Predictive Modeling for Metabolomics Data

Illustrative Example

Alex Sanchez-Pla
Genetics, Microbiology & Statistics Department
Universitat de Barcelona
2024-06-14

Outline

An Illustrative Example

Data (1)

- LC-MS metabolomics dataset from www.metabolomicsworkbench.org (Project ID: PR00038)
- Plasma from 131 subjects was collected from the Chronic Obstructive Pulmonary Disease Gene study (COPDGene) study cohort and analyzed using untargeted LC-MS (C18+ and HILIC+) metabolomics.
- Data were annotated, normalized and preprocessed using the methods described in:
  - Cruickshank-Quinn CI, Jacobson S, Hughes G, Powell RL, Petrache I, Kechris K, Bowler R, Reisdorph N (2018) Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD. Sci Rep 8(1):17132
  - Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD (2010) Genetic epidemiology of COPD (COPDGene) study design. COPD 7(1):32–43. https://doi.org/10.3109/15412550903499522

Data (2)

- COPD is an extremely heterogeneous disease comprising **multiple phenotypes**.
- The **131 subjects** were either current or former smokers with various chronic obstructive pulmonary disease (COPD) phenotypes including airflow obstruction, radiologic emphysema, and exacerbations.
- Within this set there were **56 males and 75 females**.
- **2999 metabolites**

Training and test sets

- **70% training** (93 samples)

  Fivefold CV: 5 different subsets (or fivefolds)

  - 4 fold for training
  - 1 fold as holdout-test dataset

  The algorithms were trained against each of the folds.
  The metrics were computed (average over fivefolds) for the training dataset.

- **30% test** (38 samples)

  The test dataset was used to provide an **unbiased evaluation** of the best model fit on the training dataset.

  For **model validation**, the performance of the test data was predicted using the trained models for three classifiers.

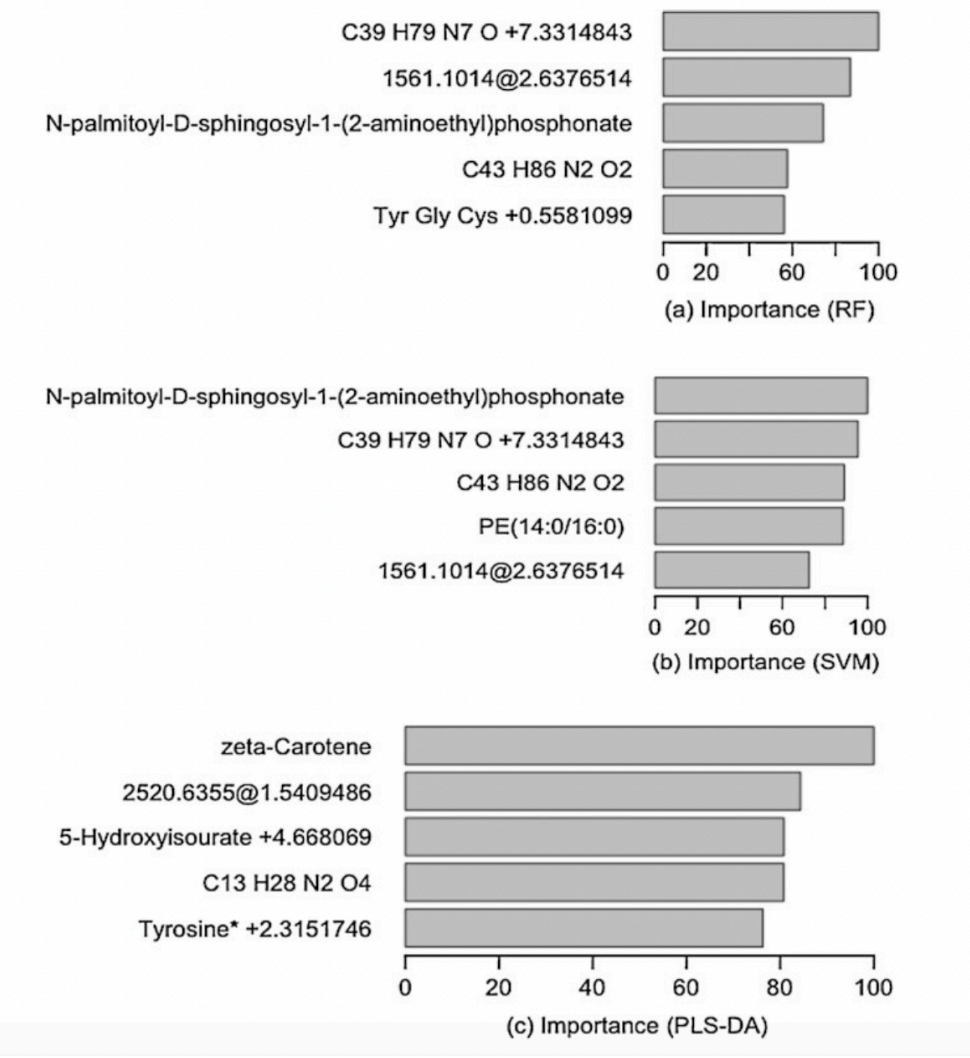
Implementing the predictive models

- Different predictive models were implemented based on the training dataset using:
    - metabolite abundances as the **predictor variables**
    - Gender (Male/Female) as the **response**
- Then, the **Variable Importance Score**, which is a measure of feature relevance to gender for each metabolite was computed.

