Abstract

Millions of people each year die from cancer worldwide. Understanding the underlying biological mechanisms that result in some individuals outliving others can help us produce better treatment strategies. Machine learning, specifically neural network-based, models can help identify otherwise hard-to-identify patterns in patient data. Graph Neural Networks (GNNs) allow the development of predictive models based on known biological network data. To aid in the development of such models, this project aims to generate an example dataset that integrates Pathway Commons (biological network) and cbioPortal (cancer patient data) for use with the popular PyTorch Geometric (PyG) library for the prediction of cancer patient overall survival. We provide scripts for processing similar datasets and example code to showcase how models can be generated. We hope this work will aid researchers in unlocking valuable insights into the role of biological pathways and genetic variation in cancer to help patients.

Results and Discussion

Table 1 describes the resulting datasets used for model development after conducting data preprocessing and converting the datasets into PyG data objects:

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>ACC</th>
<th>BRCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc_tcga_2018</td>
<td>FLAML</td>
<td>FLAML</td>
</tr>
<tr>
<td>brca_tcga_2018</td>
<td>FLAML</td>
<td>FLAML</td>
</tr>
</tbody>
</table>

In order to ensure reproducible notebooks, the final processed PyG datasets, as well as unprocessed data, were uploaded to Zenodo, a general-purpose data repository meant for long-term storage of academic datasets.

Preliminary results in Table 2 demonstrate that the GNN model performs similarly to the FLAML-based model for the smaller ACC model in predicting overall survival time. By contrast, the BRCA GNN model outperforms the BRCA FLAML model. We believe this observation between the models is based on the larger BRCA dataset. Note that given small size of ACC dataset, it was split to only train and test sets. However, for the BRCA data, a validation split was included. During the development of the baseline models, only train and test splits were used. This was done to fit the requirements of FLAML.

Conclusions/Future Work

GNNs have only in the recent few years been applied to the challenge of understanding the survival of cancer patients. To aid in this area of research, we have produced example material for dataset generation and analysis utilizing large, publicly-collected datasets using both GNN and AutoML techniques.

The preliminary models we have produced help us begin to understand issues of necessary sample size and overfitting to be addressed with future work. We believe GNNs can help researchers to understand complex relationships observed in cancer biology and be of use in personalized medicine research. Further exploration is also planned with Graph Attention Networks (GATs) techniques to aid in model interpretability.

References


Acknowledgements

We would like to thank the Google Summer of Code program for funding and the National Resource for Network Biology (nrb.org) for overall coordination.

Project Repository