Machine Learning Classifier to

Predict Liver and Pancreatic Carcinoma

Aparajita Choudhury, Anirudhan J. Rajagopalan

Courant Institute of Mathematical Sciences, New York University, New York – 10012, USA

Emails: ac5901@nyu.edu, ajr619@nyu.edu

Abstract: *Classification of cancer is an important task in proper diagnosis of the disease and subsequent treatment. The current medical procedures are not sufficient to deal with cancer identification quickly. Since DNA methylation and protein concentration of normal and cancerous cells are different, identifying the patterns in protein expression profile and DNA methylation profile will prove to be advantageous in dispensing quick medical advice and treatment. It can also help in suggesting biomarker targeted treatment to the patient in the case when conventional treatments are not very effective. In this project we have constructed predictive models using machine learning techniques to classify liver and pancreatic carcinoma based on protein expression data and DNA methylation profile and have found that one of our classifiers perform with an accuracy of over 95%.*

Keywords: *Machine Learning, Classifier, Predictive model, Liver carcinoma, Pancreatic carcinoma*

1. Introduction

In recent times cancer has been one of the highest causes of death worldwide leading to an increase in cancer research. Cancer is caused by changes in the gene that affect the functionality and division rate of a cell. Cancer cells divide continuously leading to a huge mass known as a tumor. Cancer cells differ from the normal cells in many ways like – in the ability to perform specialized function, gene expression, protein expression, DNA methylation and other biomarkers. Studies have shown that DNA methylation profile is different between cancer and normal cells and more recently it has been inferred that many cancer cells have a higher level of concentration of some proteins. As the treatment of cancer using traditional pathological process has not been much of a success, there has been a growth in attempts to use genetic information that can give leads to gene targeted treatment to cure cancer. With the growth of technology, it has become much easier to process huge sequences of genes, proteins, DNA methylation. Given the huge amount of data and its inherent mathematical structure, the problem makes itself an ideal candidate for applying machine learning algorithms to. In this project we have classified pancreatic adenocarcinoma (PAAD) and hepatocellular carcinoma (HCC) on the basis of their DNA methylation profile and protein expression. We have used supervised as well as unsupervised learning algorithms on the data to classify the type of cancer and show that DNA

methylation and protein expression can be used as classifiers efficiently by dimensionality reduction using Principal Component Analysis (PCA).

2. Data Processing

2.1. Aim

In this project we aim to compare different machine learning algorithms as classifiers of cancer. We used K-Means to verify that our dataset can be well separated and then compared the predictive performance of logistic regression, linear SVM, Gaussian SVM, AdaBoost and Gaussian Naïve Bayes in predicting hepatocellular carcinoma and pancreatic adenocarcinoma.

We also compare the computational efficiency and resource utilization of the various algorithms used and the two biomarkers we use – DNA methylation and Protein expression.

2.2. Data Collection and processing

We collected the data from The Cancer Genome Atlas (TCGA):

2.2.1 DNA Methylation

The raw DNA methylation data collected from TCGA is ~27GB for liver cancer and ~10GB for pancreatic cancer. The data consisted of patients, tumor types and the β- values for each probe, where a probe corresponds to a gene and the β- values were assayed using Illumina HumanMethylation450 Beadchip.

$$\beta = \frac{Methylated\ Intensity}{Methylated\ Intensity + Unmethylated\ Intensity} + 100$$

The number of samples (patients) available are 430 for liver cancer and 195 for pancreatic cancer and the number of features for both are 485577.

We processed the data into a format that is suitable for applying machine learning algorithms to – the format consisted of a high dimensional matrix of the samples and the features and their corresponding β- values. This step was performed using Python and Numpy. After processing and combining the data for both the cancer types, the data was around 4GB which we loaded into a numpy array. We performed imputation of the data for missing values using mean.

2.2.2 Protein Expression

The raw Protein expression data from TCGA is around 1GB each for liver and pancreatic cancer with the data consisting of proteins and their corresponding concentration in the cells.

For liver cancer the number of samples available are 184 with 193 features and 123 samples with 192 features for pancreatic cancer.

Similar to DNA methylation dataset, we processed this data too into a format suitable for applying machine learning algorithms to– the format consisted of a matrix of the samples and the features (proteins) and their corresponding protein concentration. This step was performed using Python and Numpy. After processing and combining the data for both the cancer types, the data was around 1MB which we loaded into a Numpy array. We then performed imputation of the data for missing values using mean.

