Recreation of Biologically Informed Deep Neural Network for Prostate Cancer Discovery

Noah Parker - np2833

Fu Foundation, Electrical Engineering, Columbia University, NYC

Abstract

Motivation: This project sought to recreate the results of the paper Biologically Informed Deep Neural Network for Prostate Cancer Discovery by Elmarakeby et. al.
Results: The model failed to reconstruct the results found in the original paper, over-classifying all samples as not being indicative of prostate cancer.
Availability: https://github.com/NJParker415/ECBM-4060-Final-Project

1. Introduction

Determination of which biological features correspond to the presence of prostate cancer remains a major issue in bioinformatics. The advent of machine learning models has the potential to greatly improve prostate cancer identification. This project attempted to recreate the paper Biologically Informed Deep Neural Network for Prostate Cancer Discovery (Elmarakeby et. al.) which sought to create a neural network with a design inspired by extant biological systems. The neural network itself is a deep network with a progression modeling increasing pathway complexity; the molecular profile of a patient is fed in, and the layers of the network represent abstractions of genes and pathways with increasing complexity. In the paper, this design was chosen under the idea that a neural network structured after an extant biological hierarchy would be able to better predict the presence of cancers than less specialized, more abstracted traditional models.

2. Approach

The same dataset and preprocessing methods were used as in the original paper, with the model itself being constructed using Pytorch.

3. Methods

3.1 Data Preprocessing

Information from the original paper was used to construct pathway mappings, which were in turn used to inform the structure of the neural network. Information curated from the reactome database by the paper authors was used to construct the connections between layers of the neural network informed by these pathway mappings.

This mapping was performed by using code from the original paper in order to construct a network linking genes and pathways, and then “slicing” it into layers interpretable by the model builder.

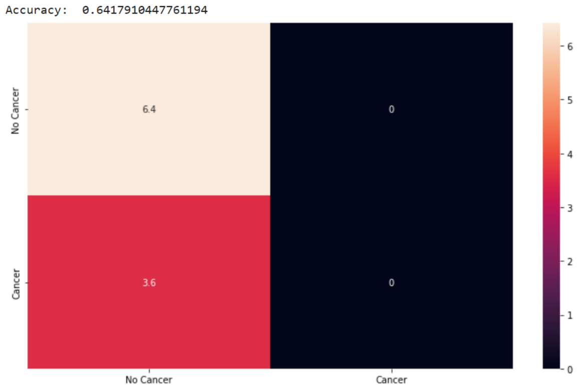
3.2 Model Construction

The model constructed was built according to the parameters identified within the original paper’s codebase as being those used to produce its final results. The model architecture was composed as follows:

- A fully connected layer, linking tpm data to nodes representing genes, followed by a tanh activation function.
- 5 occurrences of the following sequence
    - A fully connected layer, masked using the mapped pathways to only allow weights corresponding to connections along pathways.
    - A tanh activation function
    - A dropout layer with dropout probability equal to 0.5
- A final fully connected layer collapsing all inputs to a single value
- A sigmoid activation function

The model used binary cross-entropy to compute loss, the Adam optimization algorithm with a learning rate of 0.001, and was trained for 100 epochs.

4. Results



```
Accuracy:  0.6417910447761194
```

Figure 1: Confusion matrix showing model predictions on the test set.

No produced model was able to achieve the same results as found in the original paper. All models produced invariably over-classified all inputs as not indicating the presence of prostate cancer.

5. Discussion

While an attempt was made to faithfully re-create the model of the original paper, several modifications were required in order to produce acceptable results. Due to the original paper’s codebase being written in Python 2 with a Keras backbone, combined with the author’s unfamiliarity with Keras, it is likely that several nuances present in the original model were neglected.

Additionally, limitations were imposed by the scope of this project. Due to time constraints the model was unable to be trained for as long as that in the paper. In order to obtain results for iteration in a reasonable amount of time, the model in this project was only trained for a third of the epochs as in the original paper.

The fact that the model classified all samples as not indicating the presence of cancer may also be a result of the amount of training data; compared to most other machine learning problems, there was a vanishingly small amount of training data present. This almost certainly leads to over-fitting, as the training data present may not be sufficient to accurately train a model.

6. Conclusion

Absent the presence of the original paper's conclusions, the model produced in this project would suggest a trivial representation of prostate cancer from the input data. However, the fact that the original paper was able to detect a correlation between genomic data and the presence of prostate cancer indicates that the model produced in this project was simply not able to adequately detect the connection. Further experimentation is needed to identify the disconnect between the implementation of this project's model and that of P-Net.

7. References

Elmarakeby, H.A., Hwang, J., Arafeh, R. *et al.* Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 348–352 (2021). https://doi.org/10.1038/s41586-021-03922-4