BST 260 Final Project

Xiao Gu

Introduction

Ischemic heart disease (IHD) has been identified as a leading cause of death globally (Ref). Compelling evidence showed that lifestyle changes could be effective strategies for secondary preventions of IHD (Ref). Therefore, to reduce the burden of IHD mortality, an efficient tool for IHD screening and early diagnosis is warranted. A machine learning algorithm that is developed with serum metabolites, cardiometabolic biomarkers, and self-reported phenotypic data is promising in simplifying the process and reduce the cost of IHD screening/diagnosis. IHD status could be accurately detected with a simple blood draw and metabolomic profiling. In this project, I aim to develop such an algorithm using data from a European population.

I will use data from the MetaCardis consortium that recruited participants aged 18-75 years from Denmark, France, and Germany (Ref). The data was published early this year as the supplementary material of an article on Nature Medicine (Ref). The original study included 372 individuals with IHD. These IHD cases were further classified into acute coronary syndrome (n = 112), chronic ischemic heart disease, (n = 158), and heart failure (n = 102). With a case-control design, the study also included 3 groups of controls matched on varies factors. The raw data includes 1,882 observations including repeated records with the same participant ID but different case-control status.

For this project, I will use records from the 372 IHD cases and 372 controls matched on type 2 diabetes (T2D) status and body mass index (BMI). I will also extract data for age, gender, fasting plasma triglycerides, adiponectin, and CRP, systolic and diastolic blood pressure, left ventricular ejection fraction, physical activity level, and 1,513 log-transformed values of serum metabolites.

Exploratory data analysis

After reading in the data, I first filtered the observations to keep the IHD cases and their controls matched by T2D status and BMI. I then merged metabolites data with cardiometabolic biomarkers and self-reported phenotypic data to create a `main` dataset with 744 rows and 1522 columns. I noticed that several participants do not have any metabolites data, and therefore, need to be removed. Additionally, around 30% of participants had missing values for left ventricular ejection fraction and physical activity level. Many machine learning techniques could not be implemented with that many missing and it would also not be appropriate to replace the missings with any arbitrarily selected value. So I removed these two potential predictors from my analyses. Finally, for variables with less than 10% missing data, I replaced the missing values with the median of the non-missing data. The cleaned `main` dataset had 603 rows and 1522 columns.

I then preprocessed the data to remove non-informative predictors with near-zero variances. Given that I planned to train as least one of my algorithms with regression, it would be better to have more predictors normally distributed so that model efficiencies could be improved. I tested the normality of each predictor with Shapiro-wilks Test and summarized the p-values. I found that only 101 predictors are normally distributed. It is also note-worthy that the metabolite values from the raw data were all log-transformed. Obviously, log-tranformation did not normalize the distributions successfully. So I transformed all metabolite values back to the original scale and used rank-based inverse normal transformation (INT) to normalized the distributions instead. As examples, histograms showing the distributions of oleoylcarnitine (C18:1) and S-methylcysteine sulfoxide before and after the transformation were shown. I ended up having 840 predictors normalized successfully.

Methodologies to use

The outcome that my algorithm aimed to predict is the binary IHD status (non-case = 0, case = 1). Considering that I had 1422 predictors, I would use principle component analysis (PCA) to reduce dimensions. I would keep principle components that account for at least 70% of variability as new predictors, and train a model with logistic regression, and a model with K-nearest neighbor (KNN). Given that the principle components are hard to interpret and algorithms developed based on PCA could be difficult to implement, I would train another KNN model with all 1422 predictors instead. Random forest would be the 4th training method I would use. Finally, I will use ensemble to combine the results of all four algorithms. For all algorithms, I would evaluate the overall accuracy, sensitivity, specificity, and ROC curve. I would also use cross-validation and bootstrapping to tune the model parameters.

Results

Conclusion

Reference

Appendix

```
## # A tibble: 6 x 1,524
##   ID       case   age tag   adipo~1   crp   sbp   dbp Gender  lvef   act acetate
##   <chr>   <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1 x14MCx~     0    48 1.00     5.01 0.897  104   60.5      0    60  1.25   -3.91
## 2 x14MCx~     0    49 1.00     4.03 1.11   111   70        0    NA  1.02   -3.91
## 3 x14MCx~     0    54 1.48     6.26 2.05   106.  68.5      1    67  8.75   -3.51
## 4 x14MCx~     0    47 0.787    3.44 0.67   138   78        1    67  4.12   -3.91
## 5 x30MCx~     0    50 0.54     7.82 2.63   154   91.5      0    NA 20.6    NA
## 6 x30MCx~     0    66 0.59    11.0  0.427  110.  65.5      0    NA 14      -3.91
## # ... with 1,512 more variables: acetone <dbl>, artemisin <dbl>,
## #   `beta-sitosterol` <dbl>, betaine <dbl>, `betaine-aldehyde` <dbl>,
## #   butyrylcarnitine <dbl>, catechol <dbl>, cellotetraose <dbl>, choline <dbl>,
## #   `D-trehalose` <dbl>, `D-lyxose` <dbl>, `D-malate` <dbl>,
## #   `D-sorbitol` <dbl>, `D-threitol` <dbl>, decanoylcarnitine <dbl>,
## #   glyceraldehyde <dbl>, ethanol <dbl>, ethanolamine <dbl>, formate <dbl>,
## #   glucoheptonate <dbl>, glycolate <dbl>, halostachine <dbl>, ...
```

