Analysis of Human Gut Microbiomics and Metabolomics for Colorectal Cancer Prediction

Ricardo Avila

December 11, 2018

Abstract

The identification of biomarkers associated with colorectal cancer (CRC) can provide a useful method for early CRC diagnosis, and can lead to a better understanding of the affiliated risk factors. This project uses a publicly available dual-feature dataset consisting of microbiomics and metabolomics data to classify patients into CRC and non-CRC groups. Classifiers were trained using penalized logistic regression, including lasso, ridge, and elastic net, as well as linear discriminant analysis. The performance of the models built using each of these classifiers is compared when using the metabolomics and microbiomics features jointly, as well as independently.

1 Introduction

Colorectal cancer (CRC) is the third leading cause of cancer worldwide, after breast and lung cancer. It is a disease with a very slow progression from detectable precancerous lesions to the advanced stages, having a good prognosis at the early stages, but a very poor one among patients with advanced disease. For this reason, the push for early diagnosis of CRC is highly worthwhile. Nevertheless, the current robust methods for CRC diagnosis are costly and invasive. [1] The development of cheap and non-invasive methods of detection would greatly improve the prevalence of early screening, and prevention of CRC.

The composition of the gut microbiome is a unique signature that stems from a person's diet and lifestyle choices. The joint population of microbes in the gut carries more than 100 times more genes than the human genome, and plays an important role in the regulation of numerous processes, such as energy harvesting, metabolism of dietary components, immunity, and activities of host or microbial-derived chemicals. Recent studies have established links between the gut microbiome and a variety of health conditions, such as inflammatory bowel disease, malnutrition, metabolic syndrome, type 2 diabetes, and colorectal cancer (CRC). [2] It has also been established that the genetic mutations that lead to the development of CRC are driven by carcinogens related to diet. Some of the metabolites present in the human gut are also affected by the unique population of microbes, which produce many important metabolites including short-chain fatty acids, biotin and vitamin K. [3]

The dataset used in this study comes from a study published in 2015 by Sinha et al. [4] The study focused on identifying linear correlations between single

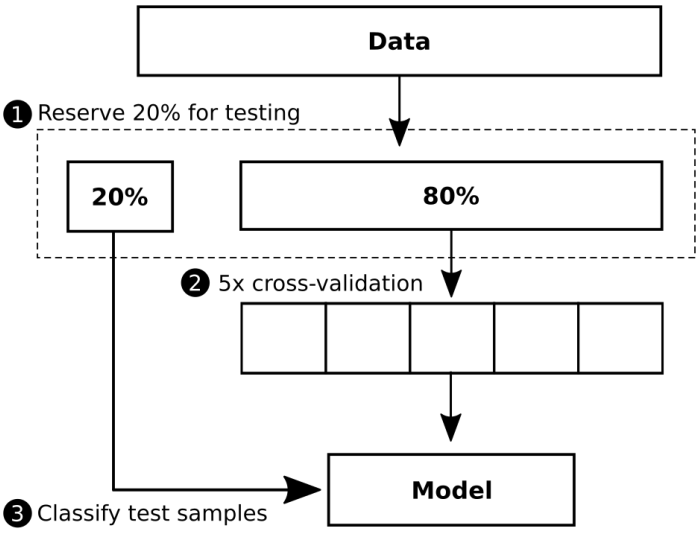


Figure 1: Schema of the cross-validation setup used for training all the different models.

microbes and multiple metabolites, either associated with CRC, or not. The authors reported several correlations: CRC was independently associated with lower levels of Clostridia, Lachnospiraceae, p-aminobenzoate and conjugated linoleate, and with higher levels of Fusobacterium, Porphyromonas, p-hydroxybenzaldehyde, and palmitoyl-sphin-gomyelin. Nevertheless, the authors did not carry out a cross-validation experiment, and did not demonstrate whether these relationships can be used for predicting CRC. This project is an attempt at building classifier models for predicting CRC cases based on microbiomics and metabolomics data. We apply two types of classifier algorithms: penalized logistic regression and linear discriminant analysis.

2 Methods

The data contains two classes of features. The first consists of the relative abundance levels — ranging from 0 to 1, of 220 categories of microbes. The second is the natural-log levels of 530 metabolites. The samples consisted of 131 patients, of which 42 were cancer patients, and 89 were controls.

