Subset Selection Ensembles

Anthony-Alexander Christidis, Stefan Van Aelst, Ruben Zamar

Department of Mathematics, University of Leuven

Two key approaches for high-dimensional regression are sparse methods such as best subset selection and ensemble methods such as random forests. Sparse methods have the advantage that they yield interpretable models. However, they are often outperformed in terms of prediction accuracy by “blackbox” multi-model ensemble methods. We propose an algorithm to optimize an ensemble of penalized regression models by extending recent developments in optimization for sparse methods to multi-model regression ensembles. The algorithm learns sparse and diverse models in the ensemble simultaneously from the data. Each of these models provides an explanation for the relationship between a subset of predictors and the response variable. To initialize our algorithm forward stepwise regression is generalized to multi-model regression ensembles. The resulting ensembles achieve excellent prediction accuracy by exploiting the accuracy-diversity tradeoff of ensembles. The ensembles can outperform state-of-the-art competitors on both simulated and real data.