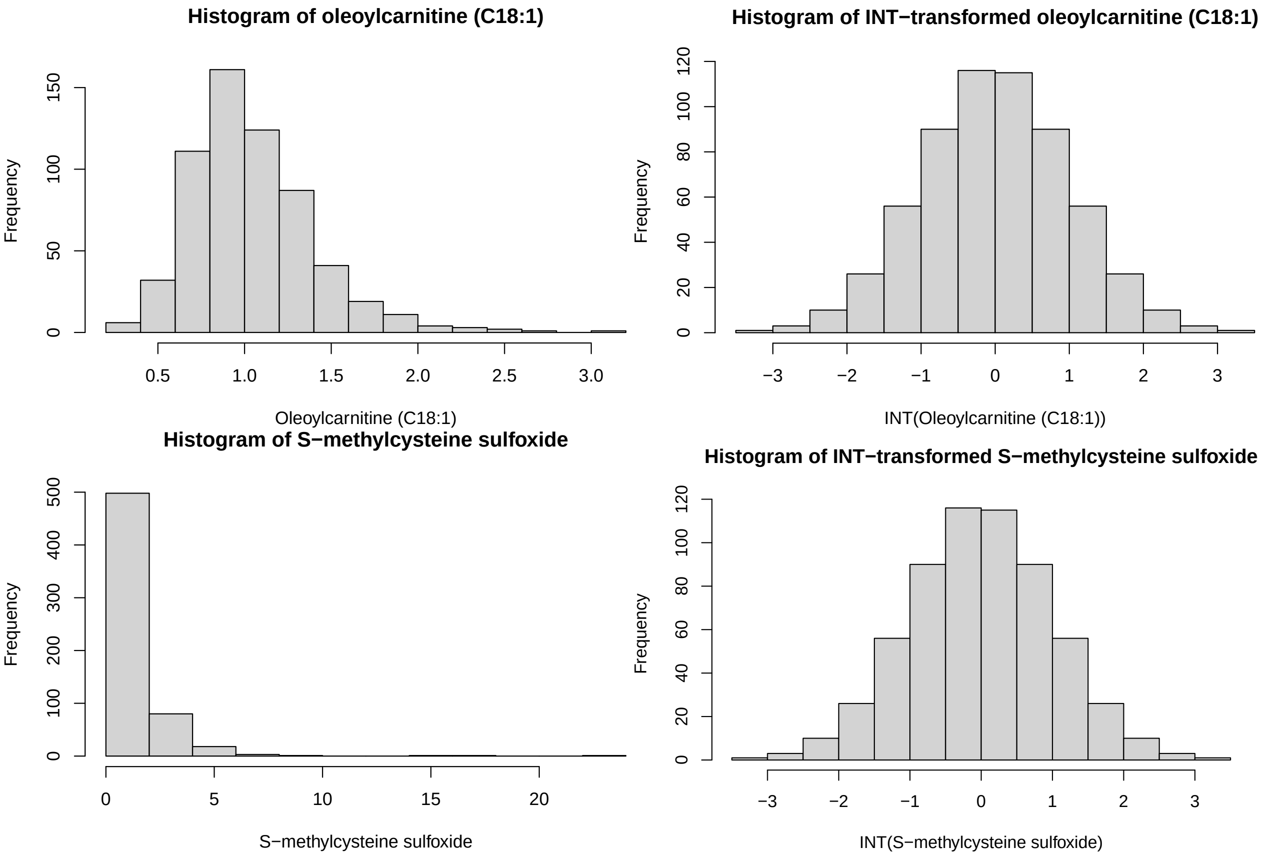
```
##              sapply(main, pctmiss)
## ID                      0.00000000
## case                    0.00000000
## age                     0.00000000
## tag                     0.04704301
## adiponectin             0.05645161
## crp                     0.05779570
```

```
## # A tibble: 6 x 1,522
##   ID          case   age tag   adipon~1   crp   sbp   dbp Gender acetate acetone
##   <chr>      <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl>   <dbl>
## 1 x14MCx1158     0    48 1.00      5.01 0.897  104   60.5      0   -3.91   -3.91
## 2 x14MCx2932     0    49 1.00      4.03 1.11   111   70        0   -3.91   -3.22
## 3 x14MCx2498     0    54 1.48      6.26 2.05   106.  68.5      1   -3.51   -3.51
## 4 x14MCx2237     0    47 0.787     3.44 0.67   138   78        1   -3.91   -4.95
## 5 x30MCx1828     0    66 0.59     11.0  0.427  110.  65.5      0   -3.91   -3.91
## 6 x30MCx1314     0    54 1.41      2.6  1.4    128.  75.5      1   -2.81   -3.91
## # ... with 1,511 more variables: artemisin <dbl>, `beta-sitosterol` <dbl>,
## #   betaine <dbl>, `betaine-aldehyde` <dbl>, butyrylcarnitine <dbl>,
## #   catechol <dbl>, cellotetraose <dbl>, choline <dbl>, `D-trehalose` <dbl>,
## #   `D-lyxose` <dbl>, `D-malate` <dbl>, `D-sorbitol` <dbl>, `D-threitol` <dbl>,
## #   decanoylcarnitine <dbl>, glyceraldehyde <dbl>, ethanol <dbl>,
## #   ethanolamine <dbl>, formate <dbl>, glucoheptonate <dbl>, glycolate <dbl>,
## #   halostachine <dbl>, hydroquinone <dbl>, isovalerylcarnitine <dbl>, ...

## [1] 1422

##
##    0    1
## 1321  101

##
##   0   1
## 582 840
```


Histogram of oleoylcarnitine (C18:1)
Frequency
Oleoylcarnitine (C18:1)
Histogram of INT−transformed oleoylcarnitine (C18:1)
Frequency
INT(Oleoylcarnitine (C18:1))
Histogram of S−methylcysteine sulfoxide
Frequency
S−methylcysteine sulfoxide
Histogram of INT−transformed S−methylcysteine sulfoxide
Frequency
INT(S−methylcysteine sulfoxide)