Overfitting and overextending -reframing the potential of machine-learning techniques in calibrating low-cost sensors

Thursday 22 June 2023 15:50 (15 minutes)

Machine-learning (ML) techniques have been recently applied to the calibration of low-cost sensors (LCS). Many studies report successes in calibration with ML techniques such as random forests (RF), neural networks (NN), and support vector regression (SVR). We find that calibrating LCS for the measurement of nitrogen dioxide (NO2) and particulate matter (PM) with ML techniques is not as beneficial as previously reported. While some hierarchical tree-based methods such as RF and gradient-boosting machines (GBM) find success, they also have substantial limitations. Others such as NN and SVR are prone to overfitting, such that prediction with these models on new data is inadvisable. Instead, we find in calibrating for NO2 and PM2.5, multiple linear regression (MLR) is the most reliable, transparent, and consistent. Though many ML techniques have potential for use in a variety of applications, they may not always be appropriate, as shown here with the calibration of LCS.

Primary author: SCHMITZ, Sean

Presenter: SCHMITZ, Sean

Session Classification: Deep/Machine learning and data science

Track Classification: Deep/Machine learning and data science