RESEARCH

Paper 2 in applied machine learning

Aakash Nepal and Cuong Gia Pham

Abstract

Goal of the project: Generate 11 data(1 original, 4 dirty, and 6 imputed). Then apply 2 Machine Learning(ML) models to those data and extract 5 metabolites.

Main results of the project: The 4 dirty datasets without imputation were causing errors. The imputation methods may improve the performance of machine learning models on the training dataset, but they may not necessarily generalize well to new, unseen data.

Personal key learning :

1 Pham Gia Cuong: One hot Encoder, making data dirty, mean and median imputation.
2 Aakash Nepal: learned about making data dirty, mean and median imputations, missingness in data, and the problems with it for ML.

Estimation working hours:

1 Pham Gia Cuong: 7 hours per week
2 Aakash Nepal: 7 hours per week.

Project evaluation: 1

Number of words : Approx. 1400(without abstract, captions and appendix)

1 Scientific Background

Metabolomics is a kind of important field of study that focuses on improving the understanding of disease mechanisms to help develop individualized medical treatments for the disease. Machine Learning algorithms are helpful for bio-marker identification and the development of predictive models for disease treatment and diagnosis. Machine learning algorithms are subjected to have an impact on many varieties of factors, including the algorithm chosen, the complexity, and also the missingness of the data being analyzed, so it is necessary to evaluate these algorithms against the novel, previously unseen data.

2 Goal of the project

The goal of this project was to find the difference in the results of two machine learning models (Random forest, Linear Support Vector Machine) between the datasets that are generated from different imputation methods.

3 Data

For this project, the dataset "MTBLS136" has been used, a serum LC-MS dataset with 949 metabolites. The dataset describes the amount or concentration of different metabolites between two groups; using medications that only impact the estrogen signaling pathway and medications that impact the estrogen plus progesterone signaling pathway in women. The estrogen group has 337 samples and has been represented as a case group. The estrogen plus progesterone group has 331 samples

and has been represented as a control group. For preprocessing, the selected data was read into two tables called "DataTable" and "PeakTable" using the function " cb.utils.load_dataXL". The "Idx" column from the DataTable was dropped and the "SampleID" column was made as an index. In the original dataset, there were already 310432 missing values (around 20%), and some of them were visualized using a heatmap plot with the help of the seaborn package. For randomly removing 10% of values, a NumPy matrix was created using a probability of 0.1 containing randoms of the condition "True" and 0.9 for "False" of the size of the DataTable. Using this matrix the mask() function from the pandas library was used to replace random values in the DataTable with NANs. The same was done for removing 50% creating a matrix which was created using a probability of 0.5 each for both of the conditions and then the mask() function was used similarly. For randomly changing 10% of values same matrix for randomly removing 10% of values was used and the mask() function was used to replace values randomly in the DataTable with values between 500000 and 8000000 as our metabolites had similar lower and upper bounds. Similarly, for randomly changing 50% of values in the data same matrix as for randomly removing 50% of values was used and the mask() was used to replace the values randomly similar to as for randomly changing 10% of values. In total 4 dirty datasets were created.

4 Data Preprocessing

4.1 Mean and Median Imputation

Mean and median imputation methods are techniques to replace missing values in the dataset. The mean imputation method uses the mean of each column of the features and the median imputation method uses the median of each column of the features to replace missing values in the dataset. If the data is skewed, then it is better to use median imputation and if the data is normally distributed, also mean imputation method can be used. For this project, the numeric data from the dataset were extracted. After that the NANs were imputed using their means using mean() and fillna() functions which are found in the pandas library. And for the non-numeric columns such as the Class column, they were imputed using mode() function, which imputes the NANs with the most frequently occurring class. Both imputed non-numeric and numeric columns were then combined. In total 2 data sets were cleaned using each imputation method.

4.2 Removing invalid rows method

For removing invalid rows those the dropna() function from the pandas library was utilized for those two datasets (with 10% removed values and 50% removed values) that we also chose to impute. The missing values in the original dataset were already quite plenty and distributed equally to all rows, hence after using this removing invalid rows method, there are no rows left to apply for the machine learning model. For that reason, we will not go further with the dataset from this method.

