Abstract

Autoimmune disorders (ADs) are a prevalent and growing concern. Despite the increasing number of cases each year, the majority of genes underlying these diseases are unknown. While efforts are being made to treat and develop therapies for these particular disorders, the methods for target discovery and identification are time and cost expensive. We propose a novel model to predict gene-autoimmune disease associations with Graph Neural Networks (GNNs).

Methods

In order to predict gene-autoimmune disease associations, we constructed a heterogeneous graph using a graph convolutional framework, with genes and autoimmune diseases as nodes and three types of edges representing their interactions. The data is sourced from the publicly available Gene & Autoimmune Disease Association Database (GAAD).

We load in 66 unique disease nodes, 4051 unique gene nodes, 183,346 gene-gene associations, 907 disease-disease associations, and 3295 known gene-autoimmune disorder associations from our datasets. Each disease node’s features are based on their known pathophysiological pathways, and each gene’s features are based on the number of known associations to ADs.

We construct our graph using the Heterogeneous Data class available through the PyTorch library. Each node and edge is assigned a unique index for identification, and the connections are stored through tensors.

Following a convolutional framework, we update a node’s features using an aggregation of its neighbor’s features. For each hidden layer, the ReLu function is applied as the activation function.

\[ x_i^t = W_l x_i + W_e \cdot \text{mean}_j cN(i)_j \]

We apply a final classifier function as the dot product between the vector of disease and gene node embeddings to obtain a measure (the prediction) of their relative “closeness”.

Within the training loop, we calculate the loss for our model.

\[ L = \frac{1}{N} \sum_{i=1}^{N} \left( \log(p_i) + (1-t) \log(1-p_i) \right) \]

Let \( N \) be the number of data points, \( t \) be the truth value of the edge, and \( p \) be the prediction given by the model.

Results

Measurements of performance:

\[
\begin{align*}
\text{Recall} & = \frac{TP}{TP + FN} \\
\text{Precision} & = \frac{TP}{TP + FP} \\
\text{Accuracy} & = \frac{TP + TN}{TP + TN + FP + FN}
\end{align*}
\]

Area Under the Receiving Operator Curve (ROC),

- The ROC curve is a metric to evaluate how well our model can distinguish between an edge versus no edge.

Conclusion & Future Directions

We built a heterogeneous graph neural network with a convolutional framework which accurately predicts novel gene-autoimmune disease associations. We discovered that increasing the amount of data points improves the model significantly. We hope to extend our model to include even more genes which we can relate to our 67 autoimmune disorders. We would like to benchmark our model in order to compare it to existing heterogeneous GNNs.

Acknowledgements

This work was funded by the National Science Foundation through Research Experience for Undergraduates (REU) award, with additional support from the National Institute of Computational Sciences at the University of Tennessee, Knoxville. In addition, we would like to thank Dr. Kwai Wong for his mentorship and advice.

References: