Bahaar analysis

Bahaar Ahuja

2025-11-05

Q4 - Use any method to find either:

a simpler panel that achieves comparable classification accuracy

an alternative panel that achieves improved classification accuracy

Benchmark your results against the in-class analysis.

Goal: Explore alternative feature selection

```
# Load libraries
suppressPackageStartupMessages(library(tidyverse))
suppressPackageStartupMessages(library(caret))
suppressPackageStartupMessages(library(glmnet))
suppressPackageStartupMessages(library(pROC))
suppressPackageStartupMessages(library(yardstick))
```

```
# 1. Load processed data

load("~/module-1-biomarker-data-table13/data/biomarker-clean.RData")

# Rename
biomarker <- biomarker_clean

# Convert group to factor (TD = control group)
biomarker$group <- factor(biomarker$group, levels = c("TD", "ASD"))
```

```
# 2. Train/Test Split (80% train, 20% test)

set.seed(123)
train_idx <- createDataPartition(biomarker$group,
                                 p = 0.8, list = FALSE)

train <- biomarker[train_idx, ]
test <- biomarker[-train_idx, ]

# Prepare matrices for glmnet
x_train <- as.matrix(train %>% select(-group, -ados))
y_train <- train$group
```

```r
x_test <- as.matrix(test %>% select(-group, -ados))
y_test <- test$group
```

```r
# 3. LASSO Classification and Feature Selection

set.seed(123)
lasso_fit <- cv.glmnet(
  x_train, y_train,
  alpha = 1, # LASSO penalty
  family = "binomial"
)

# Extract non-zero coefficients (The Alternative Panel)
lasso_coef <- coef(lasso_fit, s = "lambda.min")
lasso_features <- rownames(lasso_coef)[lasso_coef[, 1] != 0]
lasso_features <- lasso_features[lasso_features != "(Intercept)"]

cat("\nQ4 Alternative Panel (Pure LASSO features):\n")
```

```
##
## Q4 Alternative Panel (Pure LASSO features):
```

```r
print(lasso_features)
```

```
##  [1] "IgA"                          "CD59"
##  [3] "FAM3D"                        "FSTL1"
##  [5] "CXCL16, soluble"              "Macrophage mannose receptor"
##  [7] "P-Cadherin"                   "Protein S"
##  [9] "IGFBP-1"                      "LAG-1"
## [11] "Kallikrein 11"                "Met"
## [13] "Growth hormone receptor"      "ESAM"
## [15] "Siglec-3"                     "FCN1"
## [17] "HGFA"                         "a2-Macroglobulin"
## [19] "MAPK2"                        "IL-6 sRa"
## [21] "ENPP7"                        "ENTP5"
## [23] "MFGM"                         "PCSK7"
## [25] "PERL"                         "ITI heavy chain H4"
## [27] "Calcineurin"                  "IgD"
## [29] "DERM"                         "hnRNP K"
## [31] "ILT-4"                        "RELT"
## [33] "SIG14"                        "TWEAKR"
## [35] "PPID"                         "PSMA"
## [37] "SRCN1"                        "NRP1"
## [39] "Epo"                          "GDNF"
## [41] "14-3-3 protein zeta/delta"    "ANK2"
## [43] "a-Synuclein"                  "CSRP3"
```

```r
# 4. Evaluation of LASSO Classifier on Test Set (The Answer)

# Predict probabilities on the held-out test set
pred_prob_lasso <- predict(lasso_fit, newx = x_test,
                           s = "lambda.min", type = "response")
```

```
pred_class_lasso <- factor(ifelse(pred_prob_lasso > 0.5,
                                  "ASD", "TD"), levels = c("TD", "ASD"))

# Calculate AUROC
roc_lasso <- roc(y_test, as.numeric(pred_prob_lasso),
                 levels = c("TD", "ASD"))
```

```
## Setting direction: controls < cases
```

```
auroc_final <- auc(roc_lasso)

cat("\nAUROC (LASSO classifier) on Test Set: ",
    auroc_final, "\n")
```

```
##
## AUROC (LASSO classifier) on Test Set:  0.8577778
```

```
cat("Test Set Accuracy (LASSO classifier): ",
    confusionMatrix(pred_class_lasso, y_test,
                    positive = "ASD")$overall['Accuracy'], "\n")
```

```
## Test Set Accuracy (LASSO classifier):  0.7666667
```

Q4: Use any method to find either a simpler panel or an alternative panel.

We chose to find an alternative panel that achieves improved classification accuracy by utilizing the pure LASSO (Least Absolute Shrinkage and Selection Operator) penalized regression model.

The original analysis used the intersection of three methods (t-test, Random Forest, and LASSO) followed by an unregularized Logistic Regression. For this alternative approach, we used the LASSO method alone to perform both feature selection and classification. The LASSO model was trained on the 80% training partition, and the optimal penalty ($\lambda_{\min}$) automatically selected the feature panel.

Results:

1. Alternative Panel: The LASSO model selected a panel of 43 proteins (listed in the code output).
2. Test Set Performance: The LASSO classifier achieved a Test Set AUROC (Area Under the Receiver Operating Characteristic curve) of 0.8578 (Accuracy: ≈ 0.767).

```
comparison_data <- data.frame(
  Metric = c("AUROC"),
  `In-Class Analysis Benchmark (Approx.)` = c(0.825),
  `Q4 Alternative Panel (Pure LASSO)` = c(0.8578),
  Conclusion = c("Improved")
)

library(knitr)

kable(comparison_data,
      caption = "Benchmark of Classification Accuracy for Alternative Panel",
      digits = 4,
      col.names = c("Metric", "In-Class Benchmark (Approx.)",
                    "Q4 Pure LASSO Panel", "Conclusion"),
      align = 'lccc')
```

Table 1: Benchmark of Classification Accuracy for Alternative Panel

Metric	In-Class Benchmark (Approx.)	Q4 Pure LASSO Panel	Conclusion
AUROC	0.825	0.8578	Improved

The pure LASSO approach resulted in an Alternative Panel that achieved improved classification accuracy compared to the in-class intersection method. By using the penalized LASSO model for the final classification, we avoided the multicollinearity and numerical instability that would break an unregularized Logistic Regression model when handling this large, correlated set of 43 features. The 0.8578 AUROC is the highest accuracy achieved in our experiments.