EGU2020-1963
https://doi.org/10.5194/egusphere-egu2020-1963
EGU General Assembly 2020

Entropy Ensemble Filter: Does information content assessment of bootstrapped training datasets before model training lead to better trade-off between ensemble size and predictive performance?

Hossein Foroozand and Steven V. Weijs
University of British Columbia, Civil Engineering, vancouver, Canada

Machine learning is the fast-growing branch of data-driven models, and its main objective is to use computational methods to become more accurate in predicting outcomes without being explicitly programmed. In this field, a way to improve model predictions is to use a large collection of models (called ensemble) instead of a single one. Each model is then trained on slightly different samples of the original data, and their predictions are averaged. This is called bootstrap aggregating, or Bagging, and is widely applied. A recurring question in previous works was: how to choose the ensemble size of training data sets for tuning the weights in machine learning? The computational cost of ensemble-based methods scales with the size of the ensemble, but excessively reducing the ensemble size comes at the cost of reduced predictive performance. The choice of ensemble size was often determined based on the size of input data and available computational power, which can become a limiting factor for larger datasets and complex models' training. In this research, it is our hypothesis that if an ensemble of artificial neural networks (ANN) models or any other machine learning technique uses the most informative ensemble members for training purpose rather than all bootstrapped ensemble members, it could reduce the computational time substantially without negatively affecting the performance of simulation.