Supervised Machine Learning:

Algorithms for IBD Classification based on Gene Expression Profiling

Dao Feng

York University
Department of Mathematics
Prof. Xin Gao
April 22nd, 2023

1 Introduction

Inflammatory bowel disease is a difficult disease to diagnose without proper diagnostic procedures and clinical tests that include invasive endoscopy (Hübenthal et al., 2015) which can be difficult on both the patient and the doctor. Although both Crohn's Disease and Ulcerative Colitis are both considered IBD of the disease group, there are differences between them that they can appear in and affect different locations of the bowel (Mossotto et al., 2017) which can add further difficulty in the process of diagnoses.

With the use of machine learning classification, we explore potential indicators and biomarkers of inflammatory bowel disease that can lead to less invasive diagnostics. Similar classification methods have been conducted in other medical topics such as discriminating cancerous cells from non-cancerous cells (Pappu & Pardalos, 2014). This research aims to differentiate inflammatory bowel disease with genetic expression levels using unsupervised machine learning classification algorithms extended to multiple classes.

We will use Linear Discriminant Analysis and Support Vector Machines in combination on our data as they are both well established methods for classification in the computational biology and microarray gene expression analysis field.

2 Data

The Global Gene Expression data set from was sourced from the Statistical Society of Canada and originally produced for Molecular classification of Crohn's disease in an academic research paper published in 2006 by Burczynski et al.The dataset contains information of genome-wide gene expression profiles for 126 individuals using the Affymetrix HG-U133A human GeneChip array to quantitatively measure gene expression in hopes to identify candidate biomarker genes (Statistics Canada, 2017) that can selectively detect and diagnose Inflammatory Bowel Disease.

The data records 41 healthy individuals (NN) as a control group to 26 Ulcerative Colitus (UC) patients and 59 Crohn's Disease (CD) Patients. The quantitative measures of each probe set was interpreted through the use of MAS 5.0 software as numerical measures of expression levels in each individual.

Data Structure

The data is structured with 307 probesets that measure diversified gene expression among 126 individuals of three separate health classes. In additional, other primary variables detailing basic patient information was also provided under "Sex", "Ethnicity", and "Age".

Data Wrangling

Data wrangling and pre-model processing was vital to ensuring the integrity of the data set. Through The first step was to use string parsing methods to extract unique patient identifiers and original probe set IDs in the row/column names and reduce the redundant strings of information and de-clutter the labels.

Although this data set is very comprehensive in the sense that there are no NA values present, the variable types were not in accordance with their inherit nature. Thus, the

conversion of all gene expression variables from characters to numeric data was necessary in order obtain tangible results. Furthermore, disease group and other categorical variables such as ethnicity, and were converted into multi-level factors with corresponding levels.

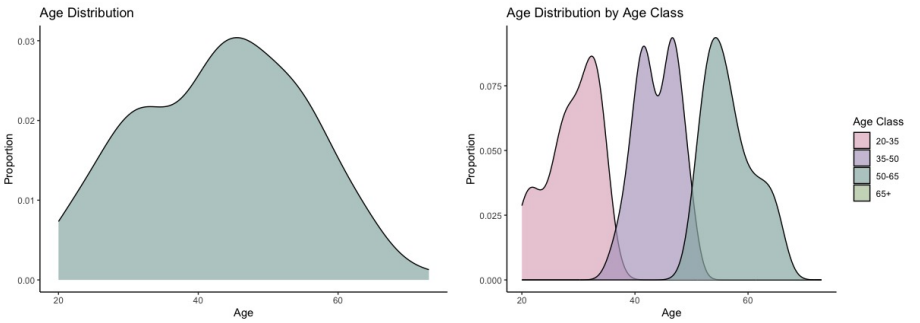
The variable Age Group was also derived from Age (numeric) for exploratory data analysis and visualization purposes. It was also important to code Sex into binary data with Male = 1 and Female = 0. Mis-spellings of categorical variables were also corrected before transposing the data frame structure such that each row was representational of an individual patient and each column was a predictive covariate to the disease group. Upon further manipulation, mismatched columns were corrected and through the use of variable manipulation, input cleaning, and data re-structuring, the data set was refinement to 126x313 data set with workable data structure and concise formatting.

