Dimension Reduction Techniques on pCR Binary Response Data

Miela Foster

5/4/2021

Motivation

Throughout STAT117 we have investigated ways to use gene expression data to better understand the functionality of biomarkers in cancer data. We worked from small individual studies to larger data sets featured in the final project. Ultimately, we built up to creating classfiers for cancer types, and even testing the performance of those classifiers. Because classfiers can often be unique depending on the studies, number participants, and batch effects investigated etc, classfiers are not consistent throughout bio-statisticians. This brings me to the goal of the final project, train a classfier for pCR response data using gene expression levels from multiple genes and studies. And even further, create a classfier that is better performance-wise than the Prosigna classfier in terms of AUC.

There are several strategies one could take when training a new classfier. Based on the course materials and other statistics courses, one could focus on the following routes:

1. Sampling Strategy
2. Machine Learning
3. In Class Strategies: k-TSP, Mas-so-menos
4. Feature Extraction/Dimension Reduction
5. Performance Metric

For the purpose of the final project, in order to pursue a novel idea that hadn't been discussed in in-class presentation I chose to focus on feature extraction and dimension techniques when building my classfier. Because of my familiarity with dimension reduction techniques and also with the amount of data I had, the correlation between the data (which will be discussed later), I felt that the most natural choice in building a robust classfier would come from dimension reduction technique. Finally, this final project is motivated by the end goal of testing AUC as the primary metric for classfier performance. However, it should be noted that there are several drawbacks to AUC, a model with high discrimination isn't necessarily well-calibrated, doesn't translate to high accuracy, ignores predicted probability, and also doesn't account for goodness of fit. This means that regardless of the results of our model, it is improtant that we understand that AUC is only the tip of the iceberg when if comes to creating a well-performing classfier.

Training & Test Split

Upon investigating the pam_50 data and the gene expression data I come across several anomalies.

1. First there were several studies that included missing data.
2. The studies are all of different size (genes and number of participants)
3. The studies had varying gene expression levels, meaning that all of the genes in the study were of equal importance.

Additionally, after listening to many in class presentations I decided to create the following set up for my training set split:

I removed studies 1, 2, 3, 17, 18, becuase they were missing pam_50 classifications. All of the these studies had NAs in the classifier list. I rationalized this because if the goal was to compare pam_50 to another classfier than it would make sense to only include studies that the pam_50 classifier was able to identify as well. While I understand at the same time I am literally removing information from a potenitally new classifier that could have utilized that data, and potentially increase its performance. However, because it was impossible to tell the reason why pam_50 did not have a classifer for studies 1, 2, 3, 17, 18, I chose to remove them for siplicity of the project.

Finally, I created a random number generator to split the remaining studies into a test and training set. I sampled 10 studies out of the set to be included in the training set, and the remaining 3 studies were included in my test set. This ultimately left me with a 80/20 split for the training and test sets.

EDA

In order to further explore anomolies and useful information in the data set I first investigated the rate of each outcome in my training sets. To do this I ploted bar graphs of the studies binary outcomes. The histograms of each study are plotted below