5 Data Analysis

Linear Kernel Support Vector Machine (SVM-Lin) and Random Forest (RF) were the two machine-learning models used in this project. The main goal of the SVM-Lin

is to find a hyperplane with a maximal margin in N-dimensional space to classify a dataset. Unlike SVM-lin, RF uses multiple unique classifiers called decision trees and averages them to generate a single prediction. For both of these model, a 5-fold cross-validation was performed to compute performance metrics(AUC and R2Q2) and select the optimal hyperparameters. After that bootstrapping was used to determine the confidence intervals of the training and test evaluation metrices.

6 Results and Discussion

6.1 Making data dirty

For all four dirty datasets, we were unable to use both Machine Learning algorithms as it got an error due to NANs in the input dataset. From Figure 1A, we can see

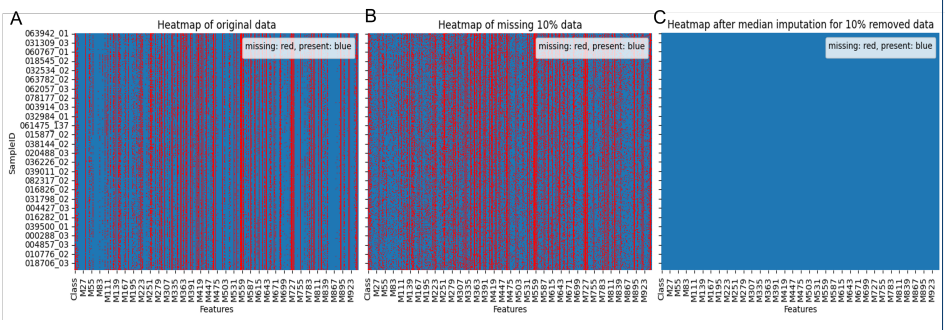

Figure 1 Heatmaps depicting the missing values of datasets. A: missing values in the original dataset, B: missing values in the dirty dataset, C: missing values in the dataset after imputation.

that the original dataset "MTBLS136" contains already so many missing values. According to our calculation, the total missing values are around 20%. And hence, adding 10% or 50% of missing values led us to a situation, where many columns contain only missing values as we see in Figure 1B. Besides that, randomly replacing 10% or 50% for making data dirty reduced the number of missing values a little bit. However, the replaced values are sometimes too high or too low. In this case, we should mark out all the outliers of each column and impute them. Because we have reached our time limit, and hence we have limited our replacing generated values between 500000 and 8000000 as our metabolites had similar lower and upper bounds.

6.2 AUC evaluation

After producing an extra 10% of missing values in the dataset and using different imputation methods, such as mean and median, two machine learning models, SVM and RF, were trained on all datasets. The results in Figure 2 show that the ROC for both the train and test datasets were similar for all imputed datasets, and even slightly better compared to the ROC on the training dataset of the original data(see. Appendix C), which was imputed using KNN with 3 clusters. However, when comparing the ROC on the test dataset of the original data, both machine learning models showed better performance.This suggests that while imputation methods may improve the performance of machine learning models on the training dataset, they may not necessarily generalize well to new, unseen data. Different

from the previous datasets, with 50% of missing values or 50% of replacing values datasets we can clearly see that the AUC values on training data are abnormally high, especially with RF algorithm(see. Appendix A). However, the confidence