Exploratory Data Analysis

Before modeling and applying the methodology, some exploratory data analysis was conducted in order to get a better overview of the distribution of the data.

	CD	UC	NN	Total
Samples	59.00	26.00	41.00	126.00
Mean Age	35.80	35.80	35.80	35.80
Median Age	38.00	38.00	38.00	38.00
Males	21.00	8.00	23.00	52.00
Females	38.00	18.00	18.00	74.00

In reference to figure 1 in the appendix, the Healthy/Normal Patients is approximately 33% of our study sample population and that out of the other 66% of IBD Patients, Crohn's Disease Patients appear twice as more frequently in our dataset than Ulcerative Colitis patients.



In the two age variables, we can visualize the distribution of sample patients by age and age class. We note that there is one age outlier patient who is in the 65+ category.

In reference to figure 2 in the appendix, we can observe that the Ethnicity in this data set is comprised of mostly Caucasian demographic, meaning the data set is not very representational of other demographics. This is a potential limitation in our study, in which

the data is predominately Caucasian, which means our results could potentially be biased towards other ethnicities.

3 Methodology

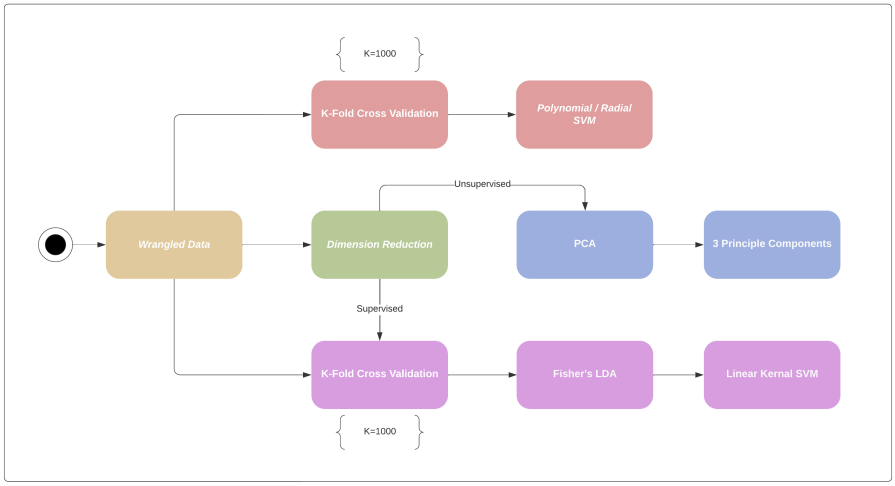
With the consideration that the data set is labeled (and thus *a priori* knowledge is known), Supervised Machine Learning Algorithms will be employed to solve this classification problem.

Although the Support Vector Machine Algorithm has shown success in bioinformatics classification, in general "classification algorithms tends to deteriorate in high dimensions due a phenomenon called the curse of dimensionality" (Pappu & Pardalos, 2014). In consideration of this problem, we explore the use of Fisher's LDA to reduce dimensionality in order to improve the SVM model such that it will have lower dimensions of input.

Principle Components Analysis

As an Unsupervised Machine Learning Algorithm, Principle Component Analysis (PCA) generally works well with non-labeled data. However, given that our data is labeled, we can still utilize the PCA method for visualization purposes, namely it's properties of dimension reduction in which data variations in the higher dimensionalities are perserved in the reduction process.

Network Diagram of the model



1. Principle Components Analysis for visualizing the data in three dimensions while retaining dimensionality variation

2. Support Vector Machine model with k-fold cross validation without dimension reduction on the original dataset to evaluate how the SVM model performs initially

3. Improved SVM model after applying Fisher's LDA as dimension reduction technique

Given that the data set has more predictive variables than observations and that it is high dimensional, Support Vector Machine methods were used to create dimensional separation in the data with certain advantages through the variety of applicable kernel functions. SVM methods were also conducted with linear discriminants of the data as an approach to making multi-class predictions. This data set is a special case of tri-class classification such that one class is neutral (NN group).

