Statistics 440: Case Study 1

Brad Smallwood, 301228034
In Collaboration with: Barinder Thind, Matthew Reyers, Ryan Sheehan

December 20, 2017

Abstract

In this case study we aimed to cluster patients together based on their health status using genetic data and to predict if they will develop Crohn's Disease, Ulcerative Colitis, or will remain healthy based on that genetic data. A K-Means Clustering method was used to cluster the patients and a Random Forest was used to predict the disease status of patients. K-Means had lackluster results but Random Forest achieve roughly 75% on average when predicting the disease group.

Keywords: Random Forest, K-Means, Ulcerative Colitis, Crohn's Disease, Predictive

1 Introduction

Ulcerative Colitis (UC) and Crohn's Disease (CD) both affect the intestinal track of the afflicted, causing inflammation. This inflammation then results in a variety of complication for individuals such as cramps, weight loss, and other symptoms. The aim of the analysis performed here is to determine if a gene-markers or combination of gene-markers will be able to determine if a person has UC or CD, and if it can be predicted whether they will develop them. More specifically this report will address the two purposed research questions: (i) Can the gene-markers be used to cluster individuals into healthy individuals, CD patients, UC patients? (ii) Can the gene-markers predict the disease state of individuals from the three biological groups?

The methodology section of this report will describe the features of the data and some of the decisions made when cleaning it. Following this, the two models used to analyze the data will be presented. K-Means clustering was used to answer the first research question while Random Forest was used to answer the second. Lastly, the results of the two models will be presented.

2 Methodology

2.1 Data Features

The dataset used in this report was provided by the Statistical Society of Canada for a case study competition. It is comprised of 126 patients with information on their age, sex, ethnicity, 309 gene-markers, and a last categorical column which details whether they have UC, CD, or are healthy. When first read in data was moderately messy and required some tidying. Luckily for us, Dr. Davis did the majority of the cleaning. Aside from general tidying it was only altered in one significant way. The ethnicity was recorded for each patient which could be Caucasian, Black, Hispanic, Asian, or Indian except there were very few patients who were not Caucasian or Black. Of all 126 patients there were only 4 who were not white or black. For this reason, those 4 patients were removed from the dataset. If we are interested in seeing how those ethnicities relate to developing UC or CD then more data should be collected with those subpopulations. There were three recorded values that could have been considered outliers in the data, however, do to my inability to conduct a follow up investigation into why those values were recorded, I decided to leave them in. The data set was unlikely to be inputted manually so these values were actually measured. Whether or not these values were the result of measurement error or not is difficult to determine when we are not able to know the process of which the data was recorded. Lastly, the data was randomly split into train and test data where 75% was allocated for training data and the remainder was allocated to the test set.

2.2 K-Means Clustering

K-means clustering partitions data into clusters which have k distinct groupings. This unsupervised learning technique will allow us to group together the patients by condition, based off the gene data we have for them.

The K-Means algorithm begins by randomly assigning a number $k \in [1, K]$ to each observation which creates the initial clustering. Iteratively, the cluster assignments continue until the clustering stops changing. More specifically the *centroid* is calculated for each cluster and observations are assigned based off of which *centroid* they are closest to. This euclidean distance is determined through,

$$\min_{C_1,...,C_k} [\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2]$$

In order to determine the optimal number of clusters k, I chose to use what is know as an elbow method. With this I added clusters until they did not contribute to the model anymore.

