Diagnosis of Alzheimer's disease using sparse logistic regression

Helena Mafalda Barros
helena.barros@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

June 2016

Abstract

The term Alzheimer's Disease (AD) is used to describe the condition of individuals with loss of cognitive functions as a result of a neurodegenerative process that is initiated around 20 years before the apperance of the first signals of decline. AD is the most frequent cause of disease and affects around 35.6 million of individuals worldwide. Nowadays, only probable AD cases are diagnosed based on neuropsychological tests, but a definite diagnosis is only possible with autopsy. In this way, new techniques of diagnosis using biomakers are under research. The main goal of this master thesis is to distinguish, AD, Mild cognitive impairment (MCI) and CN (cognitive normal) individuals, using FDG-PET images acquired every 12-months and taking advantage of spatial or temporal information. Since sparse logistic regression models allow to take into consideration that kind of information, those methods were chosen to accomplish the purpose of this thesis, more specifically Lasso, Group Lasso, Tree Group Lasso and Multi-task Group Lasso methods. From those, the Multi-task Group Lasso has shown to be the more robust and stable method, achieving an accuracy of 90.6%, 66.5% and 68.7% for AD vs CN, AD vs MCI and MCI vs CN, respectively and AUC value of 92.9%, 74.0% and 65.5% for the same samples. This method have also shown to be the one associated with higher computational costs. Finally, the most weighted features were selected in regions that doctors consider important for AD diagnosis, such as parahippocampal gyrus, hippocampus and amygdala.

Keywords: Alzheimer's Disease (AD), Mild cognitive impairement (MCI), Positron emission tomography (PET), sparse logistic regression models, Multi-task Lasso

1. Introduction

1.1. Motivation

The term dementia is applied to describe the condition of an individual with loss of some cognitive functions such as memory, orientation, ability to learn, visuospatial perception, language, capacity of planning, organizing and sequencing. Although this condition is mainly verified in people above 65 years, it is not a normal feature of ageing. Alzheimer′s Disease (AD), resultant from a neurodegenerative process, is the most common form of dementia. As time goes by, people with Alzheimer′s disease lose their independence, needing the support of their families or caregivers [10]. AD corresponds to around 65%-85% of the cases of dementia [11]. In 2010, the total number of AD patients was around 35.6 million worldwide and this number is expected to double every 20 years. There are no available treatments to stop or reverse the progression of the disease and this is why AD is such a cause of death [12]. Thus, it is desirable to develop medicines able to slow down disease progression and to prevent it before brain damage and neuronal death. However, an early and accurate diagnosis has a fundamental role to determine whose individuals need such treatments. Currently, there are no techniques available that can confirm an AD diagnosis with 100% certainty but new techniques aiming to achieve a good performance are under research [11], [12]. Recently, neuroimaging examinations have shown to provide important informations regarding biological markers that can be used as supportive features for AD diagnosis [3]. The purpose of the present research is to study different methods to extract meaningful information from neuroimaging data (PET images) to classify AD patients. According to recent studies, early AD diagnosis is more reliable and sensitive through brain image analysis than using traditional cognitive evaluation [2], [14]. Recently, many machine learning methods have been introduced and widely explored in order to automatically distinguish many types of dementia and in particular to diagnose MCI and AD patients based on their brain images [5]. The main goal of this research study is to distinguish, AD, Mild cognitive impairment (MCI) and CN (cognitive normal) individuals, using FDG-PET images

acquired every 12-months and taking advantage of spatial or temporal information. Since sparse logistic regression models allow to take into consideration that kind of information, they were chosen to accomplish the purpose of the present study. Note that those methods have been investigated in previous neuroimaging studies for selecting the relevant imaging biomarkers and have achieved very promising results on disease classification.

