Multi-view Singular Value Decomposition for Disease Subtyping and Genetic Associations

Jiangwen Sun[1], Henry R Kranzler[2], Jinbo Bi[1]

Accurate classification of patients with a complex disease into subtypes has important implications in medicine and healthcare. Using more homogeneous disease subtypes in genetic association analysis will enable the detection of new genetic variants that cannot be detected by the non-differentiated disease phenotype. Subtype differentiation can also improve diagnostic classification, which can in turn inform clinical decision making and treatment matching. Currently, the most sophisticated methods for disease subtyping perform cluster analysis on the basis of patients' clinical features. Without guidance from the genetic dimension, the resultant subtypes can be suboptimal and genetic associations may fail. We propose a novel machine learning approach based on multi-view matrix decomposition that integrates clinical features with genetic markers to detect confirming evidence in the two data sources for a disease subtype. Our approach groups patients into clusters that are consistent between the clinical and genetic dimensions of data, and also simultaneously finds the clinical features that define the subtype and the genotypes that are associated with the subtype. A simulation study validates that the proposed approach indeed identifies hypothesized subtypes and associated features. Using a dataset consisting of 1,474 African American cocaine users, we identified three cocaine use subtypes and their associated clinical variables as well as genetic variants. Moreover, the comparison between our method and the latest multi-view data analytics shows our method can identify genetically more separable clinical subtypes of a disease. In conclusion, the proposed algorithm is an effective and more advanced alternative to the disease subtyping methods employed to date. Integration of clinical/phenotypical features with genetic markers in the subtyping analysis is promising to improve the concurrent validity of identified disease subtypes and their genetic associations.

[1] Department of Computer Science and Engineering, University of Connecticut
[2] Center for Studies of Addiction, University of Pennsylvania Perelman School of Medicine