Weighted random subspace method for high dimensional data classification

Hongyu Zhao
Yale University
60 College Street
New Haven, CT 06520, USA
hongyu.zhao@yale.edu

Abstract

High dimensional data, especially those emerging from genomics and proteomics studies, pose significant challenges to traditional classification algorithms because the performance of these algorithms may substantially deteriorate due to high dimensionality and existence of many noisy features in these data. To address these problems, pre-classification feature selection and aggregating algorithms have been proposed. However, most feature selection procedures either fail to consider potential interactions among the features or tend to over fit the data. The aggregating algorithms, e.g. the bagging predictor, the boosting algorithm, the random subspace method, and the Random Forests algorithm, are promising in handling high dimensional data. However, there is a lack of attention to optimal weight assignments to individual classifiers and this has prevented these algorithms from achieving better classification accuracy. In this talk, we formulate the weight assignment problem and propose a heuristic optimization solution. We have applied the proposed weight assignment procedures to the random subspace method to develop a weighted random subspace method. Several public gene expression and mass spectrometry data sets at the Kent Ridge biomedical data repository have been analyzed by this novel method. We have found that significant improvement over the common equal weight assignment scheme may be achieved by our method.