1.2. State-of-the-art

Concerned about the high dimensional data, around 486000 voxels per image, some researchers have been focused on studying and testing different sparse models that ensure data sparsity. The objective of achieving data sparsity is to reduce costs associated with high dimensional neuroimaging data, while rejecting noisy brain image voxels i.e., voxels of regions that are not afected by the disease.
In 2012, *Liu et al.* [7] applied Least absolute shrinkage and selection operator (Lasso), a voxel-based sparse classifier using $L1$-norm regularized linear regression model to classify AD and mild cognitive impairment (MCI) using magnetic resonance imaging (MRI) data, having a better performance than classification using support vector machine (SVM). However, although being a method effectively used to promote sparsity and to allow feature selection, the features selected may be randomly distributed throughout the whole brain since no additional information about the features is introduced and thus, the features are not selected in groups. [9]. In order to overcome that fact, new studies were performed, considering optimization problems with different regularization terms, in which spatial or temporal information is introduced in order to evaluate groups of voxels with basis on temporal or spatial information. One such study was developed by Zhang *et al.* [18] that considered the multi-task Lasso model in their multimodal study, i.e., considering images from different modalities: MRI and PET. Moreover, Zhu *et* al. proposed an approach to perform longitudinal feature selection using multi-task group lasso and multi-modal data. In this case, the training samples containing only the selected features were classified using Support Vector Regression (SVR) and Support Vector Classification (SVC). The multi-modal data was composed of MRI, PET and CSF biomarkers. In another research study meant to explore the multi-task model for AD classification, Liu *et al.* [5] introduced a novel multiple kernel learning framework to combine multi-modal features. For the purpose, they imposed a mixed L_{21} norm constraint on the kernel weights to enforce group sparsity among different feature modalities. To validate the model, MRI and CSF images were used. In the same year, Liu *et al.* in [9], introduced the Tree Lasso method as a way to group brain image voxels according to spatial information. The study was based on the assumption that neighbour voxels are spatially correlated and thus, they can be grouped according to their location. In this way, the authors divided an 3D MRI image three times into consecutive smaller squares, originating a tree having the whole voxels of the image in the same group in depth zero, cubes of dimension $16 \times 16 \times 16$ in the first division (depth 1), cubes of dimension $4 \times 4 \times 4$ in the second division (depth 2) and in the last one, cubes of dimension $1 \times 1 \times 1$, e.g., the leaf nodes (depth 3). One should notice that the process of division from one depth to the other was sequential, which means that the bigger cubes are the parents of the cubes resultant from its division, e.g., the group of depth 0 is the parent of the groups of depth 1 and the nodes represented in depth 1 are the parents of the nodes of depth 2, and so far until leaf nodes are achieved (nodes of the deepest level of the tree). Later, in 2014, Liu *et al.* [8] grouped neighbour voxels and organized the resulting groups into a tree structure, applying hierarchical agglomerative clustering to encode the rich structure of voxel-wise imaging features. At first, each voxel was treated as a singleton cluster that iteratively agglomerates a pair of neighboring clusters until all clusters have been merged into the same cluster. In this way, a binary tree was produced to represent a hierarchy of clusters with each node associated with the cluster obtained by merging its children clusters. The root of the tree gathers all the voxels and the leaf nodes are the clusters, each one consisting of a single voxel. Finally, after feature selection using the tree-guided sparse learning method, classification was done with a SVM.

2. Background

In Sparse Logistic Regression methods, a set of image samples are used to train the model, $\{a_i, bi\}_{i=1}^{n}$ where $a_i \in \Re^p$ represents the features of dimension p and $b_i \in \{-1, 1\}$ the class of each sample. Using that information, a model parameter vector x of weights can be learnt by solving the optimization problem of Eq. 1 , where $L(x)$ is a given loss function, $\Omega(x)$ the regularization term (penalty) and $\lambda > 0$ 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 depends on predicted and observed values [4]. The loss function that will be considered in this research study is the Logistic Regression function, detailed in Eq. (2) and where, in general, w_i is the weight for the i-th sample, m is the total number of samples and c is the intercept (escalar).

