Schizophrenia Classification

April 20, 2018

1 Introduction

Schizophrenia is a severe and disabling mental illnesses which has no well-established, non-invasive diagnosis biomarker. Currently, due to its symptom overlap with other mental illnesses (like bipolar disorder) it can only be diagnosed subjectively, by process of elimination.

In this project you should automatically diagnose subjects with schizophrenia based on multimodal features derived from their brain magnetic resonance imaging (MRI) scans.

The features used in this project are a result from current state-of-the art developments in neuroimaging and MRI data processing. Two modalities of MRI scans are used to obtain these features: functional and structural MRI. One challenge in this competition is how to optimally combine this type of multimodal information and select features that enhance diagnosis. Optional additional information is provided that could be helpful with this particular aspect of the task.

The data used for this task had a peculiar property of having more features than available samples. Such situations are often called "High Dimensional Small Sample Size Data" (HDLSS) [4] in the literature and present a lot of challenges in both model selection and error estimation [8].

2 Data Description

2.1 Train Data

- *train_labels.csv* - Labels for the training set. The labels are indicated in the "Class" column. 0 = 'Healthy Control', 1 = 'Schizophrenic Patient'

- *train_FNC.csv* - FNC features for the training set. These are correlation values. They describe the connection level between pairs of brain maps over time.

- *train_SBM.csv* - SBM features for the training set. These are standardized weights. They describe the expression level of ICA brain maps derived from gray-matter concentration.

The test data should not be used for the project and are provided only if you want to submit your results to Kaggle competition.

2.2 Additional Files

In order to enable more principled multimodal feature selection and combination strategies, we provide files that contain the actual brain maps to which the FNC and SBM features refer to. MATLAB and R support functions have also been provided to help loading and handling the additional data. Below is a description of each additional, non-essential file:

- *rs_fMRI_ica_maps.pdf* - Contains cross-sectional images of the 28 ICA brain maps of fMRI to which the FNC features correspond to.

- *gm_sMRI_ica_maps.pdf* - Contains cross-sectional images of the 32 ICA brain maps of sMRI to which the SBM features correspond to.

- *rs_fMRI_ica_maps.nii* - Contains the 28 ICA brain maps of fMRI to which the FNC features correspond to. File is in NIfTI format.

- *gm_sMRI_ica_maps.nii* - Contains the 32 ICA brain maps of sMRI to which the SBM features correspond to. File is in NIfTI format.

- *aal_labels.nii* - Contains the the standard AAL atlas number labels in 3D space. File is in NIfTI format. The labels are described in this reference [10]: http://www.ncbi.nlm.nih.gov/pubmed/11771995

- *aal_labels_naming.m* - MATLAB script with the text labels corresponding to the number labels in the AAL atlas. (see aal_labels.nii description)

- *comp_ind_fMRI.csv* - The fMRI ICA brain map numbers with respect to this reference [1]: http://www.ncbi.nlm.nih.gov/pubmed/21442040

- *comp_ind_sMRI.csv* - The sMRI ICA brain map numbers with respect to this reference [9]: http://www.ncbi.nlm.nih.gov/pubmed/22470337
- *rs_fMRI_FNC_mapping.csv* - The fMRI ICA brain map numbers corresponding to each FNC feature provided. (see comp_ind_fMRI.csv description)
- *load_maps.m* - A MATLAB script showing how to load and display the functional (fMRI) and structural (sMRI) ICA maps. Also includes an example of how to compute the correlation between the loaded fMRI and sMRI maps.
- *show_maps.m* - A MATLAB function for use with the load_maps.m script.
- *load_AAL.m* - A MATLAB script showing how to load and display the provided AAL atlas.
- *load_maps.R* - An R script showing how to load and display the functional (fMRI) and structural (sMRI) ICA maps. Also includes an example of how to compute the correlation between the loaded fMRI and sMRI maps.

3 Helper Code

- *load_features.m* - A MATLAB script showing how to load the data from the .csv files and also how to create a new submission using the provided example.
- *load_features.R* - An R script showing how to load the data from the .csv files and also how to create a new submission using the provided example.

Note that from the above helper codes you only need the part that loads the data.

