ATLAS BASED SPARSE LOGISTIC REGRESSION FOR ALZHEIMER'S DISEASE CLASSIFICATION

Helena Barros and Margarida Silveira for the Alzheimer's Disease Neuroimaging Initiative

Institute for Systems and Robotics, Instituto Superior Técnico
Universidade de Lisboa, Lisbon, Portugal

ABSTRACT

Sparse methods are an effective way to alleviate the curse of dimensionality in neuroimaging applications. By imposing sparsity inducing regularization terms these methods are able to perform feature selection jointly with classification.

They have been used for Alzheimer's Disease (AD) and Mild cognitive impairment (MCI) classification using different approaches such as Lasso, Group Lasso and tree-structured Group Lasso. The Group Lasso approaches have relied mainly on grouping contiguous voxels, either spatially or temporally.

In this paper we propose two grouping approaches where feature groups are more disease related. We propose that features are grouped according to anatomically defined regions of the brain, as provided by a labeled atlas, and in a hierarchy that joins corresponding regions in the left and right hemispheres, so as to take into account the bilateral symmetry which typically occurs in AD.

We apply our methods to MRI images from the ADNI and compare their performance with that of other sparse methods developed for AD. Evaluation includes classification performance and the stability of the obtained feature weights when several runs of these algorithms are performed. The proposed methods attained better or equal performance but generated more stable feature weights.

Index Terms— sparse methods, logistic regression, anatomical atlas, Alzheimer's disease, Mild cognitive impairment

I. INTRODUCTION

Alzheimer's disease (AD) is a degenerative brain disorder and the most common type of dementia. It affects primarily elderly people and its incidence is expected to increase due to aging of the population. Mild Cognitive Impairment is a preclinical stage of AD, particularly challenging to diagnose and very important for the development of treatments that can delay the progression of AD.

Neuroimaging, such as Magnetic Resonance Imaging (MRI) and Position Emission Tomography (PET), has been recognized as a powerful biomarker for AD diagnosis, and also for identifying Mild Cognitive Impairment (MCI) patients. Many machine learning techniques are currently being developed to further increase the power of neuroimaging as a diagnostic tool.

One of the difficulties these methods are faced with is the high dimensionality of the brain and the comparatively small number of images which are available for training. This is a problem termed curse of dimensionality which is known to degrade performance. To reduce dimensionality, some methods partition the brain into regions of interest (ROI) and extract regional features such as volume or mean intensity. Other methods resort to feature selection techniques, usually filter or wrapper methods, which identify the most discriminative voxels and discard the others. An alternative approach is the use of sparse methods for feature selection such as the Lasso (Least Absolute Shrinkage and Selection Operator). These methods use sparsity promoting priors based on the L_1 norm which results in feature weights that are either high or zero, thus they automatically select the most relevant voxels. They are fast and theoretically sound. They have been successfully used for AD classification in [1] and for predicting MCI to AD conversion in [2].

Despite their efficacy, the features selected with these methods may be sparsely distributed throughout the whole brain and unstable. To overcome this, sparse methods have been extended to take into account the feature structure. It is natural that contiguous voxels will be grouped or that voxels along time exhibit similar behaviour. This can be exploited with the Group Lasso method which extends L_1 regularization to $L_{2,1}$ regularization. The L_1 norm is applied for the different groups to promote sparsity and the L_2 norm for voxels within each group making features within a group have similar weights. This approach has been applied to AD in [3] where multi-task learning was used with different imaging modalities (MRI, FDG-PET and CSF) as the different tasks, in order to select a common feature subset relevant to all modalities. In another multi-task AD approach [4] it was used on longitudinal data by having each task correspond to a different time instant with the goal of

This work was supported by FCT project [UID/EEA/50009/2013]. Data used in the preparation of this article were obtained from the ADNI database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

978-1-5090-2809-2/17/$31.00 ©2017 IEEE

selecting features common to all the time points.

Although Group Lasso often leads to better performance than using feature selection methods which do not explore the relations among features, it requires that the groups are disjoint and known a priori. A more flexible structure can be exploited through the tree structured Lasso where a hierarchy of relationships between features can be defined. In this hierarchy leaf nodes represent individual features (voxels) and internal nodes in the tree represent groups of neighbouring features. This approach has been investigated in [5] for AD and MCI classification using cubes as the feature groups.