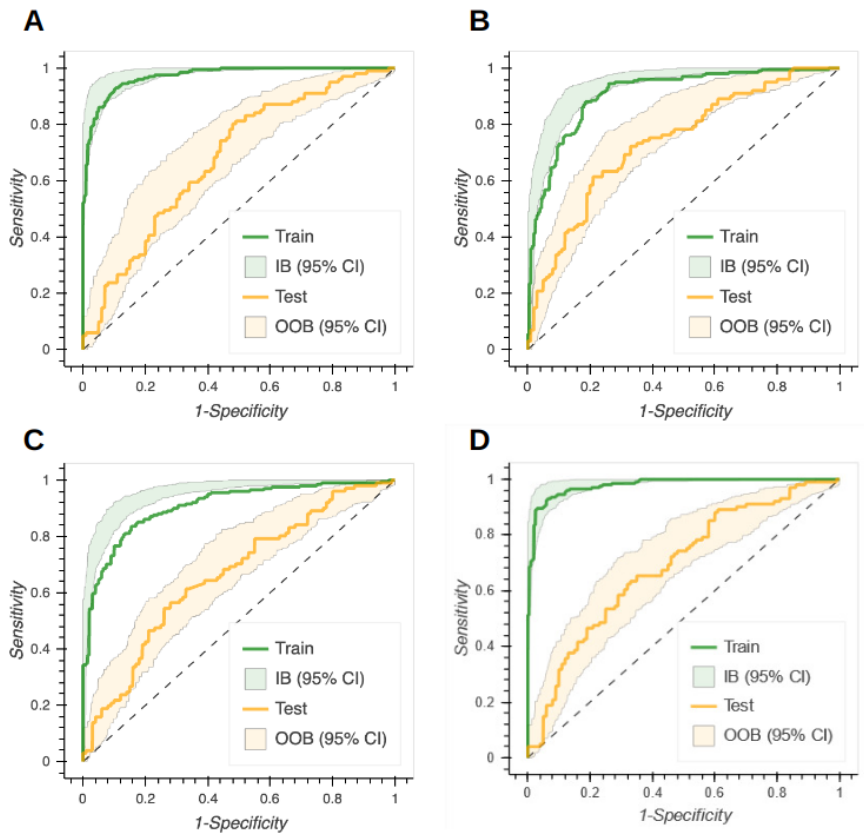

Figure 2 The ROC curves depicting the performance of support vector machine (SVM) and random forest (RF) machine learning models on different datasets

- **A**: SVM with 10% missing values dataset, where all missing values were imputed by the mean. The AUCtest value is 0.74.
- **B**: RF with 10% of missing values dataset, where all missing values were imputed by the mean. The AUCtest value is 0.66.
- **C**: SVM with 10% of missing values dataset, where all missing values were imputed by the median. The AUCtest value is 0.74.
- **D**: RF with 10% of missing values dataset, where all missing values were imputed by the median. The AUCtest value is 0.67.

interval of AUC on the test set is broad and infers an uncertainty of the result. This suggests that with so many missing values, that need to be imputed by either mean or median, the training process is always overfitted because of so many similar data points.

6.3 Important metabolites

As shown in Table 1, the metabolites detected by SVM-Lin and RF are mostly different although there are some similarities such as 1-linoleoyl-GPA (18:2)*. It

Table 1 Table showing the top 5 metabolites (M) discovered by two Machine Learning methods (SVM-Lin and RF) in the original dataset and the dataset imputed by mean.

Original data		
M	**SVM-Lin**	**RF**
1	1-linoleoyl-GPA (18:2)*	1-stearoyl-2-arachidonoyl-GPC(18:0/20:4)
2	lysine	1-linoleoyl-GPA (18:2)*
3	2-aminoheptanoate	1-linoleoyl-2-arachidonoyl-GPC(18:2/20:4n6)*
4	4-acetamidobutanoate	1-linolenoylglycerol(18:3)
5	2'-deoxyuridine	2-linoleoylglycerol(18:2)
Imputed 10% removed		
M	**SVM-Lin**	**RF**
1	1-linoleoyl-GPA(18:2)*	1-linoleoyl-GPA (18:2)*
2	lysine	1-arachidonylglycerol (20:4)
3	gamma-glutamyl-alpha-lysine	gamma-glutamyl-alpha-lysine
4	2-aminoheptanoate	21-hydroxypregnenolone disulfate
5	Isobar: fructose 6-phosphate	1-stearoyl-2-arachidonoyl-GPC (18:0/20:4)
Imputed 50% removed		
M	**SVM-Lin**	**RF**
1	theophylline	erythronate*
2	threonine	1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)
3	1-linoleoyl-2-arachidonoyl-GPC	pimelate (heptanedioate)
4	2-aminobutyrate	cholesterol
5	androstenediol(3beta,17beta)disulfate(2)	1-linoleoylglycerol (18:2)

