Discovery of Biomarkers for Coronary Microvascular Disease

Saumya Agrawal, Alicia Arredondo Eve, Justina Zurauskiene, and Zeynep Madak-Erdogan
Student Pushing Innovation Program, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

INTRODUCTION

Coronary microvascular disease (CMD) is a heart disease that affects the walls and inner lining of tiny coronary artery blood vessels that branch off from the larger coronary arteries. This research is vital because although CMD is vastly common in women, the only method of identification we have is the elimination of the other possibilities such as coronary artery disease.

Biomarkers: The term "biomarker," refers to a broad subcategory of medical signs – that is, objective indicators of medical state observed from outside the patient which can be measured accurately and reproducibly.

AIM

This research is vital because currently, the identification of CMD is subject to the elimination of the other similar possibilities such as Coronary Artery Disease (CAD).

With definite biomarkers for CMD, we could eliminate the need for a stress test, or a coronary angiography by optimizing testing and treatment.

DATA OVERVIEW

The raw data consisted of 71 patients and 594 features. These features were in the form of targeted and untargeted metabolites, and general health data.

METHOD

Pre-processing Pipeline

The data needs to be cleaned prior to the analysis. We did this programatically in Python, so that the code can be re-used for future similar datasets.

- Creating reduced models by removing features with more than 20% and 40% missing values.
- Feature Scaling using Min-Max Normalization
- Gaussian Mixture clustering to find ideal K value for K nearest neighbor imputation
- Performing KNN imputation and creating two datasets with 20% and 40% imputed values.

Biomarker Discovery

In Python, a dictionary of three data frames was created with each data frame only containing data from one of the groups.

After that, a difference in means T test was performed for each of the features between the groups.

Machine Learning

Machine learning algorithms were used to validate the efficacy of the discovered biomarkers.

Three models were created: Logistic Regression training model, Decision Tree training model, and Random Forest training model. Based on their respective efficacies, the Logistic Regression Model was picked for this study.

Logistic Regression Model: Specifications

- Penalty = ‘l1’. L1 or Lasso regularization was performed which reduces overfitting
- C = 0.8, Inverse regularization strength
- Class Weight = balanced.

Solver = ‘liblinear’ is an algorithm which applies automatic parameter selection

Receiver Operator Curve (ROC) and AUC score

A receiver operator curve shows the trade off between sensitivity (or true positive rate) and specificity (1-false positive rate).

Area under curve is a common approach to judge the ROC curve. In general, an AUC of 0.5 suggests no discrimination, 0.7 to 0.8 is considered acceptable

We can see from the above plotted curves, only the dataset with only 20% of the data imputed has an acceptable AUC score, so henceforth only that dataset is used.

Graphical Verifications

The heatmap above shows the mean values of each biomarker for the CMD and Non CMD groups. We can see that some biomarkers have a very significant difference

On the right side, we have the Principal Component Analysis performed on all the biomarkers

RESULT

This analysis discovered 65 biomarkers for Coronary Microvascular disease.

The ROC and AUC graphs help us choose the dataset with imputation performed only on the dataset with features with at most 20% missing data.

The Logistic Regression Model had a training accuracy of 87.76% and a testing accuracy of 86.72%. This indicated that overfitting was overcome in the machine learning Model.

CONCLUSIONS

One sided tests also need to be performed to identify what biomarkers lean to what level for each group. For example, does a higher or lower level of Lactamide indicate the presence of Coronary Microvascular disease? We can also use the same methodology for the discovery of biomarkers for other diseases.

ACKNOWLEDGEMENTS

We would like to acknowledge the National Center for Supercomputing Applications and their Students Pushing Innovation Program.