subsemble: Ensemble learning in *R* with the Subsemble algorithm

Erin LeDell [1,2,⋆], **Stephanie Sapp**[1,3], **Mark van der Laan**[1,2,3]

1. University of California, Berkeley
2. Division of Biostatistics
3. Department of Statistics
⋆Contact author: ledell@berkeley.edu

Keywords: machine learning, ensemble methods, cross-validation, prediction, big data

We present the **subsemble** *R* package, which implements the Subsemble ensemble machine learning algorithm (Sapp et al., 2013), a new variant of Super Learning (van der Laan et al., 2008). Ensemble methods that combine models trained on different subsets of observations have recently received increased attention as practical prediction tools for massive datasets. Subsemble is a general subset ensemble prediction method that partitions a full dataset into subsets of observations and trains a user-specified learning algorithm on each subset. Then a unique form of V-fold cross-validation is used to learn a final prediction function which combines the subset-specific fits via a user-specified metalearner algorithm. Instead of simply averaging subset-specific fits, Subsemble differentiates fit quality across the subsets and learns an optimal combination of the subset-specific fits.

This implementation allows the user to ensemble subset-specific fits which are trained using the same or different learning algorithms. The package uses the machine learning algorithm API provided by the **SuperLearner** *R* package. This user-friendly API currently provides a uniform interface to nearly 30 machine learning algorithms (e.g. `randomForest`, `gbm`) and allows the user to define custom algorithm wrappers. Each of the default algorithm wrappers can also be customized by specifying unique model parameters.

The user can either explicitly define which observations belong to each subset or simply specify the desired number of subsets. In the case of the latter, the subsets will be created randomly, with or without stratification. The package provides the ability to compute the V-fold cross-validation step as well as the model fitting across the subsets in parallel using the *R*-core **parallel** package.

The **subsemble** package will be released to CRAN soon and the current version of the package can be found here: `http://www.stat.berkeley.edu/~ledell/R/subsemble.tar.gz`

References

Sapp, S., van der Laan, M. J., and Canny, J. (2013). Subsemble: an ensemble method for combining subset-specific algorithm fits. *Journal of Applied Statistics*.
Article: `http://dx.doi.org/10.1080/02664763.2013.864263`
Tech report: `https://biostats.bepress.com/ucbbiostat/paper313`

van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2008). Super Learner. *Statistical Applications of Genetics and Molecular Biology*, 6, article 25.
Article: `http://dx.doi.org/10.2202/1544-6115.1309`
Tech report: `http://biostats.bepress.com/ucbbiostat/paper222`