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

The regularization term, introduced in the model coefficients, is responsible for leveraging sparsity and prevent the coefficients to fit so perfectly that overfit. Consequently, it is the one that will circumvent the noise resultant from the high dimensionality of PET images (hundreds or thousands of features) that would affect the performance of the model and that will guarantee that only discriminative features (the ones with a weight higher than zero) are used for classification [15].
In Sections 2.1 to 2.4, a more detailed explanation about each Sparse Logistic Regression method is addressed [6].

2.1. Lasso

Lasso method is the simplest sparse model and has attracted many researchers not only because of its sparsity-inducing property, but also due to its low complexity and high theoretical agreement shown [17]. This sparse model is based on the $L1$-norm penalty and therefore the regularization term considered in 1 is shown in Eq. 3.

$$\Omega_{lasso}(x) = \|x\|_1 \quad (3)$$

2.2. Group Lasso

Group Lasso penalizes the model parameters based on the $L2,1$-norm, as described in Eq. 4. In this type of regularization, the weights attributed to the voxels in each group are subjected to the Euclidean norm and the regularization strength is controlled by the parameter w^g [17], as shown in Eq. 4. Note that $x \in \Re^{n\times 1}$ is divided into k non-overlapping pre-defined groups $x_{G_1}, x_{G_2}, ..., x_{G_k}$ and w_i^g denotes the weight for the i-th group.

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

2.3. Tree Group Lasso

In some other applications, a special case of the overlapping group Lasso, in which there are hierarchical relations between groups, e.g., a tree structure, is relevant. In those cases, the tree structured group Lasso penalty is applied (Eq. 5), where $x \in \Re^{n\times 1}$, 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 the node j of the level i of the tree and, finally, w_j^i is the weight pre-established for that group [17],[9]. In this model, nodes in the same level of the tree do not overlap while the index sets of a child node is a subset of its parents nodes. Furthermore, if a parent node is not selected, any of its children nodes can be selected. The reverse is not applicable.

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

2.4. Multi-task Lasso

Lastly, Multi-task Group Lasso is a model in which temporal information is introduced. In this model, a matrix X of weights is learnt, with three columns (one for each of the follow ups) and the same number of lines as the number of features. The matrix A, which includes all the samples, is penalized by applying the $L_{2,1}$-norm, according to Eq. 6. The use of the L_2-norm on row vectors forces the weights corresponding to the p^{th} feature across multiple time points to be grouped together and the L_1-norm selects the features according to their strength across the different times jointly [18].

$$\Omega_{multi-task}(x) = \|x\|_{l_1 \div l_2} \quad (6)$$

In that case, the optimization problem that needs to be solved to calculate brain voxels weight is the one in Eq. 7, where a_{il} denotes the i-th sample for the l-th task, $x \in \Re^{n\times k}$, w_{il} is the weight for a_{il}^T, b_{il} is the response of a_{il} and $c \in \Re^{1\times k}$.

$$\min_x \sum_{l=1}^{k} \sum_{i=1}^{m_l} w_{il} \log(1+\exp(-b_{il}(x_l^T a_{il}+c_l))) + \lambda \|x\|_{l_1 \div l_2} \quad (7)$$

3. Implementation

The goal of this research work is to perform automatic FDG-PET neuroimaging analysis through feature selection and classification. For the feature selection task, different sparse learning models are considered to calculate the weights associated with each neuroimaging voxel, such as Lasso, Group Lasso, Tree Group Lasso and Multi-task Group Lasso (already detailed in Section 2).
The classification task was performed using the Logistic Regression classifier. Prior to feature selection and classification, intensity normalization and feature extraction are required.
Those steps are detailed in Section 3.1 to 3.4.

3.1. Pre-processing

In order to avoid artifacts in regions that are known to be conservative in AD, the normalization proposed by Yakushev et al., in 2009 [16] was applied. After that, normalized features with mean zero and standard deviation one were obtained, making it possible to compare the PET images of different patients.