can be that the choice of algorithm can significantly impact the determination of the important features. When we compare the metabolites for the data with 10% of values imputed by mean with metabolites detected for original data, we can see that it becomes hard to understand the biological significance of metabolites due to the loss of the information which was reliable for this dataset. if we compare it furthermore for 50% of values removed, it becomes even more confusing. For example, if we see lysine that is detected by SVM-Lin, it was found that it has differed significantly in the association between estrogen-only and estrogen plus progesterone users [1], when we remove 10% and impute it then it remains there but when we remove 50% and impute, this biological significance is lost. Other similar examples are 1-linoleoylglycerol(18:3) and 2-linoleoylglycerol(18:2) which are detected by RF and get lost already after the removal of 10% values and imputation. We also had similar observations for median imputation(see. Appendix B), however, SVM-Lin seems to detect the same metabolites for both of the imputations(mean and median). In overall, It seems like both mean and median imputation doesn't always help to make the data cleaner, if more information in the input dataset is missing, but also may introduce some noise in the data.

7 Contributions

1 Pham Gia Cuong: Overall 50% report and did the work with mean imputation and changing values plus ML.

2 Nepal Aakash: Overall 50% report and did the work with median imputation and removing values plus ML.

References

[1] Victoria L Stevens et al. "Serum metabolomic profiles associated with post-menopausal hormone use". In: *Metabolomics* 14.7 (2018), p. 97. DOI: 10.1007/s11306-018-1393-1.

Appendix A: 50% of missing values data

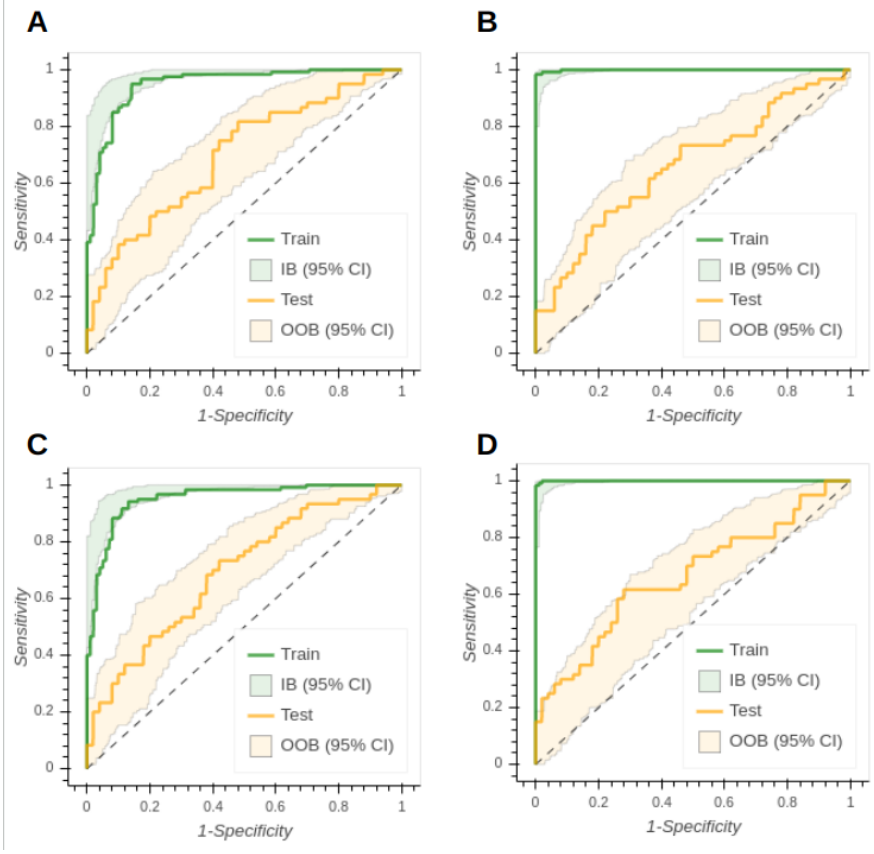

Figure 3 The ROC curves depicting the performance of support vector machine (SVM) and random forest (RF) machine learning models on different datasets

- **A**: SVM with 50% missing values dataset, where all missing values were imputed by the mean. The AUCtest value is 0.69.
- **B**: RF with 50% missing values dataset, where all missing values were imputed by the mean. The AUCtest value is 0.57.
- **C**: SVM with 50% missing values dataset, where all missing values were imputed by the median. The AUCtest value is 0.70.
- **D**: RF with 50% missing values dataset, where all missing values were imputed by the median. The AUCtest value is 0.67.

