A Method for Variables Selection and Prediction

Shaw-Hwa Lo
Department of Statistics, Columbia University

Abstract

A recent puzzle in the big data scientific literature is that an increase in explanatory variables found to be significantly correlated with an outcome variable does not necessarily lead to improvements in prediction. This problem occurs in both simple and complex data. We recently (PNAS, 2015) offered explanations and statistical insights into why higher significance does not automatically imply stronger predictivity and why variables with strong predictivity sometimes fail to be significant. We suggest shifting the research agenda toward searching for a criterion to locate highly predictive variables rather than highly significant variables. We offer an alternative approach, the partition retention method, which can effectively in reducing prediction error rates. Motivated by the needs of current genome-wise association studies (GWAS) we provide a discussion on a theoretical framework. We lay out an objective function for correct prediction rates (the best one can hope for, also as targeted parameters) for which we need to maximize.