Proteogenomics of Breast Cancer

Capstone Proposal - Machine Learning Engineer Nanodegree

Andra Tolbus
October 19, 2019

Domain Background

Breast Cancer is a the most common type of cancer among women worldwide and the second most common type of cancer overall, causing death in 672 000 women in 2018 alone [1][2]. Thus, early screening procedures as well as better understanding of the breast cancer mechanisms at a molecular level is crucial in order to increase the overall survival in these patients.

Laboratory techniques like microarray and recently DNA , Protein and RNA sequencing of tissue samples collected from patients, together with clinical information on the individual patients, give us the opportunity to explore the differences in molecular signatures which will eventually lead to personalised treatment.

I have personally chosen to explore this topic as it complements my existing knowledge in bioinformatics, as I have previously worked with understanding microRNA mechanisms within cardiovascular diseases. Moreover, I have not had the chance to work on cancer related projects in the past and the domain knowledge is highly relevant as I currently work in the Healthcare industry, where Oncology is a high focus area. Therefore, I believe that applying ML methods in this domain would be engaging, interesting and highly useful at the same time.

Problem Statement

I are trying to explore a couple of possible questions, in a step-wise manner

A) Can we use gene expression data and clinical information to divide the cancer patients in sub types (unique molecular signatures) ?

B) Can we we use gene expression data (and clinical information) to predict the cancer subtype?

Datasets and Inputs

The datasets selected on the analysis have been generated by the Clinical Proteomic Tumour Analysis Consortium and have been made available on Kaggle [3] , “Breast Cancer Proteomes” .

The data has been originally used in a Nature published study (http://www.nature.com/nature/journal/v534/n7605/full/nature18003.html) where the researchers have investigated how proteogenomics can be used to link mutations to breast cancers’ molecular signatures and how the known cancer subtype classification is reflected by the clusters identified in an unsupervised manner from the expression data.

There are 3 datasets :

- **Clinical Data -** patient characteristics and attributes related to the diagnosis (age, sex, metastasis, presence of lymph-nodes , tumour size , tumour shape , etc) . Total number of samples : 105.

Reference file: *clinical_data_breast_cancer.csv*

- **Protein expression data** - by row, protein id, gene symbol, gene name and expression levels for 83 patients, with the last 3 columns belonging to healthy individuals. This data set can be linked to the clinical data set through the TCGA (included in the column names for each of the samples)

Reference file: *77_cancer_proteomes_CPTAC_itraq.csv*

- **PAM50 protein data -** official list of genes and proteins according to the PAM50 classification system. The expression data can be enhanced with additional information from this data set by using the RefSeq ids.

We will use this data set to answer the same problem as initially investigated in the above mentioned study (Problem Statement A) , but will also go one step further to try to understand whether we can also determine the stage of breast cancer (Problem statement B)

Solution Statement

After getting a thorough understanding of the data through descriptive statistics by using the appropriate libraries, the solution for the two problem statements would be as follows:

- Solution Statement A - a clustering model where we are able to identify cancer subtypes ; this will be a simulation of the work done in by the research group. They used K-means for the task, but will aim of looking at one or two additional clustering algorithms to see if we can get better results.

- Solution Statement B - a classification model able to predict breast cancer type potentially using clinical data features (if relevant) in addition to protein expression levels .Three classification models will be explored. Initial thinking (candidates): AdaBoost , SVM, XGboost .

In both cases (A & B) , we will only work with the patient data which can be retrieved from both datasets . According to the Nature paper, several samples were contaminated, so we will select only the valid data.

In order to ensure the best possible model among the ones selected , I will experiment with feature selection (to optimise the number of features included) and will tune the parameters of these models accordingly.

By looking at the datasets, the data included in the patient clinical information is already quite well categorised so I do not expect to have to perform a significant amount of feature engineering. It could be that for some of the numerical variable such as age it might help, so I will certainly look into that. Depending on the outcome of Solution A, we could potentially use the identified molecular signatures via clustering as an input in the classification model. Alternatively, considering that he expression levels here are linked to particular genes, we can look into whether it makes sense to bin these log values for the in relevant intervals .

Benchmark model

- Benchmark model A: I will use the results as provided in the peer reviewed Nature paper where they identified 3 distinct molecular signatures.

- Benchmark model B: a dummy classifier, F1 score for this classifier to be provided in the final report.

Evaluation Metrics

For the clustering task I will be looking at the silhouette and homogeneity scores.

For the classification task, I will be evaluating the performance using the F-score and accuracy.

Project Design

Programming language : Python 3.7

Libraries: pandas, numpy, scikit-learn, AdaBoostClassifier, SVC, XGBoost.

Workflow

- Exploratory data analysis in order to get a better understanding of the data
- Data preprocessing as needed.
- For the clustering task, feature transformation as needed (eg PCA) ; implementing the two selected algorithms (K-means and Ensemble Learning or GMM) . Once satisfied with the results, compare with the learnings from the Nature paper.
- For the classification task, feature transformation and selection as needed; splitting the data between training and testing ; implementing the 3 suggested algorithms and evaluating the performance using the proposed evaluation metrics. Fine tune the models' hyper-parameters to see if we can increase performance. Finally select the best model for the task.
- Collect conclusions regarding the identified unique molecular signatures and how do they relate to the known cancer subtypes as well as the possibility to predict breast cancer subtypes using protein expression data.

References

[1] https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/
[2] https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics
[3] https://www.kaggle.com/piotrgrabo/breastcancerproteomes