2.2.3 Technical Challenges

We faced a few challenges to get the data into the right format. We had started with NYU HPC to process the data but even the HPC cluster insufficient to process it as we had to load the entire data into memory. Eventually we shifted to using NYU Crunchy machines for our processing and then eventually had to shift our processed data from Crunchy machines back to HPC clusters to run our algorithms as there is a contention problem on Crunchy systems while CPU utilization is better on the HPC clusters.

Secondly, our protein expression data did not have the same set of proteins across liver and pancreatic cancer. And as mentioned earlier we transformed the protein data into a matrix format consisting of the samples and the features and their corresponding protein concentration but the raw protein concentration data did not have the sample information. To derive this mapping we had to download the clinical data for both liver cancer (~101GB) and pancreatic cancer (~56GB) that consisted of XMLs from TCGA. Each XML corresponded to a patient and had a "bcr_shipment_portion_uuid" tag that corresponded to one protein expression file. We parsed this XML to get the mapping between the patients and the protein expression.

Lastly, we had an unbalanced dataset and we were not very keen on balancing the dataset by leaving out samples from any cancer type as in practical scenarios it is not always that we get balanced data. So we worked on the unbalanced dataset and also our sample size was much lesser than the number of features especially for DNA methylation.

3. Predictive Models and Algorithms

3.1. Unsupervised algorithm

We used K-Means using 2 and 4 centroid to cluster the samples. K-Means does the clustering around the centroids by minimizing the within-cluster sum of squares. It is usually a fast algorithm and converges quickly but has the disadvantage that it might fall into the local minima.

3.2. Dimensionality Reduction algorithm

For DNA methylation we had a feature size of 485577 and it is not always computationally feasible or efficient to work on such a huge dimension. Moreover, in the biological domain usually there are many features that are irrelevant and features that are highly correlated with each other and thus do not contribute much to the learning curve. To get an idea of the number of features corresponding to a particular explained variance we performed Principal Component Analysis (PCA). The number of principal components selected for a few explained variance for both DNA methylation and protein expression are as shown in Table 1 and Table 2 respectively.

Explained Variance (%)	Number of Principal Components
99	513
95	361
90	255
77	92
65	20
50	4

Table 1: Explained Variance and corresponding number of PCs for DNA Methylation

Explained Variance (%)	Number of Principal Components
99	116
95	66
90	44
77	22
65	12
50	6

Table 2: Explained Variance and corresponding number of PCs for Protein expression

3.3. Boosting algorithms

Boosting algorithms are supervised ensemble meta-algorithm that work by combining weak classifiers into a strong one through multiple iterations. They are usually used when the number of samples is much less than the number of features, which is the case in our data of DNA methylation. We used AdaBoost as the ensemble algorithm. On each iteration AdaBoost adjusts the weights of misclassified instances to build a strong classifier.

3.4. Supervised prediction algorithms

3.4.1. SVM

SVM works well on a high dimensional space but has the disadvantage that it is computationally expensive. There are variations of SVM and we have used Linear SVM and Gaussian SVM for prediction on our datasets.

The log2C values that we used for Linear SVM range from $2^{-9}\ to\ 2^{9}$.

Both the log2C and log2gamma values that we used for Gaussian SVM range from $2^{-9}\ to\ 2^{9}$.

3.4.2. Gaussian Naïve Bayes

Gaussian Naïve Bayes works well for unbalanced problems by itself without any adjustment required to the class weights unlike SVM. It is based on the Bayes theorem with a Gaussian distribution of $P(x_i|y)$. It is much more computationally efficient than other algorithms like SVM. Since we have an unbalanced dataset, we used it as one of our predictors.

3.4.3. Logistic Regression

Logistic regression is a traditional predictive algorithm that we used as an benchmark to compare with the performances of all our other predictive models. The log2C values that we used for Logistic Regression range from $2^{-9}\ to\ 2^{9}$.

We have all the above algorithms as a pipeline which included randomizing the order of data, stratified split of the training and test set (stratified since we have unbalanced dataset), scaling the data, performing grid search to find the best hyper-parameters using 10-fold cross-validation, fitting the model on our training set and obtaining the prediction and scores on our test set. The hyper-parameters that we did a grid-search for are number of PCA components for all the models, log2C for Linear SVM, log2C and log2gamma for Gaussian SVM, log2C for Logistic Regression, n_estimators and learning_rate for AdaBoost.