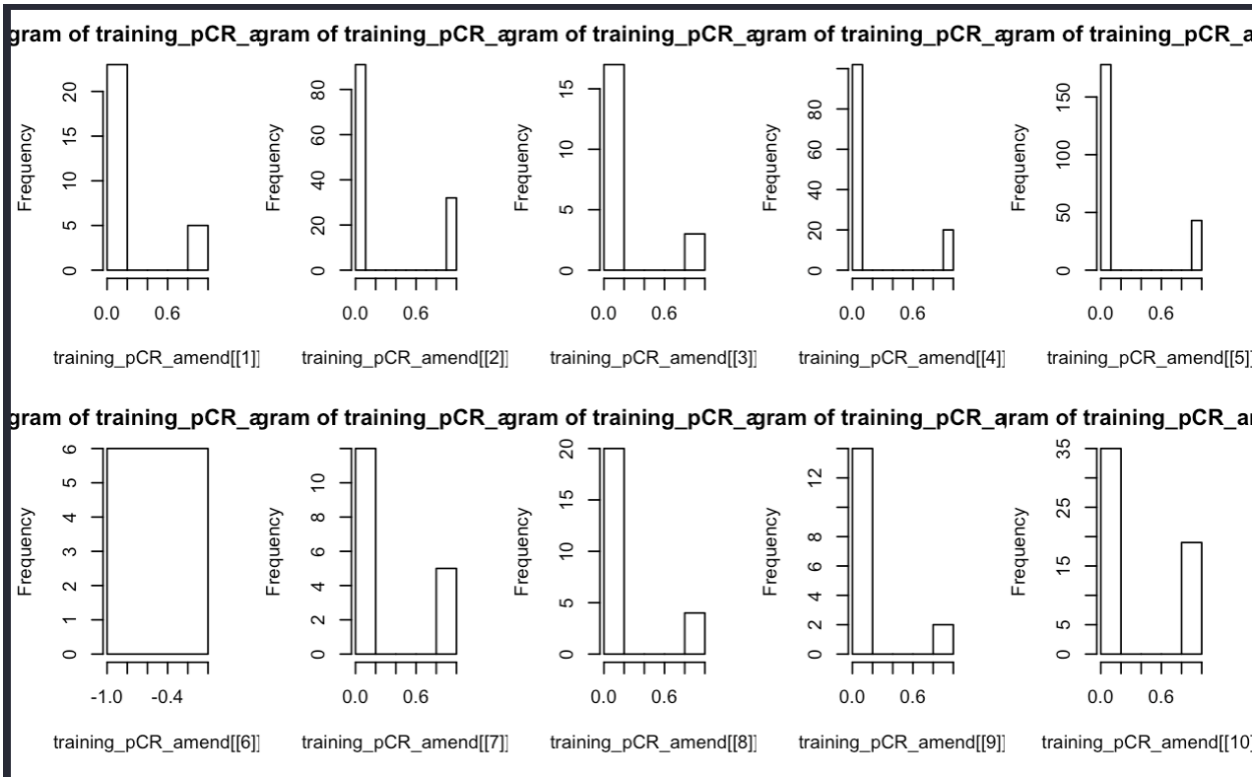

image:

As you can see there are varying frequencies of each outcome within each study. I wanted to further explore this anomoly within the dataset so I also began to plot the distribution of the each binary outcomes' gene expression level for each study. I was curious if there was some kind of correlation between gene expression levels and each binary outcome, as it might illuminate to a classifer strategy. Below are plots of two studies, these studies had the greatest amount of variation between binary outcomes, so I believed they might exaggerate discrimination between binary outcomes.

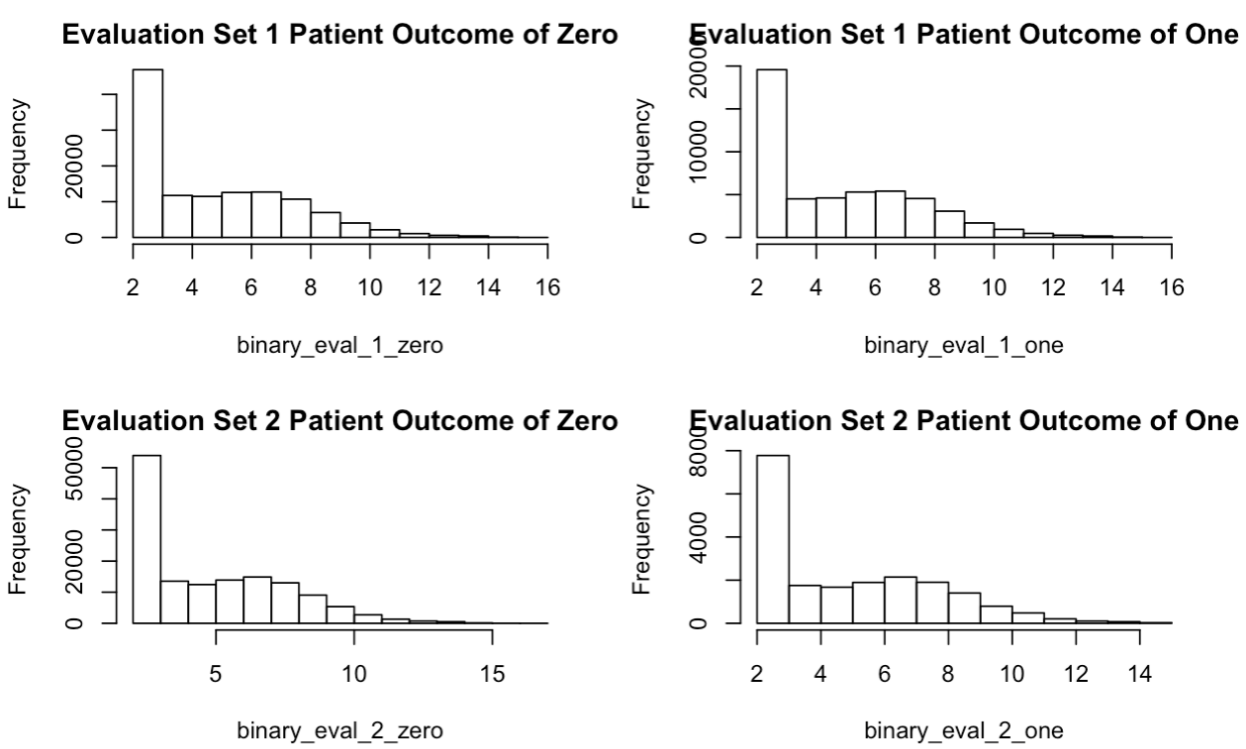

image:

