Summer Project 2020

(Duration : Jan 13 - Feb 14)

Prediction of Glioblastoma from mRNA profiles of TEPs using transfer learning approach

- Abhishek Vijayan
(Under guidance of Dr Fatemeh Vafaee)

Task : Classify GBM(brain cancer) Vs Healthy, from mRNA profiles of TEPs (Tumour Educated Platelets) using transfer learning approach.

The relevance of GBM prediction from mRNA profiles of TEPs is that, it can be obtained from liquid biopsy thereby enabling minimally invasive molecular diagnostics.
The transfer learning approach is proposed here because GBM is a rare disease and hence there are very few samples available. The idea is to use the relatively more abundant NSCLC(Non-Small Cell Lung Carcinoma) data to support prediction of GBM.

Reference Papers

(used currently for obtaining data)

- RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics (2015 Best et al)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4644263/#!po=32.1429
- Swarm Intelligence-Enhanced Detection of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets (2017 Best et al)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6381325/#!po=77.0833

Data from 2015 paper contains mRNA read counts of 285 samples with 57736 genes. It contains 6 cancer types (one of which is GBM) and healthy control
Data from 2017 paper contains mRNA read counts of 779 samples with 4722 genes. It contains NSCLC and healthy control

Pipeline

1. Preprocessing
    a. Data filtering and normalization
        Used R Bioconductor EdgeR package
2. Predictive Models
    a. Data transformation
    b. Model training
    c. Evaluation
    Used python & python packages - pytorch, scikit learn

Preprocessing

- Generated 2 sets of filtered data
- Set 1
    - On 2015 dataset, filtered out only GBM and healthy control
    - On both datasets separately applied
        i. filterByExpr
        ii. calcNormFactors of DGEList
        iii. Using 2 groups - Cancer and Non-Cancer, applied voom
- Set 2
    - Same as previous till filterByExpr
    - filter out common genes from 2015 and 2017 data
    - Perform step 2, 3 on this common gene sets

Data Size

	Original Data	**Set 1**	**Set 2** **(found 2708 common genes)**
2015 data (for GBM)	57736 x 285	57736 x 95 (only GBM and HC) 3368 x 95	57736 x 95 (only GBM and HC) 2708 x 95
2017 data (for NSCLC)	4722 x 779	3067 x 779	2708 x 779

	Cancer	**Non Cancer**
2015 data (for GBM)	40	55
2017 data (for NSCLC)	402	377

Models

- Baseline models
    - SVM
    - Logistic Regression

- 3 layer NN (3368-33-1)
- 5 layer NN (3368-1000-100-10-1)
- Using only common genes : 3 layer NN (2708-27-1)
- Using only common genes : 5 layer NN (2708-1000-100-10-1)

- 2 basic transfer learning models
        - using the common gene set, pretraining on NSCLC data and then using GBM data
    - 3 layer NN (2708-27-1)
    - 5 layer NN (2708-1000-100-10-1)

Results

Accuracy (Mean over 30 iterations)

	Logistic Regression	SVM	3 layer NN	5 layer NN	Common genes : 3 layer NN	Common genes : 5 layer NN	Transfer Learning : 3 layer NN	Transfer Learning : 5 layer NN
Mean	0.926	0.889	0.884	0.914	0.837	0.879	0.867	0.877
Standard Deviation	0.061	0.074	0.065	0.056	0.065	0.071	0.049	0.082

AUC (Mean over 30 iterations)

	Logistic Regression	SVM	3 layer NN	5 layer NN	Common genes : 3 layer NN	Common genes : 5 layer NN	Transfer Learning : 3 layer NN	Transfer Learning : 5 layer NN
Mean	0.979	0.952	0.935	0.958	0.905	0.94	0.918	0.921
Standard Deviation	0.032	0.048	0.074	0.051	0.075	0.039	0.07	0.089

Observation

Considering the accuracy values,

- Logistic regression is significantly better than 3 layer NN (p = 0.0126) and non-significantly better than 5 layer NN (p = 0.4127).
- 5 layer NN performs better than SVM (p = 0.1514), but the improvement is not statistically significant.
- On considering only the common genes, there seems to be a significant decrease in accuracy for both 3 layer NN (p = 0.0068) and 5 layer NN (p = 0.0534). This indicates that better methods for obtaining common set of features should be explored further.
- Performing transfer learning showed non-significant increase in 3 layer NN (p = 0.6228) and non-significant decrease in 5 layer NN (p = 0.9392). Transfer Learning approach has to be further improved, to obtain reasonable increase in accuracy.

Considering AUC values,

- Logistic Regression is significantly better than 3 layer NN (p = 0.0043) and 5 layer NN (p = 0.0584)
- 5 layer NN performs non-significantly better than SVM (p = 0.6888)
- On considering only common genes, there is a non-significant decrease for both 3 layer NN (p = 0.1262) and 5 layer NN (p = 0.1941)
- Performing transfer learning showed non-significant increase in 3 layer NN (p = 0.8642) and non-significant decrease in 5 layer NN (p = 0.6897)

Further Steps

- Current pipeline a very basic one - improvement required in all steps
- Better filtering and normalization
    - Use the supplementary data provided
    - Comparison among different filtering / normalization methods
- Perform NN hyper parameter tuning
- Include dropout, regularization in the network
- Try out more complex NN models for prediction
    - Currently CNN has not been considered since CNN is commonly used in image recognition to identify useful features relevant in multiple parts of the image. This kind of structure among features could not be seen in genes. This is something to be further explored.
- Better approaches for doing transfer learning
    - Currently only pre-training has been done, try out fine tuning i.e. training only the last few layers
    - To get equal num of features for NSCLC and GBM data try
        - PCA
        - UMap
        - Variable input autoencoder
    - Try to use non-GBM cancer samples from 2015 data, for transfer learning
- Try out simulations to get more data