2.1 Data Preparation

The two sets of predictor variables: microbes and metabolites were used independently to train the classifiers, and also joined to obtain combined models from the concatenated features. For each of these three setups, 20% of the data was set aside for testing, and the other 80% was used for training and tuning the models with five-fold cross-validation, as shown in Figure 1.

2.2 Penalized Logistic Regression

The most popular methods for penalized logistic regression are ridge regression, lasso regression, and elastic net. These methods reward goodness-of-fit but punish model complexity in order to avoid overfitting.

Ridge Regression. Also called weight decay, it shrinks the regression coefficients by imposing a quadratic ($L2$) penalty on their size. Ridge regression is useful for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large, so they may be far from the true value. [5] Ridge regression enhances least squares by applying a penalty to shrink the estimated coefficients, without actually assigning zero values.

Lasso (least absolute shrinkage and selection operator) Regression. Lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. It assigns zero values to non-important variables by applying an $L1$ penalty to the coefficients.

Elastic Net. The elastic net technique solves the regularization problem by using a linear combination of $L1$ and $L2$ penalties. For an α strictly between 0 and 1, and a nonnegative λ, elastic net applies the penalty:

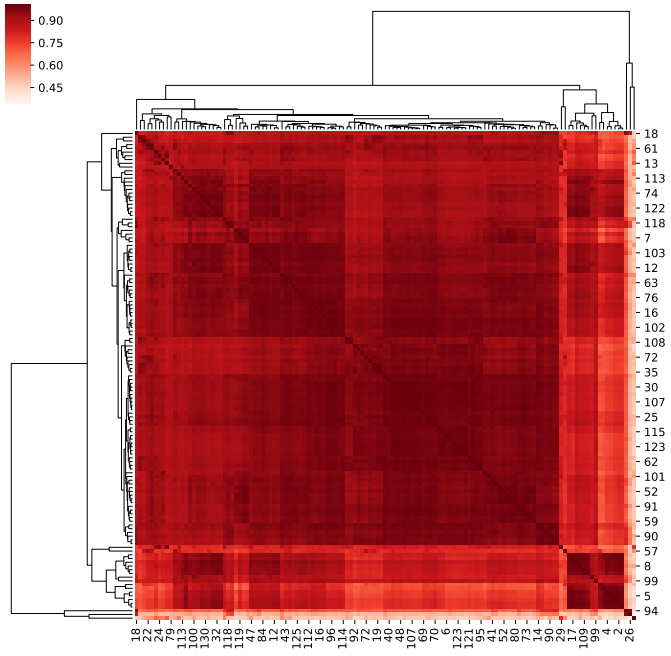
$$\lambda \cdot \sum_{j=1}^{p} \left(\alpha|\beta_j| + (1-\alpha)\beta_j^2\right) \tag{1}$$

Elastic net is the same as lasso when $\alpha = 1$. As α shrinks toward 0, elastic net approaches ridge regression. For other values of α, the penalty term $P_\alpha(\beta)$ interpolates between the $L1$ norm of β and the $L2$ norm of β.

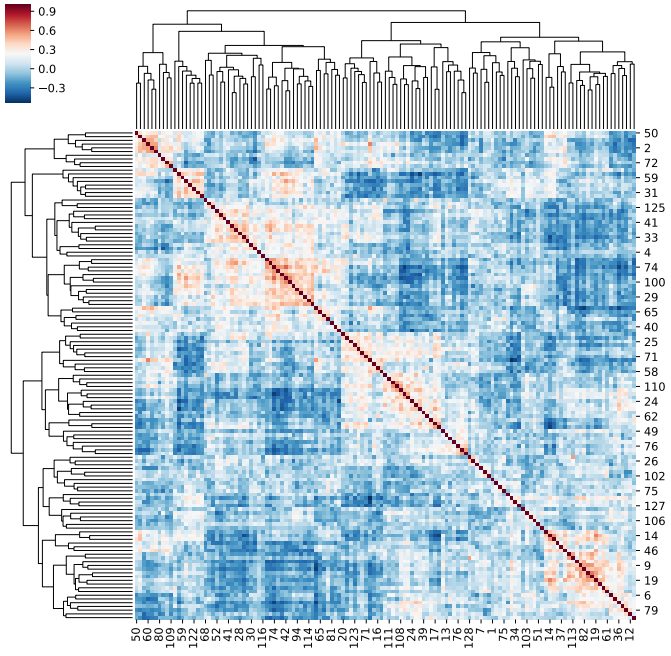
2.3 Linear Discriminant Analysis

Linear discriminant analysis(LDA), is a supervised linear transformation technique that can be used for classification and dimensionality reduction. LDA projects a linear combination of the original variables onto a lower-dimensional space in a manner similar to principal component analysis (PCA). In short, LDA projects a feature space (a dataset n-dimensional samples) onto a smaller subspace k (where $k \leq n-1$) while maintaining the class-discriminatory information. [6]

