Feature Subset Selection for Analyzing Proteomic Mass Spectrometric Data -- A parsimonious threshold-independent feature selection method

張源俊博士
中央研究院統計科學研究所

Abstract

Protein expression profiling for differences indicative of early cancer holds promise for improving diagnostics. Due to their high dimensionality, statistical analysis of proteomic data from mass spectrometers is challenging in many aspects such as dimension reduction, feature subset selection as well as construction of classification rules. Search of an optimal feature subset, commonly known as the feature subset selection (FSS) problem, is an important step towards disease classification/diagnostics with biomarkers. Motivated by this problem, we develop the parsimonious threshold-independent feature selection (PTIFS) method based on the concept of area under the curve (AUC) of the receiver operating characteristic (ROC). Starting from an anchor feature, the PTIFS method selects a feature subset through an iterative updating algorithm similar to LAS. Highly correlated features that have similar discriminating power are precluded from being selected simultaneously. The classification rule is then determined from the resulting feature subset.