Extracting chemical information in supervised learning through variable selection

유 준†, Petar Zuvela
부경대학교
(jayliu@pknu.ac.kr†)

In this work, six variable selection methods were evaluated for extraction of chemical information in supervised learning problems. Namely, Genetic Algorithm (GA), Firefly Algorithm (FA), Particle Swarm Optimization (PSO), Least Absolute Shrinkage and Selection Operator (LASSO), and Least Angle Regression Algorithm (LARS). The work consisted of two case studies: (i) prediction of soil carbonate content from spectral information, and (ii) classification of cancer patients from gene expression. Three performance measures: predictive ability, selection of true features from the full dataset, and robustness were used to evaluate the variable selection methods.
Results have shown that in order of decreasing predictive ability and robustness: GA > FA ≈ PSO > LASSO > LARS are recommended. For classification, the following trend: GA > PSO > FA ≈ LASSO > LARS has been observed. Strong robustness has been observed in the regression case for GA, FA and PSO. In the classification case, only LARS exhibited a considerable decrease in accuracy upon introduction of noise features.