4. Results and Analysis

The performance of our models are measured by their accuracy in prediction on the test set.

4.1. DNA Methylation

We observed that for DNA methylation Linear SVM had the best performance with an accuracy of 98.09% with the best parameters being log2C as 2^{-5} and number of principal components (features) as 20. So we get the best performance by reducing the 458577

dimension space to a 20 feature dimension space. The time taken by our Linear SVM algorithm was 240 minutes. Logistic Regression also performed well with an accuracy of 97.2%. We also observed that on increasing our training size, the performance of our algorithms do not change considerably but on increasing the number of principal components the accuracy drops.

4.2. Protein expression

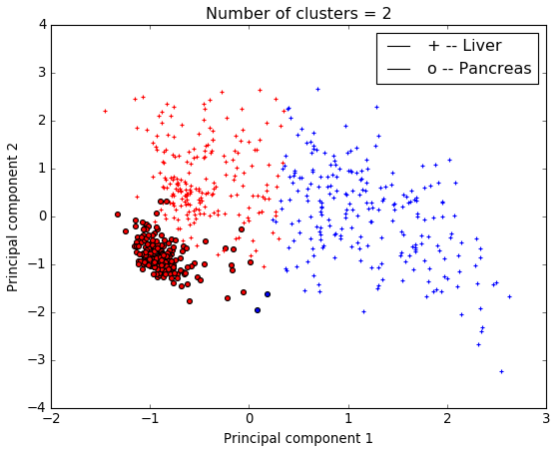
For protein expression Gaussian SVM had the best performance with an accuracy of 93.5% with the best parameters being log2C as 2^5, log2gamma as 2^{-8} and number of principal components (features) as 66. So we get the best performance by reducing the 205 dimension space to a 66 feature dimension space. The time taken by our Gaussian SVM algorithm was 15 minutes.

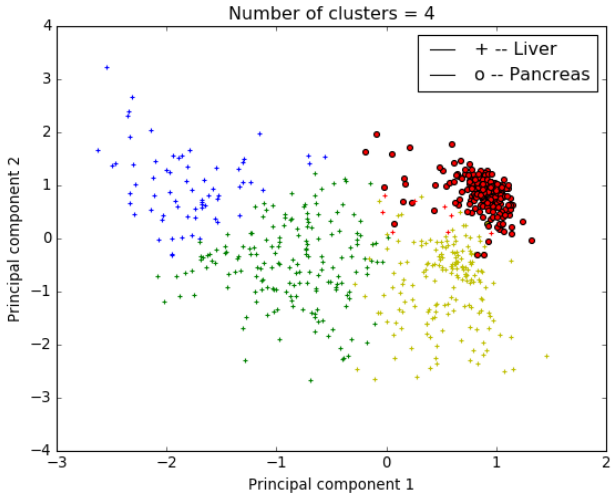
4.3. DNA Methylation vs Protein expression as an indicator of cancer

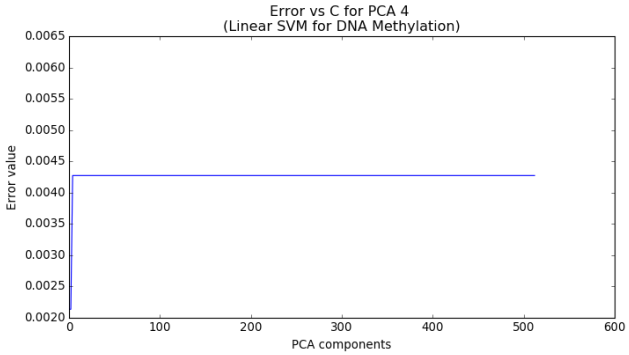
We observe that the accuracy of prediction of cancer is high using DNA methylation as compared to protein expression. But this accuracy in prediction comes at a cost of high resource utilization. For DNA methylation, the running time for the best model was 240 minutes whereas for protein expression, the running time for the best model was just 15 minutes. The data required to predict cancer using DNA methylation is around 4GB while with protein expression is around 1MB. This difference in requirement of data and running time affects the memory and CPU utilization. Thus there is a trade-off that we observe between accuracy and resource utilization.