In this paper we propose two grouping approaches where feature groups are more disease related. In our approach neighboring features are grouped according to anatomically defined regions of the brain and in a hierarchy that joins corresponding regions in the left and right hemispheres of the brain to take into account bilateral symmetry which typically occurs in AD.

We apply these methods to MRI images from ADNI and evaluate their classification performance and the stability of the obtained feature weights when several runs are performed.

The remainder of the paper is organized as follows. First, in Section II we present the methods. Then, in Section III we describe the data used, the experiments and report the results we obtained. Finally, in section IV we present our conclusions.

II. METHODS

We tested five different sparse methods of the three types: Lasso, Group Lasso and Tree-Structured Group Lasso. The ultimate goal of all these methods is to calculate the weight associated with each feature (i.e., with each voxel). These weights will reflect the importance of a given cerebral voxel for AD or MCI classification.

In this type of methods, a set of image samples are used to train the model, $\{a_i, b_i\}_{i=1}^{m}$ where $a_i \in \Re^n$ represents the features of dimension n and $b_i \in \{-1, 1\}$ the class of each sample. Using that information, a model parameter vector x of weights can be learned by solving the optimization problem in Eq. (1), where $L(x)$ is a given loss function, $\Omega(x)$ is the regularization term (penalty) and $\lambda > 0$ is the regularization parameter that regulates the trade-off between the two terms.

$$\min_{x} f(x) = L(x) + \lambda \Omega(x) \tag{1}$$

The loss function used is the Logistic Regression function, detailed in Eq. (2) and where w_i is the weight for the i-th sample, m is the total number of samples, $x \in \Re^n$ is the vector of weights associated with each feature and c is the scalar intercept.

$$L(x) = \sum_{i=1}^{m} w_i \log(1 + e^{-b_i(x^T a_i + c)}) \tag{2}$$

The following sections detail the different regularization terms that were used and in section II-D we describe the classifier.

II-A. Lasso

Lasso method is the simplest sparse model. In addition to its sparsity-inducing property it has low complexity and guaranteed convexity. This sparse model is based on the L_1-norm penalty and therefore the regularization term is the one in Eq. (3).

$$\Omega_{Lasso}(x) = \|x\|_1 \tag{3}$$

The L_1 penalty imposes sparseness on x by shrinking its coefficients towards zero. Since many feature weights will be zero, the method automatically selects the most relevant voxels.

II-B. Group Lasso

In the Group Lasso method the penalty term is based on the $L_{2,1}$ norm, as shown in Eq. (4).

$$\Omega_{GLasso}(x) = \sum_{i=1}^{k} w_i^g \|x_{G_i}\|_2 \tag{4}$$

In this expression, $x_{G_1}, x_{G_2}, ..., x_{G_k}$ denotes the k non-overlapping groups that x is divided into, and w_i^g denotes the weight for the i-th group. The effect of $L_{2,1}$ norm is to make features in the same group have similar weights and to push entire groups to have zero weights.

Two types of Group Lasso were tested.

- Cubic Group Lasso - In this approach the groups correspond to the division of the 3D brain into 16^3 non overlapping cubes, 16 along each dimension.
- Atlas Group Lasso - In this approach the groups correspond to all the regions of the Harvard-Oxford atlas [6]. This includes 21 cortical regions and 48 subcortical regions. These groups are the ones represented in the second level of the tree shown in Fig. 1 plus the brain stem in the first level.

II-C. Tree structured Group Lasso

The tree structured Group Lasso is a special case of the overlapping Group Lasso, in which there are hierarchical relationships between groups. In this case the penalty term in Eq. (5) is applied, where d is the number of tree levels, n_i is the number of nodes for a given level, G_j^i is the group of voxels in node j of the i^{th} level of the tree and, finally, w_j^i is the weight which has been predefined for that group [5].

