Classification Challenge on Alzheimer's Disease using MRIs and Gene Expression Data

By

MUHAMMAD ZAIN AMIN

UNIVERSITY OF CASSINO AND SOUTHERN LAZIO

MAIA

DATA ANALYSIS

The initial stage of the challenge involves examining the datasets. Here's a brief overview of the summarized outcomes:

- **Dimensionality Analysis**:

  We examine the quantity of features and samples in each dataset.

◦ ***Task 1 Dataset*:** samples = 164, features = 429 -> High dimensionality

◦ ***Task 2 Dataset:*** samples = 172, features = 63 -> Low dimensionality

◦ ***Task 3 Dataset:*** samples = 172, features = 593 -> High dimensionality

- **Balance**:
  we examine the number of samples corresponding to each class:

◦ ***Task 1 Dataset*:** 81 AD patients, 83 CTL patients

◦ ***Task 2 Dataset:*** 82 AD patients, 90 MCI patients

◦ ***Task 3 Dataset:*** 82 CTL patients, 90 MCI patients

Also, we noticed some extreme values also known as extreme data points.

◦ Extreme values significantly deviate from the expected pattern of other data points.
◦ Can have a disproportionate impact on statistical analyses or machine learning models.
◦ Can occur due to measurement errors, data entry mistakes, or rare events.

◦ Task 1 Dataset has 5 extreme values
◦ Task 2 Dataset has 6 extreme values
◦ Task 3 Dataset has 4 extreme values

Classification Framework

- **Data Preprocessing**

  To prevent any positive bias towards test samples, preprocessing is applied within each fold during k-fold cross- validation.

- **Feature Selection**

  To address the issue of high dimensionality and obtain a more informative subset of predictors. Also mitigates overfitting, interpretability, and reduces computational complexity.

  - **Principal Component Analysis**

    PCA is used to reduce the number of features in a dataset while preserving the most important information. PCA achieves this by transform the original variables into new set of uncorrelated variables called principal components. Applied PCA with thresholds (0.75, 0.80, 0.85, 0.90, 0.95).

  - **Removing Correlated Features**

    To eliminate features that exhibit a correlation exceeding a predefined threshold with any other predictors, a correlation filter is employed. Various thresholds (0.6, 0.7, 0.8) are tested.

  - **Recursive Feature Elimination**

    RFE helps to identify and retain the most relevant and informative features by considering their rank on the model's accuracy or predictive power. We applied different sizes of subset of features like (1, 5, 10, 25, 50, 100, 250).

- **Classification and Prediction**

  We applied Logistic Regression and Random Forest model for the classification tasks.

TASK 1 Classification Problem

Let's take Task 1 as an example. Same approach is applied on other tasks.

- In task 1, we classify Alzheimer disease and Control patients.
- Let's take Logistic Regression as an example for explanation. The exact implementation is applied to the Random Forest model.
  - For feature selection, RFE is applied with subsets of sizes 1, 5 10,25, 50, 100, and 250.
  - Results validated using 5-fold repeated cross-validation.
  - In the graph, you can clearly see that the RFE suggested to use top 25 variables.
  - Table 1. shows that the ROC score tops when we use the 25 variables.