LDA was applied to the microbes, metabolites, and combined features. Using the cross-validation scheme mentioned previously, it was determined that the best parameters for the models were the 'lsqr', or least squares solver, coupled with 'auto' shrinkage determination for the metabolite and microbe features. For the combined features, the shrinkage was manually set to a value of 0.7. This shrinkage of the variance matrix improves the estimation of covariance in situations where the number of training samples is small compared to the number of features.



(a) Microbe feature correlation



(b) Metabolite feature correlation

Figure 2: Hierarchical clustering of the inter-variable correlation for the two feature sets. The microbe data shows high correlation overall, while the metabolite data only has low-to-moderate correlation.

		Metabolites		Microbes		Combined	
Model	α	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
1	0.0	0.69	0.60	0.62	0.50	0.69	0.60
2	0.1	0.73	0.67	0.62	0.50	0.69	0.60
3	0.2	0.77	0.72	0.62	0.50	0.73	0.65
4	0.3	0.77	0.72	0.62	0.50	0.73	0.65
5	0.4	0.69	0.62	0.62	0.50	0.73	0.65
6	0.5	0.69	0.62	0.62	0.50	0.69	0.62
7	0.6	0.69	0.62	0.62	0.50	0.73	0.65
8	0.7	0.69	0.62	0.62	0.50	0.69	0.62
9	0.8	0.69	0.62	0.62	0.50	0.69	0.62
10	0.9	0.73	0.67	0.62	0.50	0.73	0.67
11	1.0	0.69	0.62	0.62	0.50	0.73	0.65

Table 1: Accuracy and area under the curve (AUC) of penalized logistic regression models for the three feature types. The value of the parameter α determines the weight of each type of penalty that is applied, where a value of $\alpha = 0$ corresponds to pure L2 regularization, and a value of $\alpha = 1$ corresponds to pure L1 regularization.

3 Results

3.1 Exploratory Data Analysis

Between-variable correlation matrices were computed for the two feature matrices to identify the presence of multicolinearity. It was observed that the microbes dataset contains a high level of correlation between most of the variables, as shown in Figure 2. The metabolites dataset only shows moderate to low inter-variable correlation.

3.2 Logistic Regression

Eleven different models were generated with different values of the parameter α, which controls the combination of $L1$ and $L2$ regularization under elastic net.

From the analysis, comparing the microbes and metabolites features using the Area Under the Curve (AUC) measurements, metabolites yielded better results. The microbes dataset was the worst, with an AUC score of 0.5 for all models, indicating that the predictions are not better than random. Combining the two feature sets reduced the performance of the metabolites-only models.

For the metabolites data, the best value of α was between 0.2 and 0.3, corresponding to an elastic net with a heavy ridge penalty. This is consistent with the fact that the features suffer from multicolinearity. On the other hand, the microbes features did not seem to be affected by the regularization parameter at all. It is hypothesized that this is because the microbes dataset not only suffers from very heavy multicolinearity, but also has many unimportant variables.

	Accuracy	Balanced Accuracy	AUC
Microbes	0.67	0.62	0.68
Metabolites	0.81	0.76	0.86
Combined	0.81	0.72	0.85

Table 2: Accuracy, balanced accuracy (average of recall obtained on each class), and area under the curve (AUC) of LDA classifiers for each of the three feature sets. Because the classes on the dataset are unbalanced, we focused on balanced accuracy, and AUC as the primary estimators of model performance.