Appendix B: Median imputation data

Table 2 Table showing the top 5 metabolites (M) discovered by two Machine Learning methods (SVM-Lin and RF) in the original dataset and the dataset imputed by median.

Original data		
M	**SVM-Lin**	**RF**
1	1-linoleoyl-GPA (18:2)*	1-stearoyl-2-arachidonoyl-GPC(18:0/20:4)
2	lysine	1-linoleoyl-GPA (18:2)*
3	2-aminoheptanoate	1-linoleoyl-2-arachidonoyl-GPC(18:2/20:4n6)*
4	4-acetamidobutanoate	1-linolenoylglycerol(18:3)
5	2'-deoxyuridine	2-linoleoylglycerol(18:2)
Imputed 10% removed		
M	**SVM-Lin**	**RF**
1	1-linoleoyl-GPA(18:2)*	1-linoleoyl-GPA (18:2)*
2	lysine	oleoyl-arachidonoyl-glycerol (18:1/20:4) [2]*
3	2-aminoheptanoate	gamma-glutamyl-alpha-lysine
4	gamma-glutamyl-alpha-lysine	2-aminoheptanoate
5	Isobar: fructose 6-phosphate	2-linoleoylglycerol (18:2)
Imputed 50% removed		
M	**SVM-Lin**	**RF**
1	theophylline	1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)
2	threonine	1-arachidonoyl-GPC (20:4n6)*
3	androstenediol(3beta,17beta)disulfate(2)	linoleoyl-linoleoyl-glycerol (18:2/18:2) [1]*
4	1-linoleoyl-2-arachidonoyl-GPC(18:2)	4-guanidinobutanoate
5	2-aminobutyrate	1-palmitoyl-GPA (16:0)

Appendix C: ML results of original data

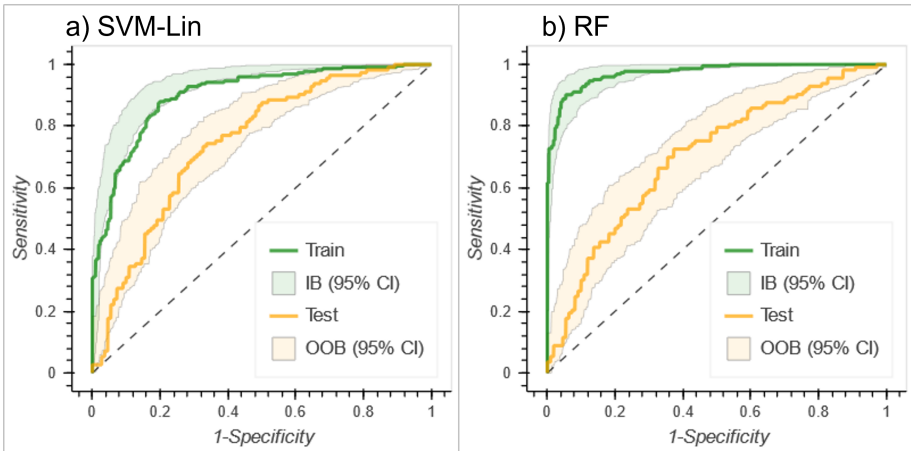

Figure 4 The ROC curves depicting the performance of three different machine learning algorithms (a) SVM-Lin, and b) RF in predicting test results. The green line with shading around it represents the in-bag 95% confidence interval for the ROCtrain and the yellow line with shading represents the out-of-bag 95% confidence interval for the ROCtest. The AUCtest values for the two algorithms are 0.75 (a and b).