Analysis of different machine learning algorithms using PIVUS data set

Shubhangi Kaushik

Abstract

The goal of the project:- It is to compare the results of three or more machine learning algorithms on metabolomics data and to find the most significant metabolites using the analysis.

Main results of the project:- 6 significant metabolites were found in the data set which had the most impact on the separation of the Male and Female samples.

Goal

The aim of this project was to create prediction models using multiple machine learning algorithms for analyzing metabolomics data and finding the most significant metabolites based on any three models of choice.

Data and Pre-processing

The metabolomics data for the project ID: MTBLS90 was collected from Metabolights database. The serum samples for people aged 70, and later aged 75 as well, were collected in Uppsala, Sweden for a study named Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS)[4]. The data set is publicly available. The Excel file was collected from CIMCB's git repository. The data set is a Microsoft Excel Worksheet consisting of 2 sheets. The first sheet named "Data Sheet" has information on 189 metabolite concentrations and their associated columns like "Idx", "SampleID", "Class" representing the sex of the sample, and "Sex". The second sheet named "Peak Sheet" contains the metadata related to each metabolite from the Data Sheet. The "Class" representing the sex of the sample was classified using the metabolite concentration.

Data was pre-processed first to make it suitable for building the models. It was split into the training and the test data set such that 1/3 of the data ended up as test data. It was done while making sure that the proportion of labels remained the same in both subsets. Each subset went through log transformation to normalize the data. Then this data was scaled using Z-scale. In case of any missing values, the data would be imputed through knn weighted mean method.

Methods

Optimized hyper-parameters were found through 5-fold cross-validation on the preprocessed data. The methods like Support Vector Machine(SVM), Principal Component Regression(PCR), Random Forest, and others, were trained and tested multiple times, on a set of user-defined parameters. The parameters giving the best results on the basis of R2Q2 and AUC plots were chosen.

In PCR, the optimal number of principal components were found through this method, whereas, for SVM, optimal C value was found, and leaf fraction was found for Random Forest.

These optimized values were then used for training the main model for each method. These models were evaluated on the basis of how well they predicted the label for both test and train data. In order to improve the working of the models, bootstrapping is performed on the training data set, which is then used to evaluate the model using the remaining data set. The evaluation done after bootstrapping seems to show improvement in the models' performance when compared to the evaluation without bootstrapping.

After this the important features(metabolites) were extracted from the models used, followed by the biological significance of those features and their correlation with the Y predicted values(Gender of the subjects). Below are the methods used:-

Support Vector Machine(SVM)

SVM[6] uses a hyperplane for classifying the data points into different groups by finding a direction of the hyperplane such that it maximizes the margin, i.e. the distance of the closest points(support vectors)from each group to the hyperplane. SVM utilizes hyper-parameters like C and gamma for finding the best margin. SVM(linear):- For determining the hyperplane, SVM linear requires the hyperparameter C. The margin for the hyperplane is inversely proportional to the value of C.

Principal Component Regression(PCR)

PCR[5] involves dimension reduction with the help of the projection of data points in such a way that it maximizes the distribution of the projection of the data points. After this step, principal components are obtained, which undergo linear regression against the dependent variable. One of the focus points of this method is dimension reduction, in order to reduce the complexity of analyzing and viewing data from higher dimensions.

Random Forest

Random Forest[1] is an ensemble technique that runs on top of the Decision Tree algorithm. Random Forest is an average of multiple Decision Trees. In addition to the random subsets of observations, random subsets of features are also used to create these Trees. The best features for splitting the data are the closest to the root node.

Results

After the training and hyper-parameter optimization of the model with the help of the training data set, AUC(Area under the curve) and R2Q2(R2:- coefficient of determination for the full data set;q2:- mean coefficient of determination for crossvalidation) graphs were obtained for all the three methods used. In Figures 1, 2 the graphs obtained from SVM are used. Likewise, the other algorithms also have AUC and R2Q2 graphs obtained.