3.3 Linear Discriminant Analysis

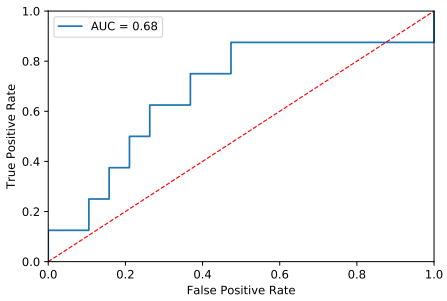
Overall, LDA performed better than all of the penalized regression methods. When comparing across the three feature sets, the performance of the metabolites data by itself was the best, with an AUC score of 0.86, and balanced accuracy of 0.76. The microbes data performed the worst, with an AUC score of 0.68, and a balanced accuracy of 0.62. Combining the microbes and metabolites features resulted in a small decrease in performance. The full results are shown in Table 2, and Figure 3 shows the receiver operating characteristic(ROC) curves for these models.

It is also worth mentioning that when analyzing the variable weights used in the LDA transformation, which can be an indicator of variable importance, we found that some of the microbes and one of the metabolites that were mentioned in the paper by Sinha et al. [4] ranked near the top. These variables were the microbes Fusobacterium, CLostridia, and Lachnospiraceae, along with the metabolite palmitoyl-sphingomyelin.

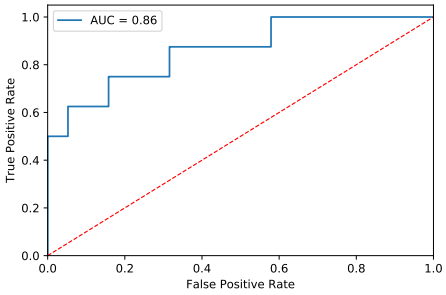
4 Discussion

The LDA methods outperformed penalized logistic regression. This was surprising, because one of the assumptions of LDA is that the variables are independent, and normally distributed. Nevertheless, it is also stated in the literature that LDA is sometimes robust even when the assumptions of common covariance matrix among groups and normality are violated. [7]

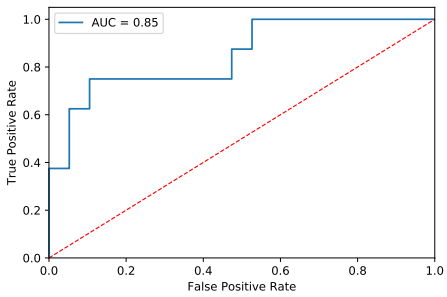
Although LDA performed reasonably well, it is still nowhere near the accuracy necessary to serve as a reliable clinical testing method. It would be interesting to evaluate the performance of non-linear discriminant analysis methods, such as quadratic discriminant analysis. It is also possible that performing some variable selection to eliminate unimportant variables would improve the model performance, so future work will involve implementing a feature reduction approach alongside LDA, such as the lasso method found in the R package 'penalizedLDA'.



(a) LDA with microbe features



(b) LDA with metabolite features



(c) LDA with combined features

Figure 3: Receiver operating characteristic (ROC) curves for each of the LDA classifiers.

References

[1] Romain Villéger, Amélie Lopès, Julie Veziant, Johan Gagnière, Nicolas Barnich, Elisabeth Billard, Delphine Boucher, and Mathilde Bonnet. Microbial markers in colorectal cancer detection and/or prognosis. *World Journal of Gastroenterology*, 24(22):2327–2347, Jun 2018.

[2] Jiyoung Ahn, Rashmi Sinha, Zhiheng Pei, Christine Dominianni, Jing Wu, Jianxin Shi, James J. Goedert, Richard B. Hayes, and Liying Yang. Human gut microbiome and risk for colorectal cancer. *Journal of the National Cancer Institute*, 105(24):1907–1911, Dec 2013.

[3] James J. Goedert, Joshua N. Sampson, Steven C. Moore, Qian Xiao, Xiaoqin Xiong, Richard B. Hayes, Jiyoung Ahn, Jianxin Shi, and Rashmi Sinha. Fecal metabolomics: assay performance and association with colorectal cancer. *Carcinogenesis*, 35(9):2089–2096, Sep 2014.

[4] Rashmi Sinha, Jiyoung Ahn, Joshua N. Sampson, Jianxin Shi, Guoqin Yu, Xiaoqin Xiong, Richard B. Hayes, and James J. Goedert. Fecal microbiota, fecal metabolome, and colorectal cancer interrelations. *PLoS ONE*, 11(3), Mar 2016.

[5] NCSS. Ridge regression.

[6] Sebastian Raschka. Linear discriminant analysis, Aug 2014.

[7] Tao Li, Shenghuo Zhu, and Mitsunori Ogihara. Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and Information Systems*, 10(4):453–472, Nov 2006.