Lasso classification

1.Introduction

Lasso classification select mOTU markers to identify the sample's category based on training set and testing set. In this report, it was based on R package (glmnet). By using K folds cross validation and a grid of values, glmnet can regularize parameter lambda. Through receiver operating characteristic curve (ROC curve), it can illustrates the diagnostic ability of a binary classifier system. The area under the curve (AUC) is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example, in most cases, combining ROC to evaluate the model's accuracy.

2.Data info

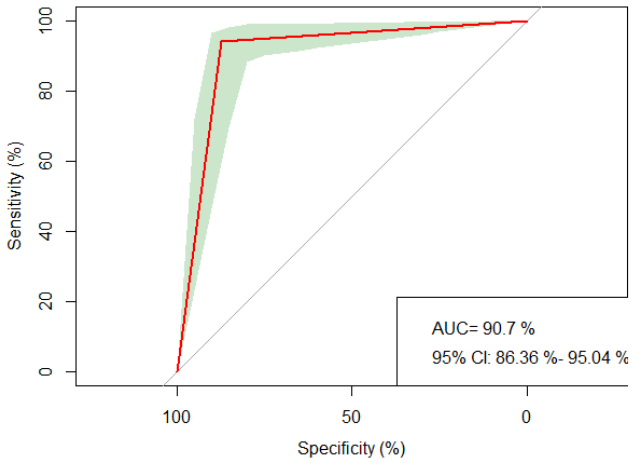
171 samples collected from community and Research center for mental disorders, includes 81 healthy controls (HC) and 90 schizophrenes (SZ). In SZ, 49 patients are first-episode, and 41 patients are relapse. 360 mOTUs profile which represents bacteria abundance was calculated by analyzing the sequencing data. As the metagenomics data, mOTUs profile contains numerous zero.

3.Procedure

There are two ways to make classifier, one is using total samples as training set, and using total samples as testing set (part I). The other is samples divided by first-episode and relapse, first-episode was used as training set and relapse was used as testing set (part II). All of the code was posted on Github.

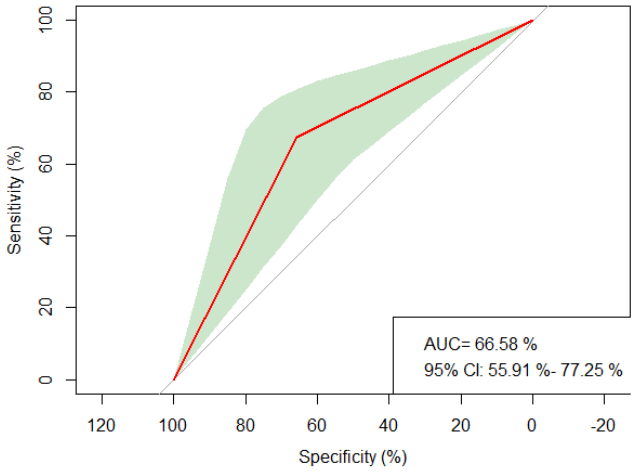
3.1 part I

5 folds cross validation and a grid of values are used to regularize parameter lambda. By setting seed 0, there are 155 samples were classified right, 11 samples were false-positive and 5 samples were false-negative. The AUC is 90.7%, and ROC curve is below. The classifier is well enough. What's more, markers got written down in file named 'Lasso_Classification_TotalSample_result'.



3.2 part II

Training HCs (44) and testing HCs (37) were randomly selected in proportion from total controls, and training HCs combined with first-episode samples as training set, testing HCs combined with relapse samples as testing set. By following part I methods, 52 samples were classified right, 12 samples were false-positive and 14 samples were false-negative. The AUC is 66.58%, and ROC curve is below. The classifier is a little bit better than stochastic prediction. What's more, markers got written down in file named as 'Lasso_Classification_TestTrainSample_result'.



4. Comparing with random forest classification

Classifier based on random forest theory has been developed as a general purpose two class classifier. With same data set as part I, 5 fold CV, random forest's AUC is about 95%, which a litter bit better than lasso. But there are small amount of intersection markers. As we discussed last few days, the correlation of these markers is high which is showed in file named as 'Lasso&rf_cc'.