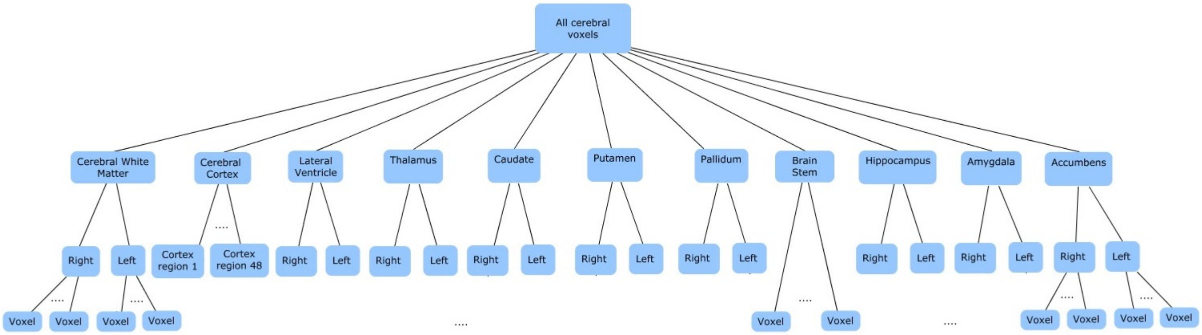

Fig. 1. Tree based on the cortical and subcortical regions of the Harvard-Oxford atlas.

$$\Omega_{Tree}(x) = \sum_{i=0}^{d} \sum_{j=1}^{n_i} w_j^i \|x_{G_j^i}\|_2 \tag{5}$$

Two types of tree structured Group Lasso were tested.

- Pyramid Tree - This tree is based on the approach used in [5] which explores hierarchical spatial relations between neighbouring voxels. The tree has 3 levels. The first level divides the brain into 4^3 cubic regions, the second level divides each region in the previous level into another 4^3 cubes (which corresponds to the groups in the Cubic Group Lasso described in II-B), and the third level is formed by single voxel leaves.
- Atlas Tree - This tree extends the Atlas Group Lasso method described in II-B by exploring hierarchical spatial relations between the brain regions, such as the union of corresponding regions in left and right brain hemispheres, which is expected as AD is typically bilateral. The tree has 3 levels: the first one groups together the cortical regions in the atlas and the subcortical regions with their counterpart in the opposing brain hemisphere (not applicable for the brain stem); the second level contains the regions in the Atlas Group Lasso and the third level contains all the leaf nodes corresponding to individual brain voxels. This tree is represented in Fig. 1.

II-D. Classification

For classification we use the Logistic Regression Classifier. This classifier computes the posterior probability of class $y \in \{-1, 1\}$ as follows:

$$\Pr(y|a_j) = \frac{1}{1 + e^{-y(x^T a_j + c)}} \tag{6}$$

where x and c are the weight vector and intercept respectively. A test sample j is classified by evaluating the probabilities of class membership in Eq. (6) and then assigning it to the more probable class.

To avoid biasing we re-estimate the model using only the voxels which were selected (weights different from zero) and assess the performance of this new model.

III. EXPERIMENTS

We tested our methods on two binary problems (AD vs. Control Normals (CN) and MCI vs. CN). The value of the regularization parameter was estimated with 5-fold nested cross-validation. The values for λ were specified as a ratio $10^{\alpha} \times \lambda_{max}$ with $\alpha \in \{-3, -2.75, -2.5, ..., 1\}$, where λ_{max} denotes the maximum value of λ above which all weights are zero. In the Group Lasso methods the weight for each group was set to the square root of the group size. The methods were implemented with the SLEP toolbox [7].

For classification evaluation we used ROC Curves and the corresponding Area under the Curve (AUC) obtained with 10-fold cross validation.

Regarding the obtained feature weights, we analyzed their distribution and their stability across the different folds. Our stability metric is based on Pearsons correlation coefficient to measure the similarity between two weight images. We evaluate the pairwise similarities over all the possible pairs and then compute the average.

III-A. Data

We used baseline 1.5T MRI data downloaded from the ADNI database (`http://www.adni-info.org/`).

The images had undergone a number of pre-processing steps by ADNI researchers to eliminate meaningless differences but they were not aligned. Therefore we warped all the images into the MNI152 standard space in the following five steps: 1) skull-stripping with FreeSurfer, 2) tissue segmentation into white-matter (WM) gray-matter (GM) and cerebrospinal fluid (CSF) with SPM12, 3) non-linear registration into a subject specific template with DARTEL toolbox from SPM12, 4) affine mapping of the template to the MNI152 atlas and 5) resampling of all the images into the MNI152 standard space using the appropriate composition of transformations. The resulting volumes were $121 \times 145 \times 121$

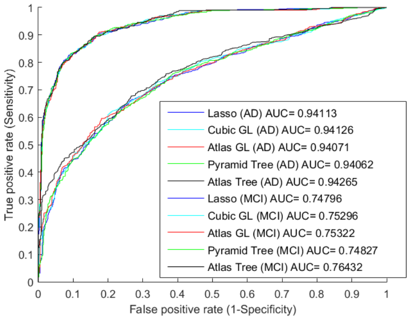

Fig. 2. ROC Curves and AUC.

but only the brain voxels were used, in a total of 486031. The GM data from these voxels were used as features.