Download all project files in: `http://www.telecom.tuc.gr/patreco/res/projects/08_Schizophrenia/`

4 Features

4.1 FNC Features

Functional Network Connectivity (FNC) are correlation values that summarize the overall connection between independent brain maps over time [2]. Therefore, the FNC feature gives a picture of the connectivity pattern over time between independent networks (or brain maps). The provided FNC information was obtained from functional magnetic resonance imaging (fMRI) from a set of schizophrenic patients and healthy controls at rest, using group independent component analysis (GICA). The GICA decomposition of the fMRI data resulted in a set of brain maps, and corresponding timecourses. These timecourses indicated the activity level of the corresponding brain map at each point in time. The FNC feature are the correlations between these timecourses. In a way, FNC indicates a subject's overall level of 'synchronicity' between brain areas. Because this information is derived from functional MRI scans, FNCs are considered a functional modality feature (i.e., they describe patterns of the brain function). More about FNCs can be found in [2].

4.2 SBM Features

Source-Based Morphometry (SBM) loadings correspond to the weights of brain maps obtained from the application of independent component analysis (ICA) on the gray-matter concentration maps of all subjects. Gray-matter corresponds to the outer-sheet of the brain; it is the brain region in which much of the brain signal processing actually occurs. In a way, the concentration of gray-matter is indicative of the "computational power" available in a certain region of the brain. Processing gray-matter concentration maps with ICA yields independent brain maps whose expression levels (i.e., loadings) vary across subjects. Simply put, a near-zero loading for a given ICA-derived brain map indicates that the brain regions outlined in that map are lowly present in the subject (i.e., the gray-matter concentration in those regions are very low in that subject). Because this information is derived from structural MRI scans, SBM loadings are considered a structural modality feature (i.e., they describe patterns of the brain structure). More about SBM loadings can be found in [9].

4.3 Feature Selection

You may do feature selection[1] in order to reduce the number of features. However, this is only an optional step and should be evaluated whether it improves the performance of the classifiers.

5 Classifiers

People that worked with these data suggest the following classifiers.

- Distance Weighted Discrimination [6, 5].
- Gaussian Processes [7].
- Support Vector Machines [3].

Another good choice is the Random Forest classifier, which exists in scikit-learn. You may also implement simple classifiers together with feature selection.

6 Project deliverables

You must provide your own implementation of every method and algorithm of the project apart from advanced classifiers such as the random forests which you may use code from matlab or python libraries. You must also write a report where you provide all your results (plots, tables, etc.) and comments on the project questions. Include an introduction where you explain the mathematical background of your implementations. Your report should provide sufficient details that are needed to future readers in case they want to repeat your experiments. Provide explanations about your empirical design decisions (if any), the insights you learn in intermediate analysis steps, as well as your final results.

References

[1] Elena Allen, Erik Erhardt, Eswar Damaraju, William Gruner, Judith Segall, Rogers Silva, Martin Havlicek, Srinivas Rachakonda, Jill Fries,

[1] https://en.wikipedia.org/wiki/Feature_selection

Ravi Kalyanam, Andrew Michael, Arvind Caprihan, Jessica Turner, Tom Eichele, Steven Adelsheim, Angela Bryan, Juan Bustillo, Vincent Clark, Sarah Feldstein Ewing, Francesca Filbey, Corey Ford, Kent Hutchison, Rex Jung, Kent Kiehl, Piyadasa Kodituwakku, Yuko Komesu, Andrew Mayer, Godfrey Pearlson, John Phillips, Joseph Sadek, Michael Stevens, Ursina Teuscher, Robert Thoma, and Vince Calhoun. A baseline for the multivariate comparison of resting-state networks. *Frontiers in Systems Neuroscience*, 5:2, 2011.

[2] Elena A. Allen, Eswar Damaraju, Sergey M. Plis, Erik B. Erhardt, Tom Eichele, and Vince D. Calhoun. Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex*, 24(3):663–676, 2014.

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[4] Peter Hall, J. S. Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B*, 67(3):427–444, 2005.

[5] J. S. Marron. Distance-weighted discrimination. *WIREs Comput. Stat.*, 7(2):109–114, March 2015.

[6] J. S Marron, Michael J Todd, and Jeongyoun Ahn. Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271, 2007.

[7] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[8] Sarunas J. Raudys and Anil K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(3):252–264, 1991.

[9] Judith Segall, Elena Allen, Rex Jung, Erik Erhardt, Sunil Arja, Kent Kiehl, and Vince Calhoun. Correspondence between structure and function in the human brain at rest. *Frontiers in Neuroinformatics*, 6:10, 2012.

[10] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 15(1):273 – 289, 2002.