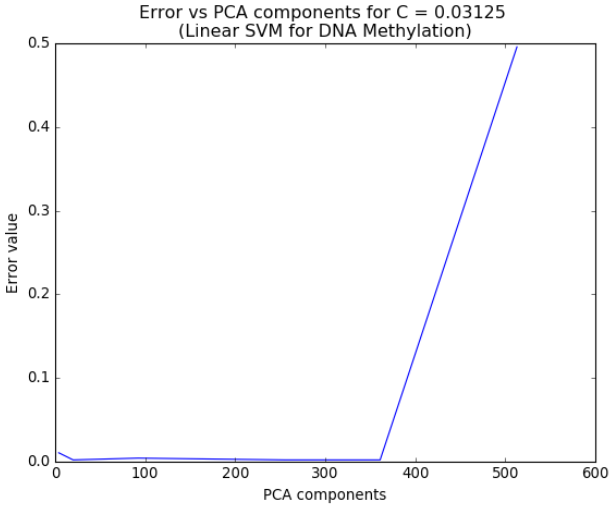
4.4. Plots

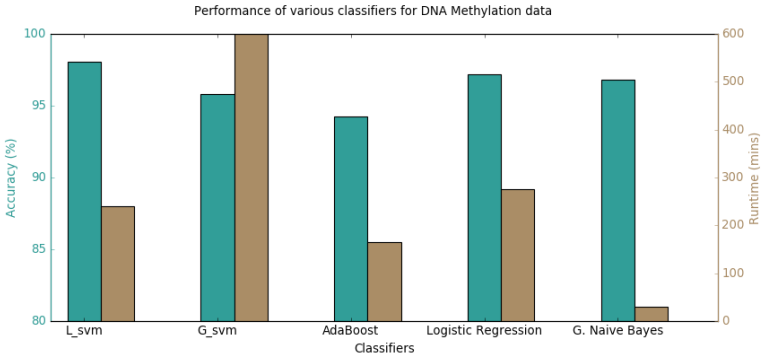
4.4.1. DNA Methylation

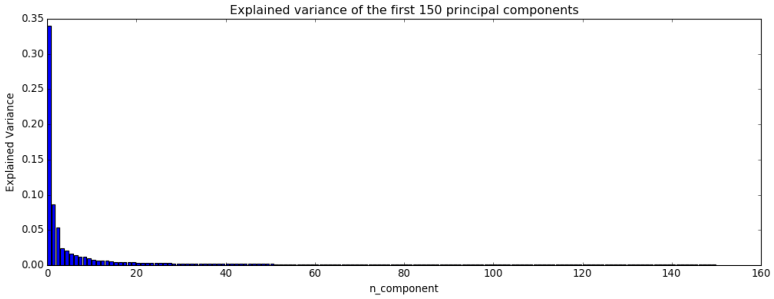



Number of clusters = 4
+ -- Liver
o -- Pancreas
Principal component 2
Principal component 1