From the results of the plot above, it was clear that this data wasn't as discriminatory as I believed. This might be why there are 5 classifications in the pam_50 data. However, I wasn't quite confident at this point in choosing a classficiation strategy based on the EDA.

Extended EDA: Correlation Explorations

My last exploration of data would include a method called canonical correlation analysis. Canonical correlation analysis is defined as a method that understands if there are correlations among variables, canonical-correlation analysis finds linear combinations of random variables which have maximum correlation with each other. I used this method to explore the correlation between the binary outcomes, this would help me understand if I should focus on reducing the size of the dataset, and strategize my classifer around dimension reduction.

I found the following:

1. The association Between Binary Outcome of 0 (using studies 7 and 9)in 1st Dimension was 0.95300842
2. The association Between Binary Outcome of 1 (using studies 7 and 9) in 1st Dimension was 0.9481966
3. The association Between Binary Outcome of 0 and Binary Outcome of 1 in 1st Dimension was 0.9412148

Unfortunately, I found that most of the data was highly related, even the data between opposite binary outcomes. This led to me finally develop my classfier which centered around feature extraction in order to reduce the relattness of the data, and further exaggerate the discrimination in the dataset.

K-Means & k-TSP

Because of the high association of the data and the also the novelty of dimension reduction I chose to begin with a K-means algorithm and then paired the K-means classifer with the k-tsp classfier used in class, thus adding another element of dimension reduction.

The setup for my classfier is as follows:

1. Pair each training set with most variation in binary outcomes (1,9,10,7) with a testing set based on the number of participants in the study, so that they approximately match. (see presentation deck for pairings)
2. Create a baseline model to compare: Run k-TSP (krange=6) for each training set (1,9,10,7), extract AUC with the assigned testing set.
3. Run the k-means classifier on studies in training set, set for five clusters.
4. Subset the data based on each cluster classification (1-5) for each study
5. Run the k-tsp classifier for each cluster subset for each study (krange= 6)
6. Extract AUC with the testng set assigned for that training set, for each cluster (1-5)

Cluster 3

I ended up finding that for the 4 training sets I ran my clustering algorithm that the cluster 3 outperformed the baseline model and all of the rest of the clsuters AUC values everytime. Please note that these are amongst the 10 original training sets. I chose these training sets to hone in on because they had the greatest potential to increase discrimination between the binary outcomes, and hopefully ultiamtely improve the classifier.

My best model is as follows: Training set 7, paired with test set 1. Data subsetted for the third cluster . AUC value = 0.756923 . Which also is better in terms of AUC than the pam_50 classifier.

I was curious why cluster 3 outperformed all of the rest, and upon looking into it more, these clusters tended to have lower standard deviations and an average mean when compared to the other clusters. I also want to be more specific, because I took a dimension reduction technique I wanted to explore how much each cluster was representing in each dimension.

The plot of the cluster in the first and second dimension in my best model is below:

