


Feature Selection Using Deep Neural Networks in Bioinformatics



Naresh Kumar Kaushal



Why we have chosen this problem statement

Cancer a fatal disease

Cancer is one of the life threatening disease which humans have not yet conquered.

Numerous methods are proposed to detect cancer and its subtypes at an early stage using various technologies and software tools but most of them suffer from the problem of unavailability of sufficient data. We are trying to develop methods and techniques to tackle this low sample size problem in deep neural networks.



Summary of DNP (deep neural pursuit) algorithm

I chose to work on DNP algorithm which tackles the problem of low sample size. It has 4 main phases:

There are 2 sets S and C set and initially S contains only bias and C contains all the features.

1. Train Phase - Train the subnetwork and learn the weights and use this trained subnetwork as input in next iteration (boosting).
2. Select Phase - Dropout multiple times and take average to select new features based on norm of gradients.
3. Update Phase - Update the learning rates according to Adagrad.
4. Transfer Phase - Transfer the newly selected feature from the C set to S set.

Datasets

I have chosen 4 cancer datasets to work with
i.e. Breast, Prostate, Leukemia and Lung.

All these datasets have 2 tables one with gene expression values and other with subtype information.

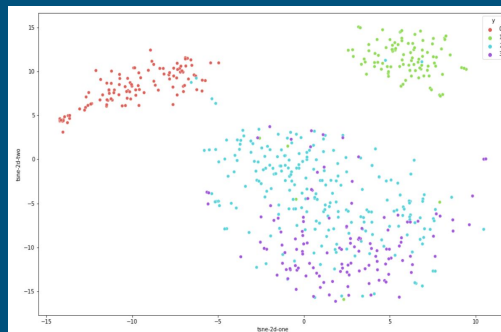
Breast Cancer subtypes - Luminal A, luminal B, Basal.

Lung Cancer subtypes - Non small cell, small cell, lung carcinoid, metastatic, Mesothelioma.

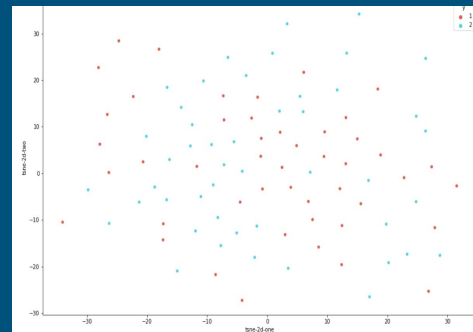
Prostate Cancer subtypes - Acinar adenocarcinoma, ductal adenocarcinoma.

Leukemia - Acute Lymphocytic Leukemia(ALL), Acute myelocytic leukemia(AML).

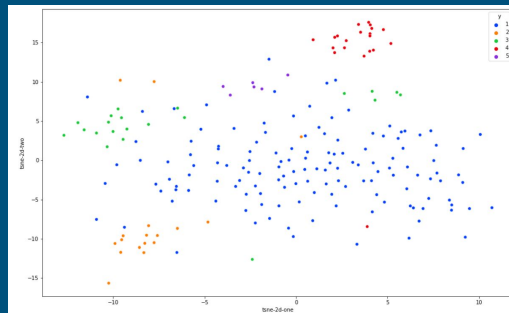
I am using mRNA-Seq based expression values.



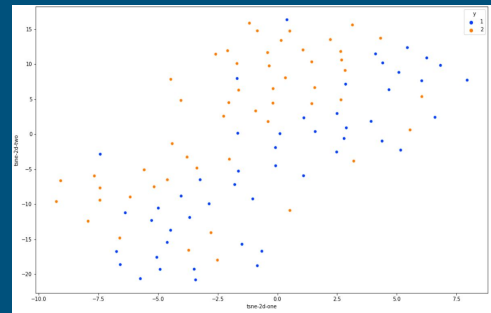
Breast Cancer t-SNE plot



Leukemia t-SNE plot



Lung Cancer t-SNE plot



Prostate Cancer t-SNE plot

Subtype Cardinality

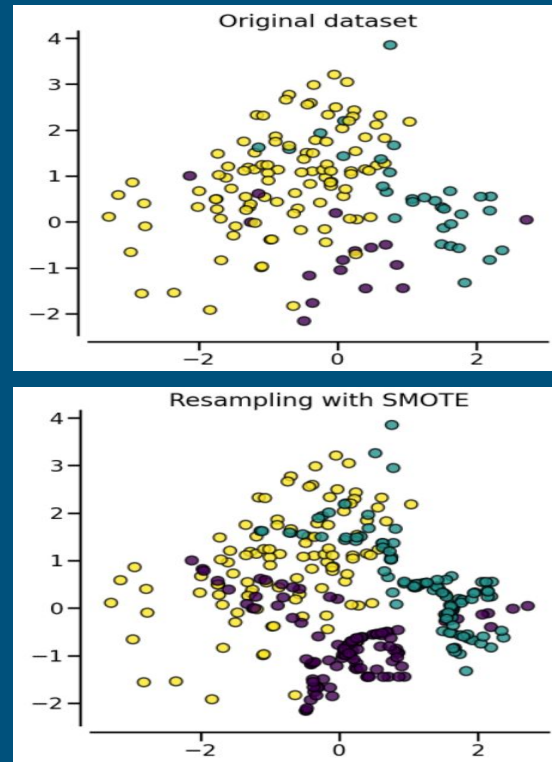
Subtype Cardinality					
Dataset	subtype_1	subtype_2	subtype_3	subtype_4	subtype_5
Breast Cancer	113	101	233	127	-
Prostate Cancer	50	52	-	-	-
Lung Cancer	139	17	21	20	6
Leukemia Cancer	47	25	-	-	-

Lung and Leukemia suffers from class imbalance problem !