Table 1 presents demographical and clinical information of the data used. Age of the different clinical groups does not vary significantly (pvalue > 0.05) according to the *t*-test performed between the different group pairs.

Group	CN (75)	MCI (135)	AD (58)
Age (mean±sd)	75.9±4.9	75.1±6.6	76.0±6.6
Sex (M/F)	49/26	88/47	34/24
MMSE (mean±sd)	29.1±1.0	27.2±1.6	23.5±1.9
CDR (mean±sd)	0±0	0.5±0.0	0.8 ±0.2

Table I. Demographic and clinical characteristics of each group. The number of images is shown in parentheses.

III-B. Results

Fig. 2 shows the ROC curves and corresponding AUC obtained for AD vs CN and MCI vs CN classification. As can be seen, in both cases the performance of all the methods is very close, with AUC around 0.94 for AD vs CN and worse, as expected, for MCI vs CN where AUC values are around 0.75. For MCI vs CN there is a small improvement of the Tree Atlas (0.76) over the others. The number of features (results not shown) is, on average, smaller for Lasso than for the remaining methods and smaller for MCI than for AD.

Table II shows the stability of the feature weights across folds for both classification tasks. In both cases Tree Atlas outperformed the other methods and for AD vs CN, Atlas Group Lasso was the second best. The improvements in stability of the Tree Atlas method are particularly remarkable for MCI vs CN. Despite the fact that Lasso has the lower number of features, and therefore more zero weights which may contribute to higher stability, overall the stability result of this method is among the worse.

We also looked at the regional distribution of the selected features, i.e. of the weights different from zero (results not shown). In general, most of the features are selected in the amygdala and hippocampus. The parahippocampal gyrus has shown to be one of the most important cortical regions. This is in agreement with previous studies [5]. The weights in these regions the weights are mostly symmetrically distributed, more so in the Tree Atlas methods.

Methods	AD vs CN	MCI vs CN
Lasso	0.8776	0.7078
Cubic Group Lasso	0.9070	0.7162
Atlas Group Lasso	0.9112	0.6934
Pyramid Tree	0.8842	0.7147
Atlas Tree	0.9335	0.9152

Table II. Stability of weights across folds.

IV. CONCLUSIONS

We proposed atlas based sparse Logistic Regression for Alzheimer's disease classification from MR images. We compared the proposed methods (Group atlas and Tree Group atlas) with other Lasso, Group Lasso and Tree structured Group Lasso approaches previously proposed for AD. All the methods were effective in the two classification tasks (AD vs CN and MCI vs CN) but the Atlas Tree method, which explores a hierarchy of brain regions and bilateral symmetry, generated more stable, and therefore more interpretable, feature patterns.

V. REFERENCES

[1] A. Rao, Y. Lee, A. Gass, and A. Monsch, "Classification of Alzheimer's disease from structural MRI using sparse logistic regression with optional spatial regularization," in *EMBC, 2011 Annual International Conference of the IEEE*, Aug 2011.

[2] J. Ye, M. Farnum, E. Yang, R. Verbeeck, V. Lobanov, N. Raghavan, G. Novak, A. DiBernardo, and V. Narayan, "Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data," *BMC Neurology*, vol. 12, no. 1, 2012.

[3] D. Zhang and D. Shen, "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease," *NeuroImage*, vol. 59, no. 2, 2012.

[4] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group Lasso," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2012, KDD '12, ACM.

[5] M. Liu, D. Zhang, P. Yap, and D. Shen, "Tree-guided sparse coding for brain disease classification," in *MICCAI 2012*. Springer, 2012.

[6] R. S. Desikan, F. Sgonne, Bruce Fischl, Br.T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, M. S. Albert, and R. J. Killiany, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, no. 3, 2006.

[7] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009.