Overlapping Group Screening for Binary Cancer Classification with TCGA High-dimensional Genomic Data

Jie-Huei Wang

Department of Mathematics, National Chung Cheng University

Abstract

Precision medicine has been a global trend of medical development, wherein cancer diagnosis plays an important role. With accurate diagnosis of cancer, we can provide patients with appropriate medical treatments for improving patients' survival. Since disease developments involve complex interplay among multiple factors such as gene–gene interactions, cancer classifications based on microarray gene expression profiling data are expected to be effective, and hence, have attracted extensive attention in computational biology and medicine. However, when using genomic data to build a diagnostic model, there exist several problems to be overcome, including the high-dimensional feature space and feature contamination. In this paper, we propose using the overlapping group screening (OGS) approach to build an accurate cancer diagnosis model and predict the probability of a patient falling into some disease classification category in the logistic regression framework. This new proposal integrates gene pathway information into the procedure for identifying genes and gene–gene interactions associated with the classification of cancer outcome groups. We conduct a series of simulation studies to compare the predictive accuracy of our proposed method for cancer diagnosis with some existing machine learning methods, and find the better performances of the former method. We apply the proposed method to the genomic data of The Cancer Genome Atlas related to lung adenocarcinoma (LUAD), liver hepatocellular carcinoma (LIHC), and thyroid carcinoma (THCA), to establish accurate cancer diagnosis models. (This is a joint work with Prof. Chen from the Institute of Statistical Science at Academia Sinica.)

Keywords: Cancer diagnosis; gene–gene interaction; logistic regression; overlapping group screening; precision medicine; TCGA.