Improving Prediction Accuracy of Regression Problems with Optimization-based Ensemble Learning and a Two-layer Feature Selection Method

Fatemeh Amini, Mohsen Shahhosseini, Guiping Hu, Hieu Pham

This study proposes two state-of-art optimization-based methodologies to improve prediction accuracy for regression problems. We first considered blending as one type of ensemble creating method and designed an optimization-based ensemble learning algorithm that not only intends to reduce variance, but also aims at decreasing the prediction bias. To this end, a nested algorithm based on bi-level optimization that considers tuning hyperparameters as well as finding the optimal weights to combine ensembles was proposed. The experimental results show that the proposed algorithm outperforms base learners as well as benchmark ensembles (average ensembles and stacked regression) in 9 out of 10 datasets. In the second chapter, the problem of ultra-high-dimensional datasets, in which the number of predictors exceeds the number of observations, is studied and an optimization-based model using Genetic Algorithm (GA) was proposed. The two-layer optimization model considers minimizing prediction RMSE and number of selected predictors using GA with Elastic Net regularization as its fitness function, in the first layer. In the second layer, the best subset of predictors is used to apply simple Elastic Net on, intending to eliminate more predictors. The real-world Maize genetic datasets from NAM population have been used to test the performance of the proposed hybrid approach and the experimental results confirms its superiority over simple Elastic Net and Linear Regression combined with GA.

Keywords— Ensemble, Bias-Variance tradeoff, Genetic Algorithms, Feature Selection, Bi-level Optimization