## The Arborist: a Scalable Decision Tree Implementation

Mark Seligman<sup>1,\*</sup>

1. Rapidics LLC \*Contact author: mseligman@rapidics.com

**Keywords:** machine learning, high performance, big data

**The Arborist** is an implementation of the Random Forest algorithm (Breiman 2001) invocable through an *R* package interface. The software is tailored for high performance, with execution time scaling linearly in both predictor and row count. Both regression and categorical cases are supported, with no limit on the number of factor levels in either the response or the predictors. In addition to standard features, **The Arborist** offers quantile regression, missing-value handling, and automatic resampling.

The Arborist's interface with *R* employs the Ropp template extension, allowing the software to be invoked by package. Specialized GPU versions are under development, but a general-purpose, multicore version is being made available under MPL-2 license.

We describe the organization of the software and compare performance with other implementations of the Random Forest algorithm.

## References

- [1] Breiman, L. (2001). Title of an article. *Machine Learning* 45, 5–32.
- [2] Breiman, L. and A. Cutler (2006). Random forests. http://www.stat.berkeley.edu/~breiman/RandomForests/.
- [3] Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. R News 2(3), 18–22.