Variable Screening in Machine Learning - Analysis Pipelines for a Metabolomic Dataset

Ngoc Duong - nqd2000

Introduction

Alzheimer's Disease (AD) is a common and serious brain disorder that destroys memory and thinking skills, especially among the elderly.

With the advancement in high-throughput technologies, we are able to collect and analyze the characteristics of this complex disease at a molecular level.

The aim of this project is two-fold:

- Utilize metabolomic data to predict incident AD
- During this process, investigate potential analysis pipelines (with and without different variable screening steps) and observe the resulting behaviors of classification performance.

Data

The data comes from PPG-DS study in which patients with Down Syndrome were followed over the duration of 9 visits (1-1.5 years apart).

Untargeted plasma metabolomic data were obtained from 297 of the patients when they enroll. There are measurements on about 2700 metabolites' expression levels for each subject.

The outcome is incident Alzheimer Disease (AD), which is obtained from clinical diagnoses (gold standard). Among 297 patients with available metabolomic data, 267 have clinical diagnoses (60 AD and 183 no AD)

Methods

For the first goal (prediction): compare classification performances of ML methods such as Elastic Net, PLS, SVC, Random Forest, and XGBoost

For the second goal: look at the performance of the methods on each subset of variables selected at the previous screening step(s).

Metrics include cross-validated ROC AUC and Precision-Recall (10-fold CV), and elapsed time for the second goal.

Screening methods considered

Univariate screening:

- Differential Expression Analysis (Bioconductor's limma)
- EBayes moderated t-statistic (correlated metabolites)
- Select metabolites with FDR-adjusted p-values < 0.05

Multivariate screening:

- PCA - top 100 metabolites with highest loadings

Due to speed and variable selection ability, PLS and Elastic Net are also considered.

- PLS - top 100 metabolites with highest VIP scores
- Elastic Net - selected features with coefficients shrunk to 0

Notes: these are built-in, or embedded feature selection, meaning that the model will only include predictors for a specific optimization problem at hand. Upside: fast implementation.

Potential Pipelines

1) Univariate screening -> Multivariate screening -> Learner/Classifier
2) Univariate/Multivariate screening -> Learner/Classifier
3) Learner/Classifier on full data

Reduced time and computational cost if relevant features are selected during univariate/multivariate screening step.