Feature Ranking and Variable Importance

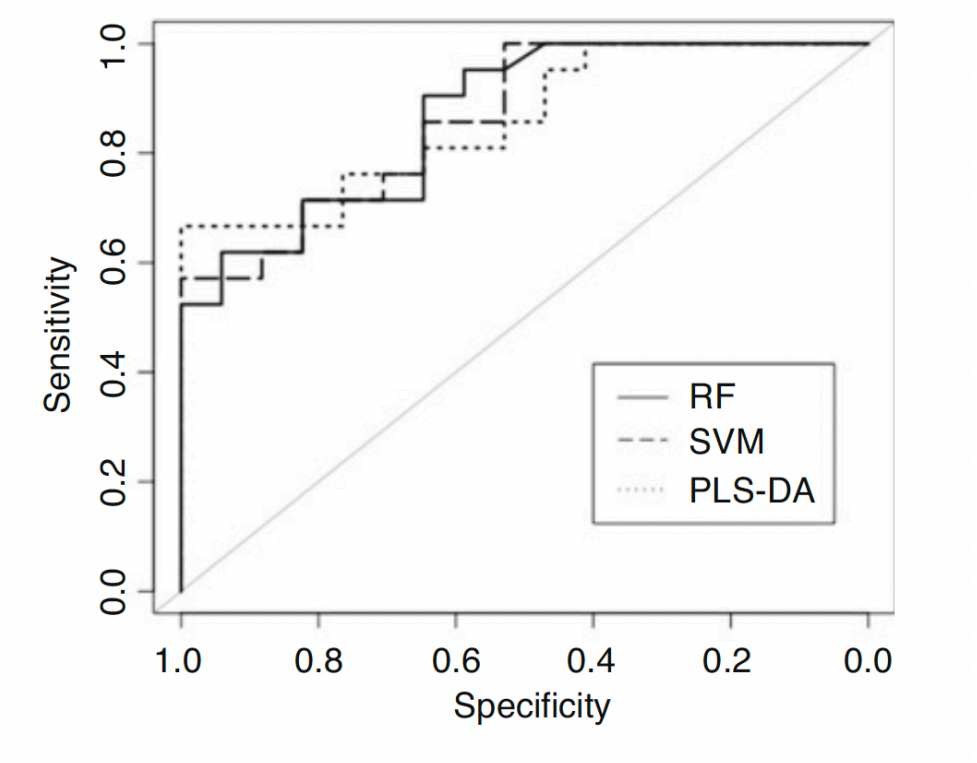
Metabolite relevant feature ranking bar plots (top five metabolites) using **Variable Important Scores** ranging from 0 to 100.

(a) Random Forest
(b) Support Vector Machine (SVM)
(c) Partial Least Square-Discriminant Analysis (PLS-DA) for the training dataset



Model Validation (1)

ROC curves of the testing dataset obtained from three classification algorithms (RF, SVM, and PLS-DA)



Model Validation (2)

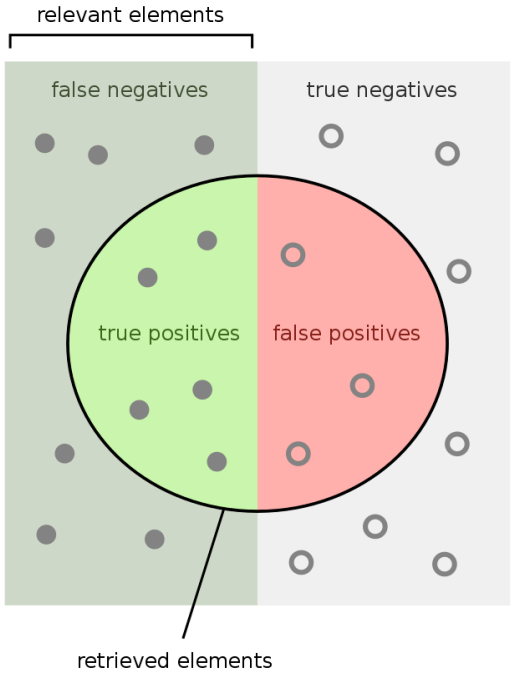
Metrics to evaluate the performance of classification on testing dataset:

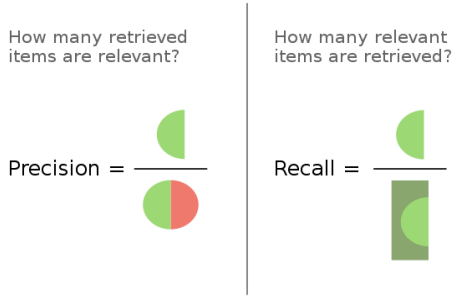
$$sensitivity = \frac{TP}{P}$$

$$specificity = \frac{TN}{N}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$





Model Validation (3)

Metrics to evaluate the performance of classification on testing dataset:

- area under curve (AUC)
- sensitivity (SENS)
- specificity (SPEC)
- precision (PREC)
- recall (REC))

Metrics/methods	AUC	SENS	SPEC	PREC	REC
RF	0.87	0.71	0.64	0.71	0.71
SVM	0.86	0.76	0.71	0.76	0.76
PLS-DA	0.86	0.81	0.65	0.74	0.81

References and Resources

Resources

- Predictive Modeling for MetabolomicsData Tusharkanti Ghosh, Weiming Zhang, Debashis Ghosh, Katerina Kechris
- Metabolomics datasets: www.metabolomicsworkbench.org
- R code