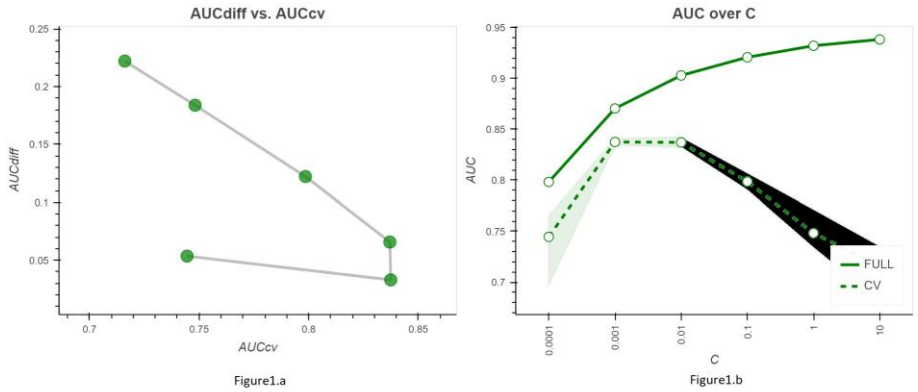


Figure 1 The AUC(Area under the curve) Score for each value of parameter C for the SVM. (1.a)The AUCdiff vs AUCcv shows the correlation between the AUC score for cross-validation, and the difference in AUC score of the full data set and cross-validation(1.b) Whereas the graph for AUC over C represents the AUC score for each C parameter ranging from 0.0001 to 10 for the full data set and cross-validation data set. In the AUC over C graph, where the performance of the model for the full data set seems to be very subtly growing with the increased value of C, the cross-validation line seems to experience a fall with the increase in the value of C Thus, figure 1 helps in understanding the performance of the model for each value of the parameter C.

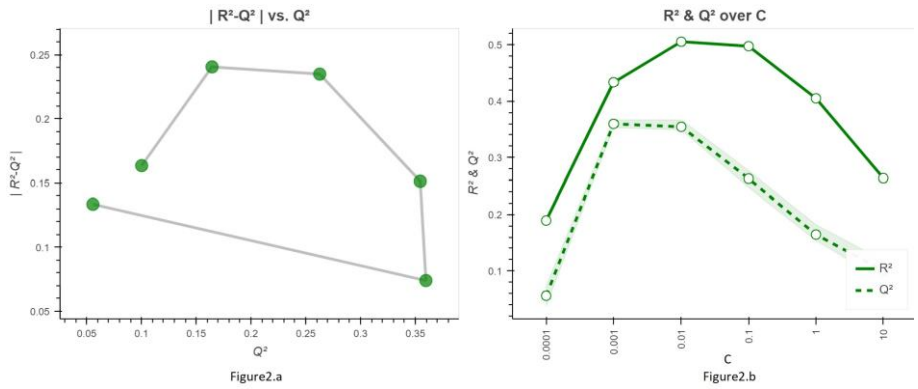
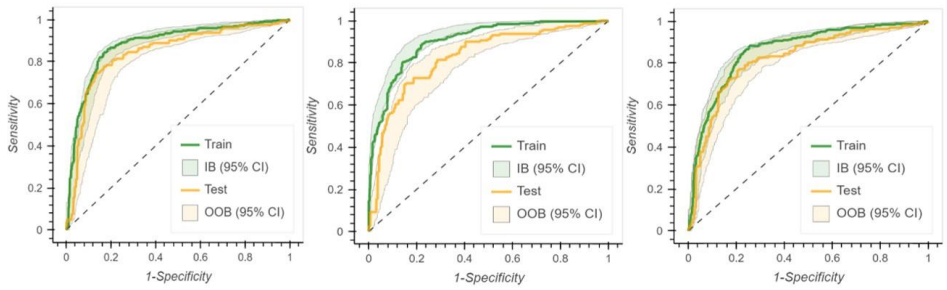


Figure 2 The R2 vs Q2 graph for each C parameter value using SVM. The optimal C value is chosen based on the inflection point of the graph[?]. This C value is later used for the model evaluation using a test data set

After the determination of the parameter values, the model evaluation with and without bootstrapping produced ROC(Receiving operation characteristic) graphs for all the methods described in Figure 3 with AUC scores as in Table 1



(a)Support vector machine (b)Random forest tree (c)principal component regression

Figure 3 ROC curves for (a)SVM, (b)RFT, (c)PCR describing the performance of the methods after bootstrapping. The green line indicates the train data set, and the yellow indicates the test data set. The green region indicates the in-bag samples and the yellow region depicts the out-of-bag samples.

