PittComputing&Information

Dissertation Defense

Doctor of Philosophy in Intelligent Systems

"**Deep Learning for Causal Structure Learning Applied to Cancer Pathway Discovery**"
by **Jonathan D. Young**

Date: February 20, 2020
Time: 10:00 – 12:00
Place: Room 536 B, 5607 Baum Boulevard,
Pittsburgh PA 15206

Committee:

- Dr. Xinghua Lu, Professor, Intelligent Systems Program
- Dr. Gregory Cooper, Professor, Intelligent Systems Program
- Dr. Vanathi Gopalakrishnan, Associate Professor, Intelligent Systems Program
- Dr. Harry Hochheiser, Associate Professor, Intelligent Systems Program
- Dr. Songjian Lu, Assistant Professor, Department of Biomedical Informatics

Abstract:
In general, the cellular mechanisms leading to cancer in an individual are heterogeneous, nuanced, and not well understood. It is well appreciated that cancer is a disease of aberrant signaling, and the state of a cancer cell can be described in terms of abnormally functioning cellular signaling pathways. Identifying all of the abnormal cellular signaling pathways causing a patient's cancer would enable more patient-specific and effective treatments—including targeting multiple abnormal pathways during a treatment regime. Here we interpret the cellular signaling system as a causal graphical model and apply a modified deep neural network (DNN) to learn latent causal structure that represents the cancer cellular signaling system.

Most causal discovery algorithms have been developed to find causal structure and parameterizations of causal structure relative to the observed variables of a dataset. A smaller number of casual discovery algorithms also find latent causal structure, but these methods are often highly constrained or less applicable to the problem explored here, suggesting that new methods are needed. In this dissertation, we address a problem for which it is known that a set of variables *X causes* another set of variables *Y* (e.g., mutations in DNA *cause* changes in gene expression), and these causal relationships are encoded by a causal network among a set of an unknown number of latent variables. We develop a modified deep learning model, referred to as redundant input neural network (RINN), with an L_1 regularized objective function to find causal relationships between input (*X*), hidden, and output (*Y*) variables. More specifically, our model allows input variables to directly interact with all latent variables in a neural network to influence what information latent variables encode in order to generate the output variables accurately. In a series of simulation experiments, we show that the RINN model successfully recovers latent causal structure from various simulated datasets with different levels of noise better than other models.

We hypothesize that training a RINN on multiple omics data will enable us to map the functional impacts of genomic alterations to latent variables in a deep learning model, allowing us to discover the hierarchical causal relationships between variables perturbed by different genomic alterations. We apply the RINN to cancer genomic data, where it is known that genomic

alterations cause changes in gene expression. We show that differentially expressed genes can be predicted from somatic genome alterations with reasonable AUROCs by a RINN (or DNN). We also show that a RINN is able to discover many real cancer signaling pathway relationships, especially relationships between genes in the *PI3K*, *Nrf2*, and *TGFβ* pathways, including some causal relationships. In this setting, the connections between input and latent variables make the latent variables partially interpretable, as they can be easily mapped to input space. However, despite relatively large levels of regularization, the returned causal graphs were still somewhat too dense to be easily and directly interpretable as causal graphs. Future versions of the RINN, with differential regularization, autoencoder pre-trained representations, and optimization with parallelized and constrained evolutionary algorithms, will have a high probability of capturing more easily interpretable cancer pathways.