3.2. Feature Extraction

In the research work, the voxels of the whole brain will be used as features (voxel-based approach) and a matrix will be created containing brain voxels from different subjects, called matrix A. Moreover, longitudinal information will be considered, i.e. data from the same subjects at different points in time. In this study we used data from baseline (first visit), 12-month (visit 12 months after the first visit) and 24-month (visit 24 months after the first visit) follow-up scans of each subject, once that the CMR_{gl} decline along time is possible to be observed through the analysis of those follow-ups.

3.3. Feature Selection

In this research work, Lasso, Group Lasso, Tree Lasso and Multi-task Lasso were used to calculate the weights associated with each brain voxel and, consequently, for feature selection.
Bellow, are detailed the approaches considered for feature selection.

1. Lasso: In this approach, the samples are the FDG-PET images of different patients and from different follow-ups, e.g., follow-ups from the same patient are considered different samples as if they were from different patients and no spatial or temporal information is included. However, note that images of patients selected for training are all included in the training set and images from the ones selected for testing are all included in the testing set.

2. Group Lasso Atlas: The groups considered in Group Lasso Atlas are based on cortical and subcortical atlases. These groups are the ones represented in the first level of the tree, named Depth 1, in Fig. 2.

3. Group Lasso Big Cubes: The groups considered result from dividing the whole 3D PET image of dimension $121 \times 145 \times 121$ into cubes of dimension $31 \times 30 \times 37$, $31 \times 31 \times 37$, $30 \times 30 \times 37$, $30 \times 30 \times 36$, $30 \times 31 \times 37$, $30 \times 31 \times 36$ and $31 \times 31 \times 36$ as represented in Figure 1 in yellow, red, green, purple, pink, blue and orange, respectively.

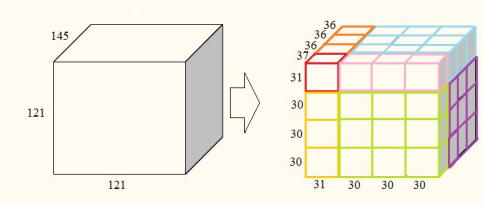

Figure 1: Groups originated based on the assumption that neighbour voxels are spatially correlated. The numbers represented correspond to the dimensions of the 3D image and, consequently, to the number of voxels. The cubes represented with the same color, have the same size.

4. Group Lasso Small Cubes: Groups that result from the division of the groups described in Group Lasso Big Cubes, i.e., groups that result from the division of the ones represented in Fig. 1. In this second division, smaller cubes are originated from the ones represented with different colors. Note that:

   - Edges with size 31 are divided in 3 edges of size 11, 10 and 10, respectively.
   - Edges with size 30 are divided in 3 edges of size 10.
   - Edges with size 37 are divided in 4 edges of size 10, 9, 9 and 9, respectively.
   - Edges with size 36 are divided in 4 edges of size 9.

5. Tree Atlas: Tree based on cortical and subsortical atlas. This tree has 4 distinct levels. The first one, from the bottom to the top, corresponds to leaf nodes and each one incorporates a single voxel. In a second level, each node contains a region of the righ subcortical atlas (except cerebral cortex), left subcortical atlas or cortical atlas. In a third level, the nodes correspond to subcortical atlas regions. Finally, the higher level consists on a single node that includes all brain voxels. This tree is reprsented in Fig. 2.

6. Tree Cubes: This tree was constructed with basis on the assumption that neighbour voxels are spatially correlated and thus, the 3D brain image is divided into smaller cubes, originating a tree with different levels where the highest level corresponds to the whole brain inside a node.

