Abstract

Sequential Lasso is a superior method for feature selection in additive and interaction models compared with penalized likelihood methods. However, when the number of covariates p is extremely large, even the sequential Lasso method is computationally inefficient. In this thesis, a greedy version of sequential Lasso is proposed with the aim to ease the computation burden.

The greedy version of the sequential Lasso algorithm ”throws away” the covariates that have negligible correlations with the current residual at each step. That is, the covariate will not be considered at later steps if its correlation with current residual is small. Since the number of interaction-effect covariates are much larger than the number of main-effect covariates, the percentages of covariates that are ”thrown away” are different. Extended BIC (EBIC) which has been proved to have the property of selection consistency is used as the feature selection criterion in this thesis.

Two sets of computer programs using R have been written to carry out the comparison study. One set of computer program is written for the original sequential Lasso and the other set is written for the greedy version of it. To assess the influence on the accuracy of the feature selection, 12 sets of simulation studies are done. From the results of the simulation studies, we found that the greedy version of sequential Lasso is indeed computationally more efficient. However, one has to be careful in selecting the threshold value of deletion in order to have a decent accuracy. We have found that when p is relatively small, i.e. p=107, a 10% deletion of main-effect features and a 20% deletion of interaction-effect features at each step is a proper choice of deletion. And a 25% deletion of main-effect features and a 35% deletion of interaction-effect features at each step is a proper choice of deletion when p is relatively large, i.e. p=365 and 1706.