**Sparse conditional generate adversarial networks (CGAN) for individualized biomarker selection and treatment effect estimation**

Yuanyuan Liu1, Shicheng Guo2, Qiyang Ge1, Steven J. Schrodi2,3, Momiao Xiong1

1Department of Biostatistics and Data Science, University of Texas School of Public Health

2Center for Precision Medicine Research, Marshfield Clinic Research Institute

3Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison

**Abstract**

Next generation genomic, epigenomic, sensing and image technologies produce ever deeper multiple omics, physiological, imaging and phenotypic data with millions of features. Integrating omics, physiological and imaging data provides invaluable information for identification of individualized biomarkers that will be used for estimation of individualized treatment effects and optimal selection of individualized therapy. The classical methods for biomarker identification use average treatment effect information. However, treatment response is heterogeneous. Only using average treatment effect information presents a problem in selecting the optimal treatment for each individual to ensure that the right therapy is offered to “The right patient at the right time.” Unfortunately, estimating individualized treatment effects and design of individualized therapy is beyond the state-of-the-art of the current biomarker selection paradigm. A key issue for individualized treatment estimation is to estimate counterfactuals of treatment. However, counterfactuals are unobserved and are therefore a missing value problem. The classical treatment effect estimation methods cannot accurately estimate the counterfactuals due to lack of methods for missing value estimation. The recently developed conditional generative adversarial nets (CGAN) are accurate tools for estimating the counterfactuals. Imaging, omics and physiological data involve millions of features. Since neural networks (NN) are complicated nonlinear functions, identifying biomarkers from omics and imaging data in CGAN is a challenging task. To identify individualized biomarkers and estimate individualized treatment effects, we combined a novel instance-wise feature selection method that consists of three neural networks: a selection network, a prediction network and baseline network, and CGAN. The algorithms are applied to the TCGA Uterine Corpus Endometrial Carcinoma (UCEC) dataset with histological imaging, gene expression (mRNA and miRNA), methylation data and 145 chemotherapy response records from 67 unique individuals. Real data analysis results show that the proposed algorithms substantially outperform the state-of-the-art methods. We find our prediction model has strong predictive performance on mislabeled drug response (N=134) compared with overall survival (OS).