Supervised Heterogeneous Multiview Learning

for Joint Association Study and Disease Diagnosis

Shandian Zhe

Dept. of Computer Science, Purdue University, U.S. szhe@purdue.edu

Zenglin Xu*

Dept. of Computer Science, Purdue University, U.S. benjaminxu@purdue.edu

Yuan (Alan) Qi

Dept. of Computer Science & Dept. of Statistics, Purdue University, U.S. alanqi@cs.purdue.edu

Abstract: Given genetic variations and various phenotypical traits, such as Magnetic Resonance Imaging (MRI) features, we consider two important and related tasks in biomedical research: i) to select genetic and phenotypical markers for disease diagnosis and ii) to identify associations between genetic and phenotypical data. These two tasks are tightly coupled because underlying associations between genetic variations and phenotypical features contain the biological basis for a disease. While a variety of sparse models have been applied for disease diagnosis and canonical correlation analysis and its extensions have been widely used in association studies (e.g., eQTL analysis), these two tasks have been treated separately.

To unify these two tasks, we present a new sparse Bayesian approach for joint association study and disease diagnosis. In this approach, common latent features are extracted from different data sources based on sparse projection matrices and used to predict multiple disease severity levels based on Gaussian process ordinal regression; in return, the disease status is used to guide the discovery of relationships between the data sources. The sparse projection matrices not only reveal interactions between data sources but also select groups of biomarkers related to the disease. To learn the model from data, we develop an efficient variational expectation maximization algorithm. Simulation results demonstrate that our approach achieves higher accuracy in both predicting ordinal labels and discovering associations between data sources than alternative methods. We apply our approach to an imaging genetics dataset for the study of Alzheimer's disease (AD). Our method identifies biologically meaningful relationships between genetic variations, MRI features, and AD status, and achieves significantly higher accuracy for predicting ordinal AD stages than the competing methods.

Key Words: Multiview learning, ordinal regression, association study, group selection, Alzheimer's disease