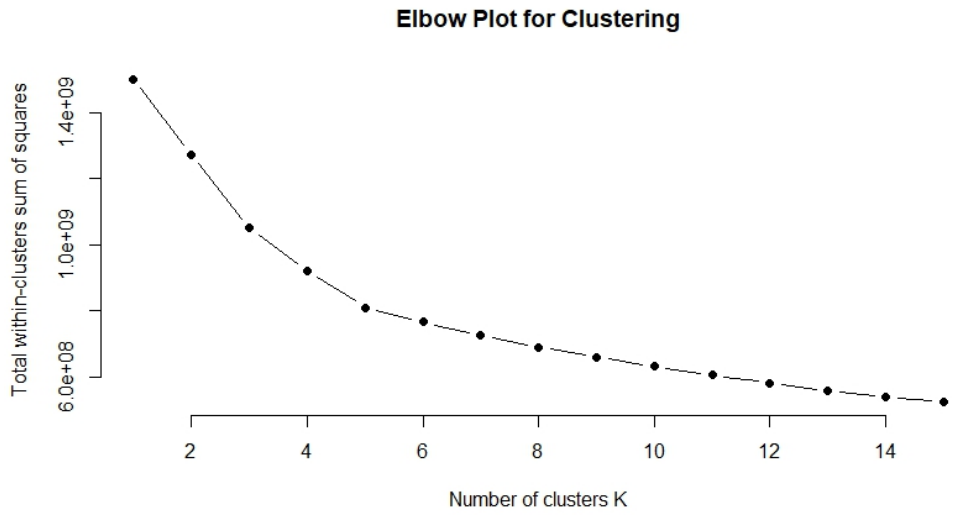


Figure 1: Elbow Method Plot

Here we can see that at roughly 6 clusters do not appear to change much anymore. The results from this clustering are discussed in Section III.

2.3 Random Forest Model

A Random Forest is a machine learning model that can be used for classification or regression problems. It is composed of many randomly generated decision trees from bootstrapped data. In each of these decision trees it makes a prediction and then averages all those prediction in the final model. This model was chosen for two reason, firstly because it is able to handle having a large number of predictors (as we have in the data), and secondly because decision trees are less sensitive to outliers, like the ones I chose to keep in the dataset. The number of trees were chosen by holding the randomForest() function parameter mtry constant at the default value while varying the number of trees from 100 to 1000. Upon inspection it was determined that the out-of-bag error was lowest on average when 600 trees were used. Following this the function tuneRF() was called to determine the optimal value for the parameter mtry. This revealed that the optimal value was the default of 17. The results of this method will be discussed in Section III. In the graph below we can see how the error for each of the classification varied as we increased the number of trees.

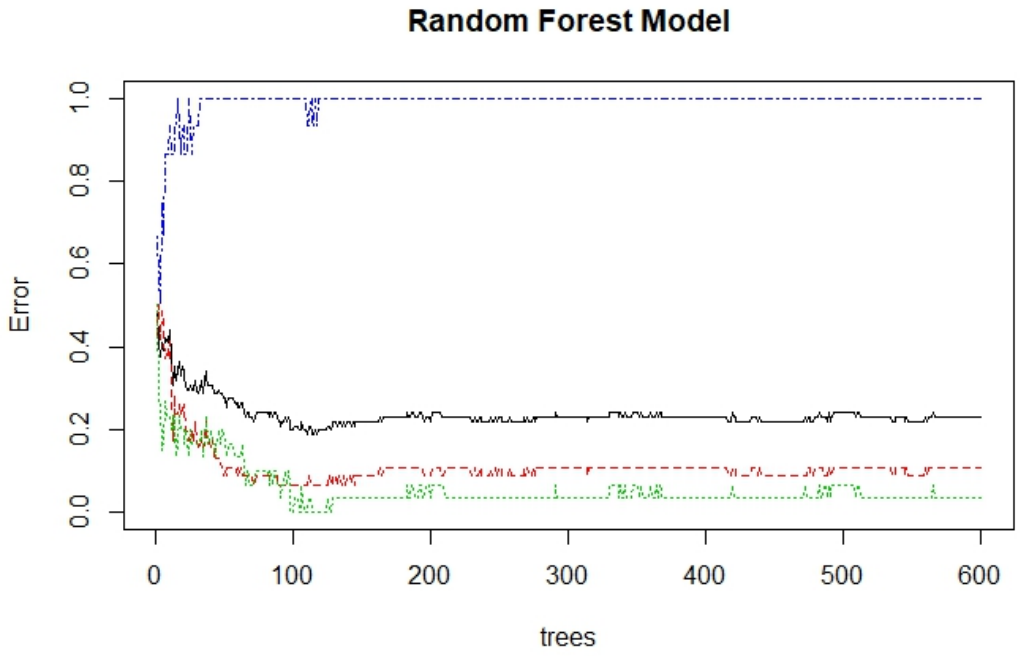


Figure 2: Plot of the Error of the Classifications.