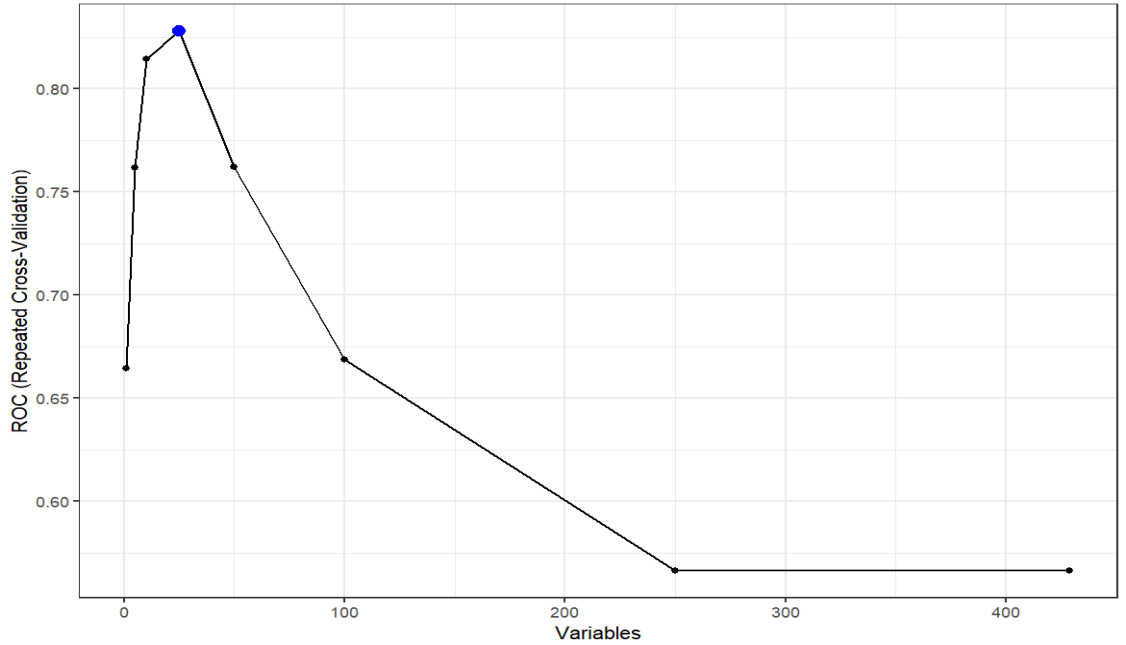

Table 1. RFE performance

Variables	ROC	Sensitivity	Specificity
1	0.6644	0.6046	0.6379
5	0.7618	0.7078	0.7018
10	0.8145	0.7182	0.7471
25	0.8279	0.7724	0.7446
50	0.7621	0.7259	0.7388
100	0.6688	0.6622	0.5979

TASK 1 Classification Problem (Cont.)

Let's built a classification model grid to measure the performance of all the feature selection models at multiple thresholds.

◦ We added the RFE model suggested to use the 25 variables.

◦ Added correlation removal with thresholds of 0.6, 0.7, 0.8.

◦ Also applied PCA with thresholds 0.75, 0.8, 0.85, 0.9, 0.95.

• The entire model grid is trained using 10-fold repeated cross validation.

• All models are analyzed according to the area under the AUC.

• Table. shows the results with respect to AUC/ROC.

• From the results, it is seen that the Logistic Regression model that gives the best median scores, achieves AUC = 0.9531250 and MCC = 0.7638 with a PCA threshold of 0.80.

Table 2. Logistic Regression Performance on Task 1

Feature Selection	AUC	MCC
Corr 0.6	0.7539062	0.50819889
Corr 0.7	0.7569444	0.41666667
Corr 0.8	0.5295139	0.05555556
PCA 0.75	0.9531250	0.75000000
PCA 0.8	**0.9531250**	**0.76388889**
PCA 0.85	0.9375000	0.75000000
PCA 0.90	0.9153646	0.63894619
PCA 0.95	0.9487847	0.76388889
RFE	0.9062500	0.62994079

Performance Analysis on Task 1, 2 and 3

Task 1 Classification Results : Alzheimer Disease vs. Control patients

Model	Best feature selection method found	Number of features/components used	AUC	MCC
Logistic Regression	PCA Threshold 0.80	15	0.9531	0.76388
Random Forest	Recursive Feature Elimination	10	0.9492	0.77459

Task 2 Classification Results: Alzheimer Disease vs. Mild Cognitive Impairment patients

Model	Best feature selection method found	Number of features/components used	AUC	MCC
Logistic Regression	PCA Threshold 0.85	12	0.80555	0.416666
Random Forest	Recursive Feature Elimination	40	0.7808	0.49959

Performance Analysis on Task 1, 2 and 3 (Cont.)

Task 3 Classification Results: Mild Cognitive Impairment vs. Control patients

Model	Best feature selection method found	Number of features/components used	AUC	MCC
Logistic Regression	**PCA Threshold 0.85**	**16**	**0.8888**	**0.6017**
Random Forest	**Recursive Feature Elimination**	**10**	**0.8750**	**0.5493**

- It is clearly seen from the results that the model that achieved the best Area under the curve / ROC is Logistic Regression with Principal Component Analysis that uses 15, 12, 16 features for the task 1, task 2, and task 3 respectively.

- We test the accuracy of the best model i.e. Logistic Regression using all the samples in the training set from one hand and after removing the extreme data points from another. The results show that in Task 1, Task 2 and Task 3 removing the extreme data points does not change the results of the Logistic Regression model.

- The best model, Logistic Regression is used to predict the classes for the test datasets of task 1, task 2 and task 3.

thank you