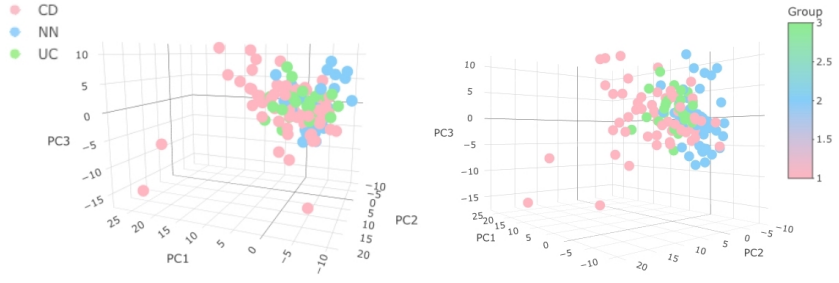
K-Fold CV

K-Fold Cross Validation was utilized in all of the algorithms with an 80/20 percent split for training and validation sets in order to avoid over-fitting the model. K=1000 was the number of iterations conducted in all of the methods. The average of results over 1000 iterative processes are then reported.

4 Analysis

Principle Component Analysis

In reducing the data to three principle components, we are able to model the clusturing of datasets conditioned on IBD class while maintaining a high proportion of variation through the three principle axes alone. Through reducing 313 dimensions to the 3 principle components we are able to retain 55.8% of variation in the un-reduced data.



Interactive 3D Models of PCA generated with plotly

We note that in our case, 55.8% of variation is sufficient for a visualization of the data, however it is not sufficient for modeling and analysis. PCA is not a good approach for this specific situation as we would need 8 principle components to capture 81.4% and 14 principle components to capture 90.7% of variation the original data variation.

Initial SVM

Now, we construct an SVM model with K-fold cross validation on the wrangled data set and test out a variety of non linear kernels on the model. Considering the high dimensionality of our data, in this initial step we specifically tested out the polynomial kernel and radial basis kernel functions under the assumption that the data is linearly separable in higher dimensions.

The kernal function for SVM models $K(x_i, x_j)$ maps features x_i, x_i for $i, j = 1, 2, ..., m$ into a higher dimensional feature space using such that $\phi(x_i)$ and $\phi(x_j)$ are non-linear mappings of our features and is defined as the inner product of the non-linear mappings:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \phi(x_i)^T \phi(x_j)$$

In our model, we used the Polynomial Kernal:

$$K(x_i, x) = (\langle x_i, x_j \rangle + 1)^p \qquad \text{where } p \text{ denotes the order of the polynomial,}$$

and the Gaussian/Radial Basis Kernal:

$$K(x_i, x) = exp\{-\sigma ||x_i - x||^2\}$$

Using k-fold cross validation to test each initial SVM model using Gaussian Radial Basis and polynomial kernels, our model predictions had an mis-classification rate of 53.56% and 18.50%, respectively.

The Gaussian Radial Basis kernel did not perform very well mis-classifying over half of the validation data, while the polynomial kernel produced a decent result with an 83% accuracy rate. This is a good initial accuracy rate to improve upon given that the SVM model can accurately classify IBD patient groups without any dimension reduction at an accuracy of over 80%.

Fisher's Linear Discriminant Analysis for Dimension Reduction

In using Fisher's LDA, we seek to project out data x_i onto a single dimension such that the classes are well separated. Class separation is denoted:

$$S_B = (\mu_1 - \mu_2)^T (\mu_1 - \mu_2)$$

$$S_W = S_1 + S_2$$

where S_B is the between-class variation and S_B is the within-class variation.

In order to achieve optimal separation, we need to maximize the between-class separation S_B in terms of their projected means $\tilde{\mu}_i$ for each class. At the same time, we want the within-class separation S_W to be minimized after the projection such that the observations in each class are tightly clustered to the class mean for the separation boundaries to be more decisive.