AUC Scores				
ML Methods	Without Bootstrapping		With Bootstrapping	
	Test	Train	Test	Train
SVM	0.84	0.88	0.84(0.78,0.89)	0.88(0.84,0.91)
Random Forest	0.83	0.90	0.83(0.76,0.87)	0.90(0.90,0.94)
PCR	0.82	0.86	0.82(0.77,0.88)	0.86(0.81,0.89)

Table 1 The AUC scores helps in evaluating which model worked the best for the provided data set. The scores are noted for the methods with and without bootstrapping for both train and test data sets. In Table 1, the scores inside the brackets for 'with bootstrapping' column showcases the range of AUC score when bootstrapping is conducted

Built-in attributes for each model were used to find the most important features. "feature importances " was used for Random Forest. It is the measure of reduction in impurity when a branch splits on a particular predictor. "coef" was used for PCR and SVM. These are the coefficients obtained after fitting the model and represent the relationship between the input variables and the target variables.

Top ten most significant metabolites were found using the three models. 6 out of 10 metabolites from Table 2 were common among the top 10 features for all three.

6 common Metabolites found among the ML Methods used				
Metabolites	Labels	Random Forest	SVM	PCR
M142	Phosphatidylcholine(28:2)	0.097414	0.077677	0.023482
M141	Ceramide phosphoethanolamine(35:2) Sphingomyelin(32:2)	0.086343	0.078281	0.023616
M10	Creatine	0.078261	0.105341	0.022035
M2	Creatinine	0.070354	0.092506	0.019407
M152	Sphingomyelin(d18:2/18:1)	0.045366	0.065498	0.020830
M15	2-Ketohexanoic acid	0.032682	0.063428	0.019914

Table 2 The most significant metabolites common for the three models. For Random Forest, the values represent the measure of reduction in impurity when the decision is made on that metabolite. For SVM and RBF, the values represent the absolute value for the coefficient obtained after fitting each of the models.

Discussion

Phosphatidylcholine(M142), a vital component of cell membranes, is involved in the metabolism of choline. Notably, females have been found to possess higher levels of choline compared to males[3], which has been associated with estrogen regulation.

Sphingomyelins(M141 and M152) play a crucial role in maintaining the structural integrity and fluidity of cell membranes. The levels of sphingomyelins are found to be higher in older women[7] and increase rapidly in females around menopausal age[2] which is also consistent with our data set's demographics. This change with age cannot be observed in males.

Even though there is not enough evidence suggesting the role of Ceramide phosphoethanolamine(M141) suggesting differences in its level based on the sex, however, it might be higher in males because of its functional and structural similarities with Sphingomyelins.

Creatine(M10) is mainly stored in muscle tissue and has a significant role in the energy metabolism of short, intense muscle activities. It also contributes indirectly to the production of Creatinine(M2). While there is no conclusive evidence indicating different levels of these metabolites between males and females, it is possible that males, who generally have more muscle mass than females, may have higher levels of these metabolites.

No evidence suggesting a difference in 2-Ketohexanoic acid (M15) levels in the two sexes could be found. Even the score in Table 1 was the lowest for it.

Based on the analysis, Phosphatidylcholine(M142) and Sphingomyelins(M141 and M152) were identified as the most noteworthy metabolites, as not only those were the top common metabolites predicted by the ML methods used but I were also able to find research papers supporting the claim. Hence providing substantial evidence to support the validity of our results obtained through the algorithms.

References

1. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
2. Xiujuan Cui, Xiaoyan Yu, Guang Sun, Ting Hu, Sergei Likhodii, Jingmin Zhang, Edward Randell, Xiang Gao, Zhaozhi Fan, and Weidong Zhang. Differential metabolomics networks analysis of menopausal status. *PloS one*, 14(9):e0222353, 2019.
3. Leslie M Fischer, Kerry Ann DaCosta, Lester Kwock, Paul W Stewart, Tsui-Shan Lu, Sally P Stabler, Robert H Allen, and Steven H Zeisel. Sex and menopausal status influence human dietary requirements for the nutrient choline. *The American journal of clinical nutrition*, 85(5):1275–1285, 2007.
4. Stacey N. Reinke David I. Broadhurst Kevin M. Mendez. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolics*, 2019.
5. RX Liu, J Kuang, Q Gong, and XL Hou. Principal component regression analysis with spss. *Computer methods and programs in biomedicine*, 71(2):141–147, 2003.
6. William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.