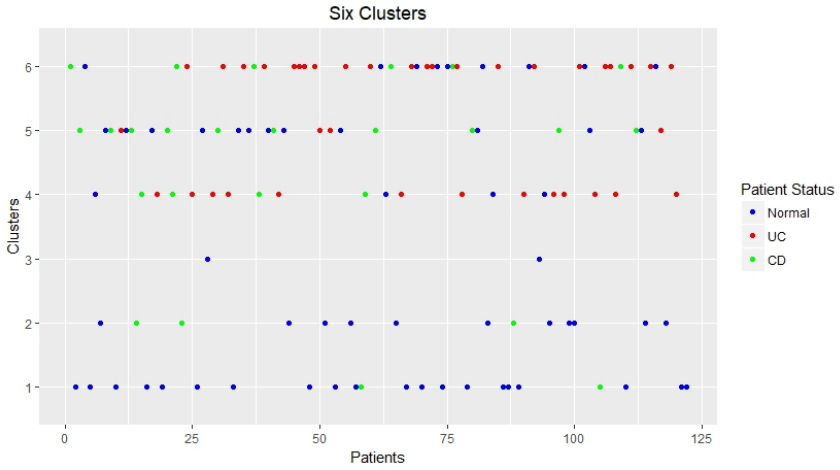
Clearly there are many issues that will need to be addressed, namely the extremely high error for one of the classification. This will be further discussed in Section III. Using the specifications described above, a final simulation was performed which randomly sampled from out of the data the test and train sets, fit a model against the training data, made predictions against the test data (even though randomForest() splits those values in train into test and train), and then calculated the average prediction accuracy across all iterations.

3 Results and Discussion

3.1 K-Means

The K-Means models utilized six clusters. As seen in the figure above, patients who were healthy were more commonly grouped in the first two clusters. K-Means had a much harder time distinguishing between patients who had UC and CD but those patients were grouped together in clusters four, five, and six. Even then there were still numerous patients without CD or UC who were grouped in clusters five and six.

Figure 3: K-Means: Patient Clustering.



3.2 Random Forest

The final random forest model resulted in approximately 21% estimated out-of-bag error, on average. Overall the model was fairly successful when classifying whether the patient was healthy or had CD, however, it grossly misclassified patients with UC, as seen below in **Table 1**.

Table 1: Confusion Matrix

Observed/Predicted	Crohn's Disease	Normal	Ulcerative Colitis	Class Error
Crohn's Disease	43	3	0	0.0652
Normal	2	28	0	0.0666
Ulcerative Colitis	14	1	0	1.0000

The dataset had relatively few occurrences of UC compared to either healthy patients or CD patients. Random Forest aims to maximize the accuracy of the overall model, and when a class within the data appears infrequently, the algorithm can achieve the highest accuracy by simply not categorizing predictions in that under represented class. This problem can be accounted for by changing the weighting of that class or by choosing a model that will account for this such as a gradient boost model.

4 Conclusion

The goal of this case study was to use gene-marker data and determine if we can cluster patients together based on whether they have UC, CD, or are healthy, and secondly to see if we can predict whether or not a patient will develop one of those diseases. K-Means clustering was a fairly unsuccessful model for clustering patients while Random Forest achieved approximately 75% accurate for classifying the three conditions. Possible future methods to be used to improve the prediction accuracy could include a Weighted Random Forest or used a more sophisticated method such as Gradient Boosting or Extreme Gradient Boosting.[1]

References

- James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R. Springer, 2017.

[1]Both Gradient Boosting and Extreme Gradient Boosting were experimented with but Gradient Boosting provided no better results over Random Forest and additional difficulties were encountered when attempting to use xgboost. To keep this report a reasonable length, neither were included.