MetaForest: Exploring heterogeneity in meta-analysis using insights from machine learning.

Lissa, van C. J.[1*]

[1] Erasmus University Rotterdam, NL

* Presenting author

Suggested talk duration:

30 minutes

Summary (max. 500 words):

Meta-analysis often presents a small sample problem: The number of studies on any given topic is typically low, because conducting research is cost- and time-intensive. Human behavior, however, is notoriously complex (Earp & Trafimow, 2015), and consequently, subject to a host of potential moderators (Cesario, 2014). Additional moderators are introduced because similar research questions are examined in different labs, sampling from different populations, using idiosyncratic methods and instrumentation (Higgins et al., 2009). Even replication studies, by definition designed to be equivalent, typically display heterogeneity due to unforeseen moderators (Maxwell et al., 2015; Simmons et al., 2011). Finally, the paucity of theory regarding sources of heterogeneity at the between-studies level, makes it hard to whittle the list of potential moderators down to a manageable number (Thompson & Higgins, 2002). Meta-analysts are thus faced with what is known as the “curse of dimensionality”: The problem that arises when the number of variables to be considered is large, relative to the number of cases in the data. Such cases do not fit comfortably into the classic meta-analysis paradigm, which, like any regression-based approach, requires many cases per parameter. This may partly explain why, despite the fact that software to conduct meta-analysis with multiple moderators is readily available (Viechtbauer, 2010), most published meta-analyses do not account for more than a few moderators, if any. In many cases, the sample size is simply too low to obtain the power required to reliably examine heterogeneity (Riley, Higgins, & Deeks, 2011).

Three approaches have been proposed to deal with between-studies heterogeneity (Higgins et al., 2009): First, if studies are assumed to be different, they should not be meta-analyzed. Secondly, if they are similar, a random-effects model can estimate the distribution of the true effect size. Thirdly, if known differences between studies introduce heterogeneity, these moderators can be accounted for using meta-regression. What is currently lacking is a “fourth approach”, for cases where heterogeneity is suspected, but the causes are unknown. This calls for an exploratory technique which can perform variable selection — indentifying which moderators most strongly influence the observed effect size.

MetaForest aims to address this need. This technique applies random-effects weights from classic meta-analysis to random forests’ bootstrapping procedure. Random forests are a powerful learning algorithm, flexible yet relatively robust to overfitting. Simulation studies show that, even in datasets as small as 20 cases, MetaForest has excellent performance in terms of three metrics: 1) Predictive performance, in terms of cross-validated R^2_{cv}; 2) power, as evidenced by the proportion of datasets in which the algorithm achieved a positive R^2_{cv}; and 3) the ability to distinguish relevant moderators from irrelevant moderators, using variable importance measures. My presentation will

cover these simulations, and provide a short tutorial. Although MetaForest constitutes a fully-fledged paradigm for meta-analysis, it can also be integrated in existing workflows, as a final check to ensure that important moderators have not been overlooked. We hope that this approach will be of use to researchers, and that the convenient R package will facilitate its adoption. Please `install.packages("metaforest")`

Relevance to conference theme:

Meta-analyses are a classic case of the small sample problem, yet rarely discussed as such. Although the Bayesian approach is promising, it relies heavily on informative priors when sample size is small – and these may be hard to come by, given the lack of theory regarding moderators at the between-studies level. The machine-learning approach described here, which applies effective regularization through the bootstrap aggregation of randomly permuted regression trees, might present a viable alternative approach.

Keywords (max. 3):

Meta-analysis; machine learning; random forests