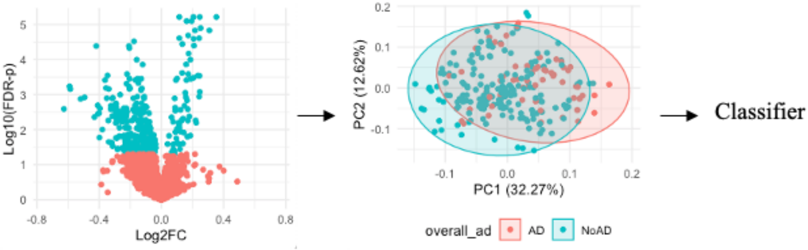

Figure 1: Pipeline

Preliminary Resutls

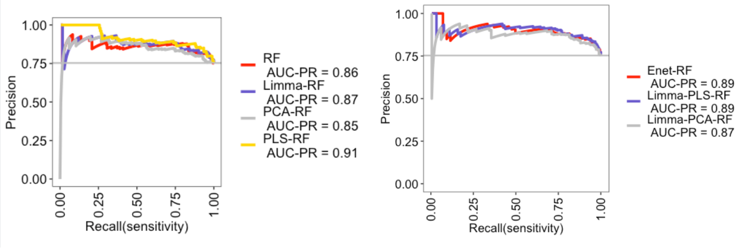

Figure 2: PR AUC for RF on different pipelines

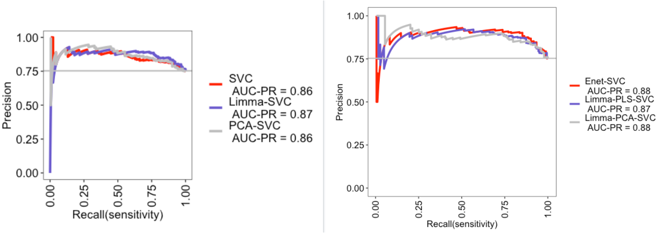

Figure 3: PR AUC for SVC on different pipelines

Preliminary Results

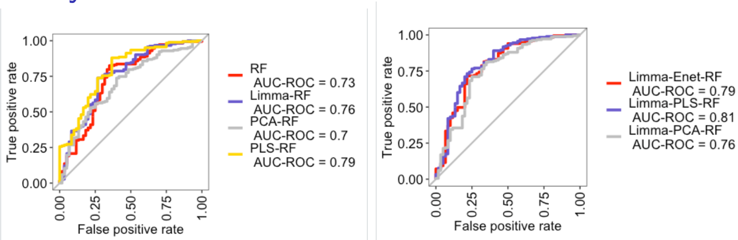

Figure 4: ROC AUC for RF on different pipelines

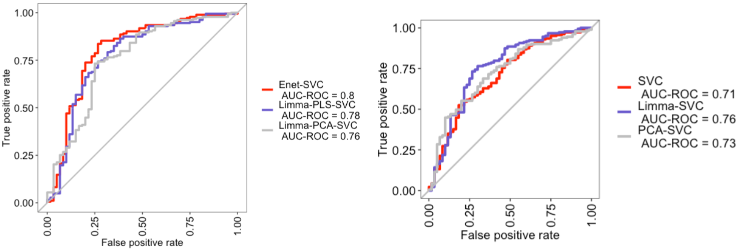

Figure 5: ROC AUC for SVC on different pipelines

Preliminary Results

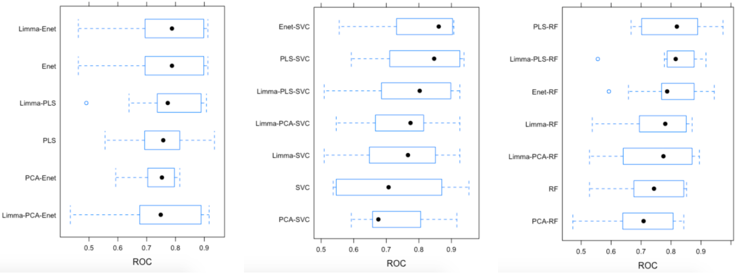

Figure 6: ROC AUC from resamples on different pipelines

Discussion

Univariate screening step can help improve prediction and lessen cost (potential to select important features and reduce computation time)

Using learners with embedded variable selection to screen variables might not guarantee good properties (if the final learner evaluates variable importance differently) -> risk of losing important variables. However, this approach is still employed (1-3).

For this dataset, there seems to be improvement in classification performance when using learners to screen variables, e.g., feeding the selected features from Enet to SVC, compared to limma-SVC or SVC. More importantly, computation time is reduced not at the cost of performance.

Ensemble methods like RF or GBM are more resistant, but can be costly when there are too many features.

Wrapper method might be preferrable if goal is predictive performance and computation power is not an issue, e.g., RFE, where a sequential search is conducted for the optimal combination of features. Embedded methods such as l1-penalized SVM can also be considered.

Reference

1. Machado, G. et al. (2015). “What variables are important in predicting bovine viral diarrhea virus? A random forest approach” Veterinary Research [https://veterinaryresearch.biomedcentral.com/articles/10.1186/s13567-015-0219-7]
2. Grissa, D. et al. (2016). “Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data.” Frontiers in Molecular Biosciences [2. https://www.frontiersin.org/articles/10.3389/fmolb.2016.00030/full]
3. Ghaffari, M., et al. (2019) “Metabolomics meets machine learning: Longitudinal metabolite profiling in serum of normal versus overconditioned cows and pathway analysis” Journal of Dairy Science. [https://www.sciencedirect.com/science/article/pii/S0022030219308380]
4. Atla, A., et al. (2011). “Sensitivity of different machine learning algorithms to noise”. CCSC: Mid-South Conference [https://dl.acm.org/doi/pdf/10.5555/1961574.1961594]