Prediction of schizophrenia based on Gray Matter

Yanis Rehoune, Lucie Gabagnou
February 1, 2023

Abstract

In this study, Machine Learning techniques were used to classify schizophrenia patients from neuroimaging data, including Voxel-based morphometry (VBM) and Region of interest (ROI) data.

1 Data Exploration and Processing

Data Exploration and Processing Initially, an exploratory analysis of the data was performed to understand the data structure for each of the ROI and VBM databases. Due to the large number of variables and the complexity of their relationships, one of the first approaches has been to reduce dimensionality. A principal component analysis (PCA) was performed on both databases after standardizing the variables to evaluate the variance explained by the principal components. On the ROI database, the first 8 components explained 70 % of the data variance, suggesting that the variables were relatively correlated with each other and followed a linear trend. However, on the VBM database, the PCA was not conclusive, with the first 10 axes explaining only 20% of the data variance. Manifold learning algorithms such as T-SNE, Isomap, MDS, and LEE were tested to reduce these variables, but these approaches were not satisfactory. The complexity of the VBM structure suggests that it would be better suited for deep learning techniques, however, this approach requires a significant amount of data to train the neural networks and make accurate predictions. By focusing only on the ROI database, the modeling process is simplified and can be performed with the available data.

2 Modeling

Cross-validations were used to evaluate the robustness and generalization ability of each tested model. Initially, Principal Component Analysis (PCA) was employed on standardized data to further reduce the dimensionality, and the number of components was selected through a GridSearch. Simple and largely linear classification models, such as logistic regression with and without regularization, were then applied. As a secondary step, to address the high dimensionality of the ROI data, Dimensionality reduction was then performed without the use of PCA by incorporating regularizations (of L2 type). First, Support Vector Machines (SVM) were considered as an appropriate choice for the analysis. As it is commonly used method in Multi-Voxel Pattern Analysis (MVPA), SVM allows SVM allows analyzing patterns in ROIs and discriminating patterns that are spread across the entire brain. Secondly, ensemblist methods like XGboost, and Random Forest were also tested on ROI. These models also account for potential non-linear effects in the data, and are well known for their robustness and they ability to learn from non linear data. An alternative method was tested by selecting variables (by feature importance from Random Forest) and then applying a second XGBoost model. Finally, the model with the best performance turned out to be a logistic regression on principal components with L2 regularisation. The model showed promising results with an AUC of 0.82 and a balanced accuracy of 0.52 on a public test dataset. However, if we were interested in other elements than the metric, such as overfitting, it is likely that the second set of methods would perform better.

```
CV fold 0
        score   auc  bacc      time
        train  0.83  0.55  4.884957
        valid  0.85  0.57  0.003631
        test   0.80  0.51  0.003476
CV fold 1
        score   auc  bacc      time
        train  0.86  0.53  2.290457
        valid  0.77  0.43  0.002677
        test   0.80  0.45  0.002242
CV fold 2
        score   auc  bacc      time
        train  0.84  0.53  2.256710
        valid  0.84  0.54  0.001780
        test   0.81  0.49  0.000696
CV fold 3
        score   auc  bacc      time
        train  0.84  0.53  2.312180
        valid  0.79  0.52  0.001868
        test   0.82  0.45  0.000751
CV fold 4
        score   auc  bacc      time
        train  0.84  0.50  2.291331
        valid  0.86  0.49  0.001789
        test   0.81  0.43  0.000725
----------------------------
Mean CV scores
----------------------------
        score          auc          bacc         time
        train  0.84 ± 0.008  0.53 ± 0.014  2.8 ± 1.04
        valid  0.82 ± 0.034  0.51 ± 0.047  0.0 ± 0.0
        test   0.81 ± 0.004  0.47 ± 0.028  0.0 ± 0.0
----------------------------
Bagged scores
----------------------------
        score   auc  bacc
        valid  0.82  0.51
        test   0.81  0.45
```

Figure 1: *CV results on public dataset for selected model (L2 Logistic Regression with C = 0.1)*