To find the Linear Discriminant, we maximize the objective function $J(v)$ where:

$$J(v) = \frac{v^T S_B v}{s v^T S_W v}$$

Due to the scale-invariance property of the Signal to Noise Ration S_B/S_W, we can transform the maximization problem to a constraint opitmization problem with Lagrange Multipliers by solving the Lagrangian:

$$L(v, \lambda) = v^T S_B v - \lambda(v^T S_W v - 1)$$

In which we find $max_v(v^T S_B v)$ with the constraint $v^T S_W v = 1$

Thus, by solving the Lagrangian, Fisher's optimum projection v is found through:

$$v = {S_W}^{-1}(\mu_1 - \mu_2)$$

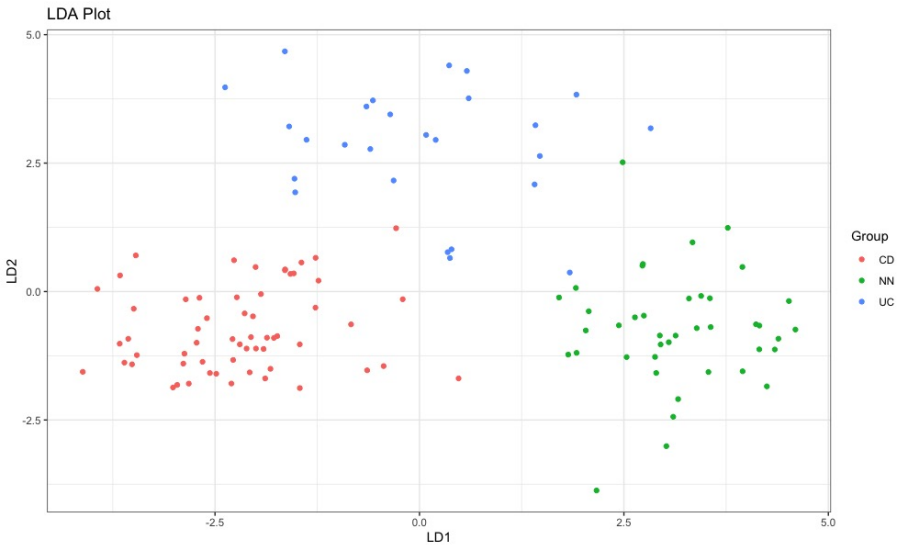
Extending this method to multiple classes, (in our case K=3) we find a second projection orthogonal to the first projection that maximizes the separation of the projected samples.

For K classes, we find r=2 projections such that

$$r = min(p, k - 1)$$

where k is the number of classes and p is the dimensions before reduction

Through this method we transform our data into $(LD1, LD2)^T$. We can plot our data on two dimensional space as follows:



From the plot of the two linear discriminants, we can visually observe that there is good separation between the Crohn's patients (CD group) and Normal patients (NN group) along the first linear discriminant. Similarily, we can observe that Ulcerative Colitis patients (UC group) are well-separated against both Crohn's patients (CD group) and Normal patients (NN group) along the second linear discriminant.

SVM after LDA Reduction

Through Fisher's LDA, we are able to create linear separation amongst the clusters on two dimensions. We then use the data LD_1 LD_2 and construct a SVM using a linear kernal which is defined as:

$$K(LD_1, LD_2) = (LD_1, LD_2)$$

We perform SVM using the data set obtained from dimension the LDA dimension reduction. In 2-Dimensional Space, the data becomes more linearly separable, and thus we use the linear kernel to classify IBD diseases.

Through K-fold cross-validation over 1000 iterations, we achieved an overall accuracy rate of 96.92% and an overall mis-classification rate of 3.08%. Comparing this to our model without dimension reduction, this is a significantly more accurate model with only a 3% misclassification rate.

5 Conclusion

PCA analysis showed that we needed 14 principle components to maintain 90% of variation in the original dataset. In terms of dimention reduction, 14 dimensions is still relatively high.