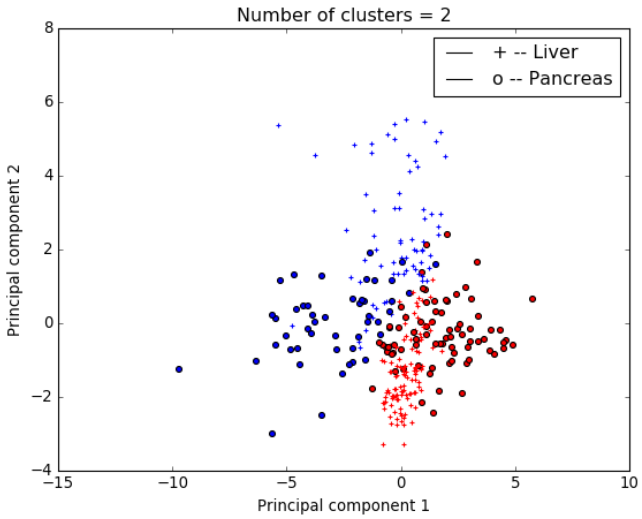

Error vs C for PCA 4
(Linear SVM for DNA Methylation)
Error value
PCA components

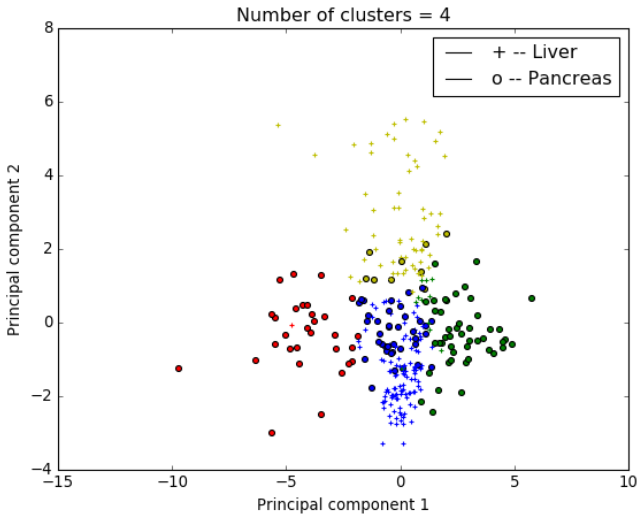

Error vs PCA components for C = 0.03125
(Linear SVM for DNA Methylation)
Error value
PCA components

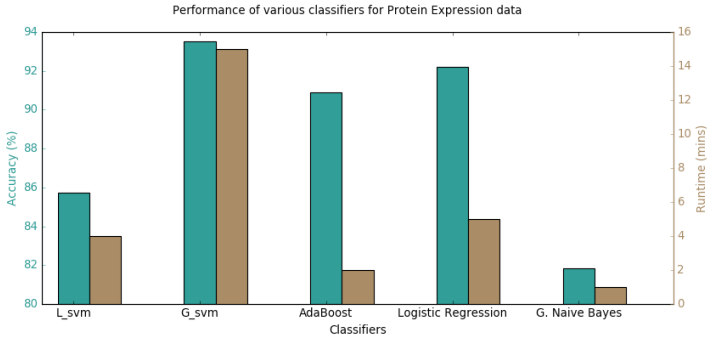


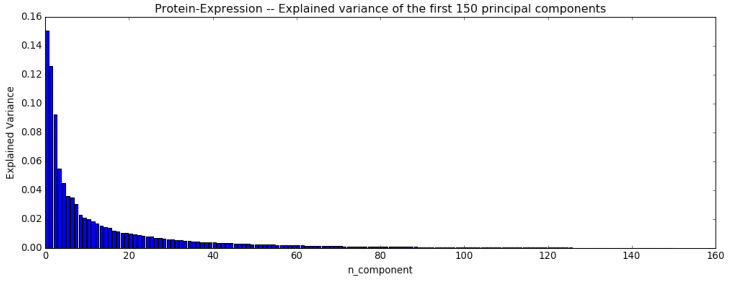


4.4.2. Protein expression









5. *Future work*

There is a lot of scope for improvement in combined terms of accuracy and computational efficiency. In future we would like to improve upon the accuracy of prediction using protein expression by varying the feature set and also try improving the computational efficiency of prediction using DNA methylation by experimenting with various other representation of the data and features. We also have in mind to combine the DNA methylation profiles and the protein expression together and measure the performance of our predictive models.

Code and Data

https://github.com/rajegannathan/Pancreatic-carcinoma-classifier

https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm

References

1. John Hayward, Sergio A. Alvarez, Carolina Ruiz, Mary Sullivan, Jennifer Tseng, Giles Whalen – "Machine learning of clinical performance in a pancreatic cancer database", Artificial Intelligence in Medicine 49 (2010) 187–195
2. Krzysztof Pawlowski, Frederic Pio, Zhi-Liang Chu, John C. Reed and Adam Godzik – "PAAD – a new protein domain associated with apoptosis, cancer and autoimmune diseases", TRENDS in Biochemical Sciences Vol.26 No.2 February 2001
3. PeiWei Zhang, Lei Chen, Tao Huang , Ning Zhang , XiangYin Kong , YuDong Cai – "Classifying Ten Types of Major Cancers Based on Reverse Phase Protein Array Profiles", Published: March 30, 2015 DOI: 10.1371/journal.pone.0123147
4. Vivek Jain, Weizhuang Zhou, Yifei Men – "Machine Learning Classification of Kidney and Lung Cancer Types", Stanford University
5. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011
6. Yuan Yuan et al. – "Assessing the clinical utility of cancer genomic and proteomic data across tumor types", Nature Biotechnology 32, 644–652 (2014) doi:10.1038/nbt.2940
7. Data: TCGA: https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm