Abstract

Background: Feature selection, also known as variable selection, is a technique that selects a subset from a large collection of possible predictors to improve the prediction accuracy in regression model. First objective of this project is to investigate in what data structure LASSO outperforms forward stepwise method. The second objective is to develop a feature selection method, Feature Selection by L_1 Regularization of Subset of Parameters (LRSP), which selects the model by combining prior knowledge of inclusion of some covariates, if any, and the information collected from the data. Mathematically, LRSP minimizes the residual sum of squares subject to the sum of the absolute value of a subset of the coefficients being less than a constant. In this project, LRSP is compared with LASSO, Forward Selection, and Ordinary Least Squares to investigate their relative performance for different data structures. Results: simulation results indicate that for moderate number of small sized effects, forward selection outperforms LASSO in both prediction accuracy and the performance of variable selection when the variance of model error term is smaller, regardless of the correlations among the covariates; forward selection also works better in the performance of variable selection when the variance of error term is larger, but the correlations among the covariates are smaller. LRSP was shown to be an efficient method to deal with the problems when prior knowledge of inclusion of covariates is available, and it can also be applied to problems with nuisance parameters, such as linear discriminant analysis.