SMOTE (Synthetic Minority Oversampling Technique)

A technique to tackle class imbalance problem

All 4 datasets have significant variability but special attention needs to be given for Lung Cancer and Leukemia. Both these datasets suffer from class imbalance problem and using SMOTE we can generate synthetic data points in our train set using interpolation.



Comparison of DNP and PCA.

Deep Neural Pursuit			
Dataset	accuracy	auc score	f1_score
Breast Cancer	77.6	90.8	77.0
Prostate Cancer	81.8	88.4	81.8
Lung Cancer	87.14	98.49	80.75
Leukemia Cancer	85.45	83.33	87.91
Principal Component Analysis with Classification Trees			
Dataset	accuracy	auc score	f1_score
Breast Cancer	79.95	85.32	79.92
Prostate Cancer	80.36	80.25	79.14
Lung Cancer	88.67	87.50	87.36
Leukemia Cancer	81.96	83.02	84.47

Why PCA is not good for HDLSS...

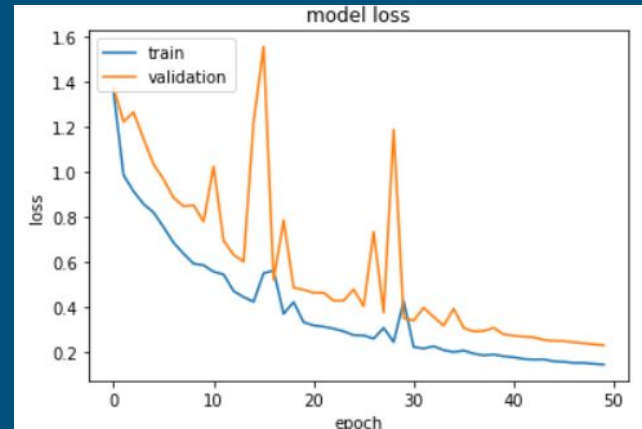
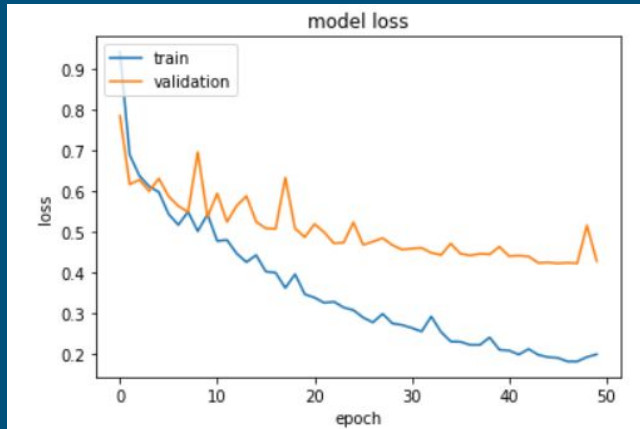
Stability is one such factor where PCA fails in comparison to DNP.

PCA with Classification Trees										
Dataset	fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	fold10
Breast Cancer	87.93	82.75	81.03	72.41	77.19	78.94	77.19	78.94	82.45	80.7
Prostate Cancer	90.9	63.63	60.0	90.0	90.0	70.0	100.0	80.0	70.0	60.0
Lung Cancer	85.29	91.17	89.55	-	-	-	-	-	-	-
Leukemia Cancer	75.0	87.5	100.0	85.71	85.71	71.42	57.14	85.71	85.71	85.71

Comparison of DNP and simple DNN

Overfitting is one such factor where simple DNN fails in comparison to DNP.

If we use simple DNN on HDLSS data then our model will suffer from severe overfitting and it will not give stable results. Below 2 are the loss curves for Lung and prostate Cancer.



Improvements..

DNP takes huge amount of time as compared to PCA or simple Deep Neural Network because the bottleneck is **Selection** phase. We can improve this phase by first reducing dimensions using PCA or some other dimensionality reduction technique (t-SNE) and then selecting features based on DNP but this method needs to be verified for its stability. I have explored 4 types of cancers with significant variability and now next step will be to analyse the hyper parameters and model architecture properly so as to get a more generalised DNP model.

My contributions.....

From data cleaning to comparison of DNP with PCA and simple DNN

1. Searched datasets on TCGA website with mRNA-seq expression values.
2. Presented DNP algorithm research paper.
3. Preprocessed the datasets and tackled with class imbalance problem.
4. Implemented DNP algorithm and tested it on 3 cancer datasets Breast, Prostate and Lung.
5. Compared DNP algorithm with PCA and simple DNN.
6. Proposed possible improvements in DNP to reduce its time complexity and increase its accuracy.

Thanks

Supervisor: Dr. Clint P. George