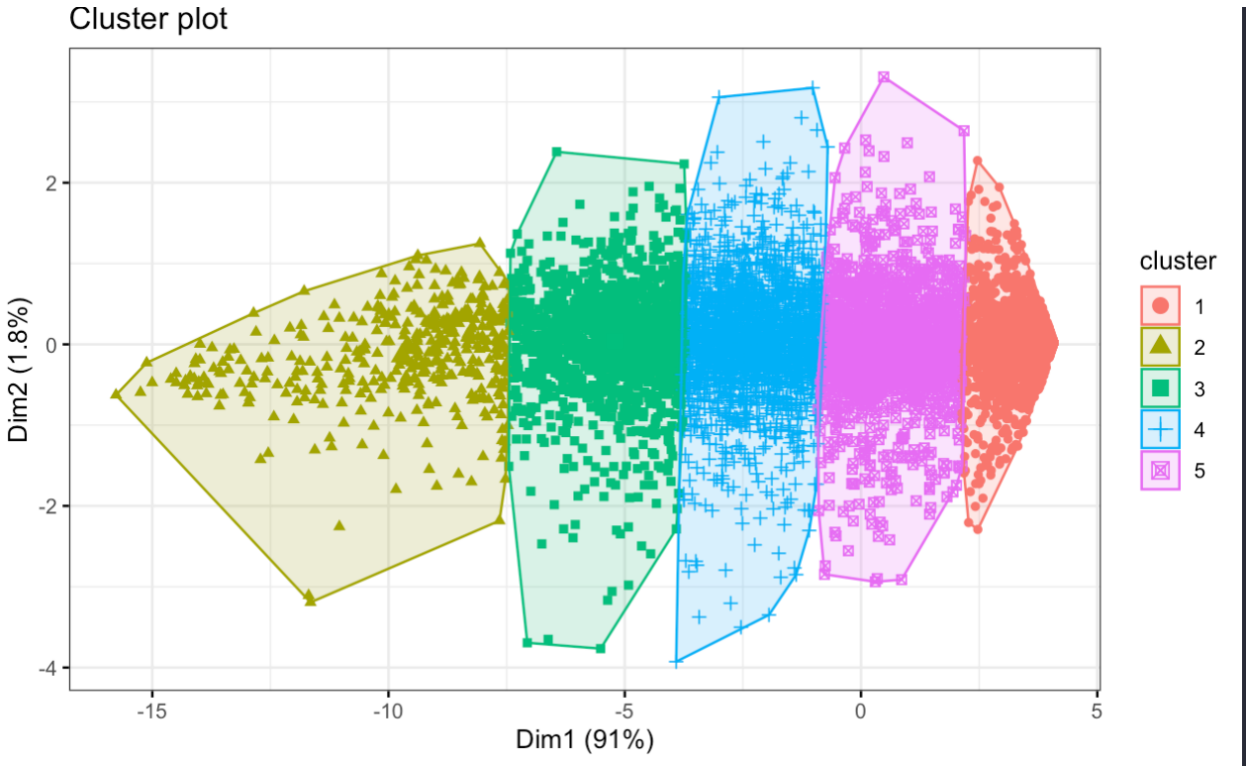

image:

The visual shows that most of the interesting variables are in the first dimension, but even more the variables represented in cluster 3 are never high in either dimension. Additionally, k means in a linear algorithm, and it seems like we were able to capture these clusters well, meaning there may not be as many non-linearities in the data set.

Discussion

For the purposes of novelty and the guidelines of the project I decided to take a dimension reduction technique. I did this for a couple reasons. First, there was a alot of data, and each study varied tremendously. Secondly, there was no clear discrimination between binary outcomes. ANd finally, most of the data was also highly correlated. All of these signaled that feature extraction might be the best route forward. Additionally, because of my familariity with dimension reduction techniques I felt very comfortable with this decesion.

I choose to use a linear method called k-means to better extract features. I chose to do five cluster for each training set because the pam-50 data had five classfications. I also chose to only focus on 4 training sets. These training sets represented those studies that had the most variability in binary outcomes. While this might not represent real world data, I chose this route because I was already pursing dimension reduction, and had previously removed 5 studies when orgiginally cleaning the data. So, it was important for me to keep as much "interesting" information as possible. Finally, I paired the k means classifer with a k-tsp classifier because of familiarity with course material.

There are several questions I have yet to answer from my analysis. 1. Did removing 5 studies also remove valuable information from my classifier? 2. Was it wise to pair each training set with a test set that matched in size (number of participants) 3. Should I have tried a non linear dimension reduction technique, like kernel PCA? Are there other non-linear extentions that would have been well suited for the data? 4. Was it wise of me to choose 5 clusters for the k means algorithm? 5. Why did clsuter 3 perform so well? What kinds of genes are those, do they also appear in well-performing classifiers in other student's presentations?

Ultimately, I am glad that I took a novel appraoch to classification, would want to change cluster size, and also try non linear dimension reduction techniques if I had more time to complete the project. Lastly, I'd like to go back to my original point in the motivation regarding if AUC is the best way to test performance of a classifier. There are several other metrics we could have used. And even further AUC while preferencing discrimination it does not tell us the probability if that discrimination will ultimately work. To me, this is the most important step in biostatistics, as if the classifer doesnt have a a high probablity of working in implementation then its meaninignless. Many statisticians can hack their way to a good classifier, manipulating data, only using some studies (like I did) however this is just the tip of iceberg when it comes to implementing in the real world.