and enables unbiased model selection in high-dimensional feature spaces (Hofner, Hothorn, Kneib, & Schmid, 2012).

The `R` (Team, 2016) package `compboost`, which is actively developed on GitHub (https://github.com/schalkdaniel/compboost), implements component-wise boosting in `C++` using `Rcpp` (Eddelbuettel, 2013) and `Armadillo` (Sanderson & Curtin, 2016) to achieve efficient runtime behavior and full memory control. It provides a modular object-oriented system which can be extended with new base-learners, loss functions, optimization strategies, and stopping criteria, either in `R` for convenient prototyping or directly in `C++` for optimized speed. The latter extensions can be added at runtime, without recompiling the whole framework. This allows researchers to easily implement more specialized base-learners, e.g., for spatial or random effects, used in their respective research area.

Visualization of selected effects, efficient adjustment of the number of iterations, and traces of selected base-learners and losses to obtain information about feature importance are supported.

Compared to the reference implementation for component-wise gradient boosting in `R`, `mboost` (Hothorn, Buehlmann, Kneib, Schmid, & Hofner, 2017), `compboost` is optimized for larger datasets and easier to extend, even though it currently lacks some of the large functionality `mboost` provides. A detailed benchmark against `mboost` can be viewed on the project homepage and on GitHub.

The modular design of `compboost` allows extension to more complicated settings like functional data or survival analysis. Further work on the package should include parallelized boosting, better feature selection, faster optimization techniques such as momentum and adaptive learning rates, as well as better overfitting control.

References

Eddelbuettel, D. (2013). *Seamless r and c++ integration with rcpp*. Springer. doi:10.1007/978-1-4614-6868-4

Hofner, B., Hothorn, T., Kneib, T., & Schmid, M. (2012). A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics*, *20*(4), 956–971. doi:10.1198/jcgs.2011.09220

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2017). *mboost: Model-based boosting*. Retrieved from https://CRAN.R-project.org/package=mboost

Sanderson, C., & Curtin, R. (2016). Armadillo: A template-based c++ library for linear algebra. *Journal of Open Source Software*, *1*(2), 26. doi:10.21105/joss.00026

Schmid, M., & Hothorn, T. (2008). Boosting additive models using component-wise p-splines. *Computational Statistics & Data Analysis*, *53*(2), 298–311.

Team, R. C. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/