This could be attributed to the fact that unsupervised learning methods were inadequate given the high degree of biological variation in gene expression amongst individuals as identified by Ramaswamy et al. in their 2001 paper: "Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures". The initial SVM model performed at an acceptable level with 82% accuracy with a polynomial kernel. However, we achieved a lower misclassification error rate in the Support Vector Machine model after dimension reduction which is not surprising due to past research from Brown et al. has demonstrated that SVM method have shown to out-performs all other methods and given that SVM has been successfully applied to high dimensional classification of microarray data in the past (Pappu & Pardalos, 2014).

Limitations

Ethinicity is heavily dominated by caucasian patients. This can likely be attributed to the possibility that the data collection methods did not cover a demoghraphic that is diverse enough to capture the general population.

Other Classes of IBD

There are patients whose disease characteristics cannot fit precisely into either of these two subtypes, such as 'IBD unclassified' (IBDU), which are more common in children (Mossotto et al., 2017). Our dataset does not reflect upon pediatric cases as it does not

include the IBDU class, and thus, there is a loss of generality when it comes to the IBD disease as a whole.

Multiple Classes where $K > 3$

Lastly, SVM can be computationally expensive with limited interpretability (Stankovic et al., 2021), especially when it comes to multiclass classification. In our case with tri-class classification, the model preformed well, however SVM could not be approproate when there are more classes added.

In conclusion, the SVM model was adequate in classifying microarray data of gene expression levels in inflammatory bowel disease patients. Further exploration and analysis is needed for categorizing IBD disease groups if the IBDU class is considered. Modifications to the SVM approach should also be adjusted in accordance with cases where classes exceed three.

References

[Abida] Abida, K. 1 Fisher Discriminant Analysis For Multiple Classes.

[2] Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267. Publisher: Proceedings of the National Academy of Sciences.

[3] Burczynski, M. E., Peterson, R. L., Twine, N. C., Zuberek, K. A., Brodeur, B. J., Casciotti, L., Maganti, V., Reddy, P. S., Strahs, A., Immermann, F., Spinelli, W., Schwertschlag, U., Slager, A. M., Cotreau, M. M., and Dorner, A. J. (2006). Molecular Classification of Crohn's Disease and Ulcerative Colitis Patients Using Transcriptional Profiles in Peripheral Blood Mononuclear Cells. *The Journal of Molecular Diagnostics*, 8(1):51–61.

[4] Chaudhary, A., Kolhe, S., and Kamal, R. (2016). An improved random forest classifier for multi-class classification. *Information Processing in Agriculture*, 3(4):215–222.

[5] Chih-Wei Hsu and Chih-Jen Lin (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425.

[6] Hu, P. (2017). Identify Patients with Inflammatory Bowel Disease?

[7] Hübenthal, M., Hemmrich-Stanisak, G., Degenhardt, F., Szymczak, S., Du, Z., Elsharawy, A., Keller, A., Schreiber, S., and Franke, A. (2015). Sparse Modeling Reveals miRNA Signatures for Diagnostics of Inflammatory Bowel Disease. *PLOS ONE*, 10(10):e0140155.

[8] Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., Ripke, S., Lee, J. C., Jostins, L., Shah, T., Abedian, S., Cheon, J. H., Cho, J., Daryani, N. E., Franke, L., Fuyuno, Y., Hart, A., Juyal, R. C., Juyal, G., Kim, W. H., Morris, A. P., Poustchi, H., Newman, W. G., Midha, V., Orchard, T. R., Vahedi, H., Sood, A., Sung, J. J. Y., Malekzadeh, R., Westra, H.-J., Yamazaki, K., Yang, S.-K., International Multiple Sclerosis Genetics Consortium, International IBD Genetics Consortium, Barrett, J. C., Franke, A., Alizadeh, B. Z., Parkes, M., B K, T., Daly, M. J., Kubo, M., Anderson, C. A., and Weersma, R. K. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, 47(9):979–986.

[9] Mossotto, E., Ashton, J. J., Coelho, T., Beattie, R. M., MacArthur, B. D., and Ennis, S. (2017). Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. *Scientific Reports*, 7(1):2427.

[10] Niu, Y. and Ye, S. (2019). Data Prediction Based on Support Vector Machine (SVM)—Taking Soil Quality Improvement Test Soil Organic Matter as an Example. *IOP Conference Series: Earth and Environmental Science*, 295(2):012021.

[11] Pappu, V. and Pardalos, P. M. (2014). High-Dimensional Data Classification. In Aleskerov, F., Goldengorin, B., and Pardalos, P. M., editors, *Clusters, Orders, and Trees: Methods and Applications*, volume 92, pages 119–150. Springer New York, New York, NY. Series Title: Springer Optimization and Its Applications.

[12] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., Poggio, T., Gerald, W., Loda, M., Lander, E. S., and Golub, T. R. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154. Publisher: Proceedings of the National Academy of Sciences.

[13] Selvaraj, S. and Natarajan, J. (2011). Microarray Data Analysis and Mining Tools. *Bioinformation*, 6(3):95–99.

[14] Stankovic, B., Kotur, N., Nikcevic, G., Gasic, V., Zukic, B., and Pavlovic, S. (2021). Machine Learning Modeling from Omics Data as Prospective Tool for Improvement of Inflammatory Bowel Disease Diagnosis and Clinical Classifications. *Genes*, 12(9):1438.

[15] Yuan, F., Zhang, Y.-H., Kong, X.-Y., and Cai, Y.-D. (2017). Identification of Candidate Genes Related to Inflammatory Bowel Disease Using Minimum Redundancy Maximum Relevance, Incremental Feature Selection, and the Shortest-Path Approach. *BioMed Research International*, 2017:1–15.

Appendix A

Figure 1

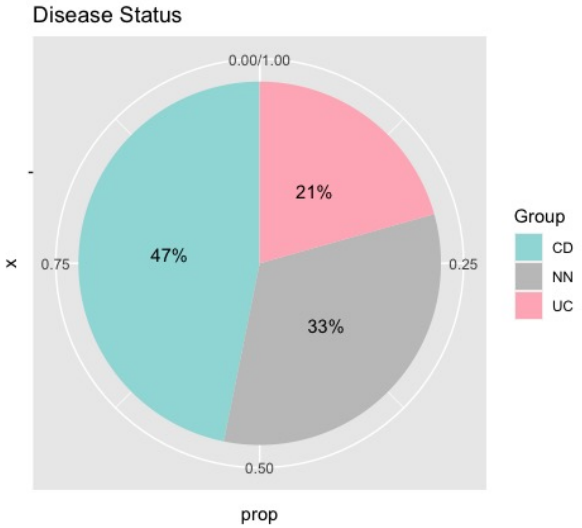
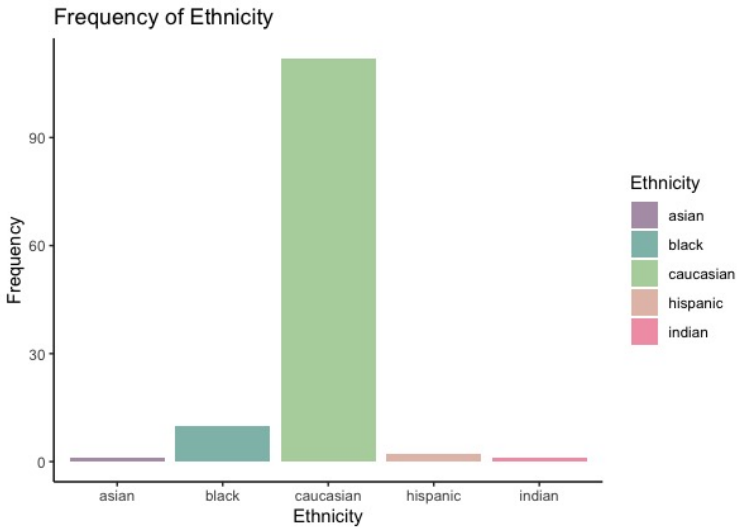


Figure 2



Appendix B