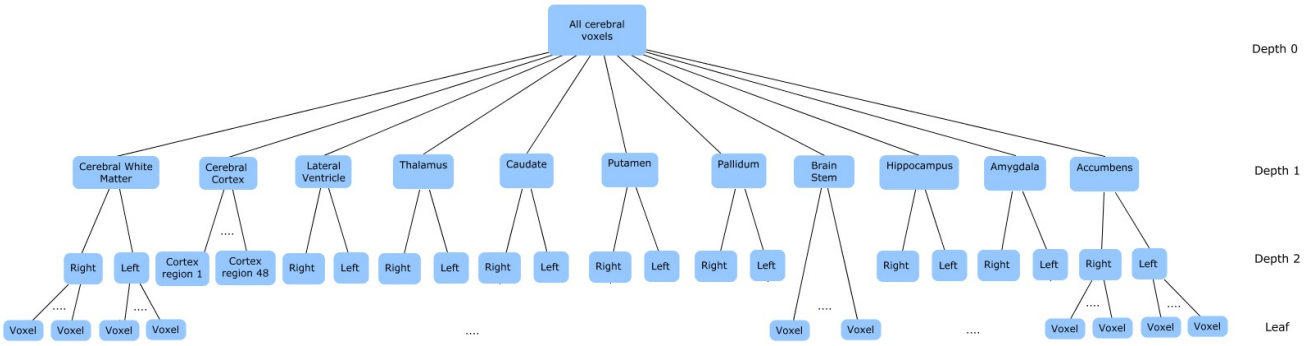

Figure 2: Tree originated with basis on cortical atlas and subcortical regions.

The following levels correspond to nodes incorporating a cube resultant from the subdivision of the whole 3D brain. This subdivision occur until the lowest level, i.e., the level formed by single voxel leaves, is achieved.
Note that this tree is created using regions obtained from cube division, as already described in Group Lasso big cubes and group lasso small cubes. However, in that case, a hierarchical relation is introduced between the big and small cubes. In this way, the cubes resulting from the first division - Big cubes - are placed in depth 1 of the tree and the smaller cubes,i.e. the ones resulting from the second division, are placed in depth 2. Finally, the voxel-based leaf nodes are placed at the bottom of the tree and in depth 0 (top of the tree) there is a region that contains all the voxels of the image, as it happened for the tree represented in Fig. 2.

7. Hierarchical Clustering: Tree that results from a hierarchical cluster analysis, where the clusters are the subcortical atlas regions. In this approach, the clusters are joined together in a hierarchical way from the closest ones, i.e., the most similar ones, to the furthest apart, i.e. the most different ones [13]. The distance between clusters is calculated with basis on the mean value of the voxels included in a given cluster, for each observation (Fig. 3).
It is not represented in the dendrogram, but there is an additional level in the bottom that considers each leaf node as an isolated voxel.
It is important to mention that the images used to construct that tree were the ones from CN patients.

8. Multi-task Lasso: Three different tasks (dimension t of Fig.4) were considered, for incorporating Baseline, 12-Month and 24-Month data.

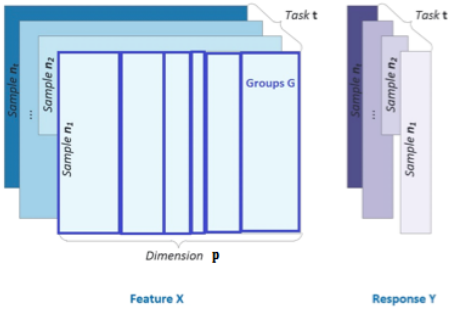

Figure 4: Data dimensions. Adapted from [6].

Note that, in Group Lasso approaches, the weight attributed to each group was equivalent to the square root of the number of features inside that group.
In it's turn, the weight attributed to each group of Tree Group Lasso approaches, was also equivalent to the square root of the number of features inside a given group (group size). However, after attibuting that weight to the groups, the algorithm normalizes the weight of all the groups of the tree between 0 and 1. In this way, the node in the root of the tree is considered with weight 0 otherwise, almost all brain features would be selected, independently of the value of λ, because the λ would not be enough to penalize such a big group. However, once the objective of that study is to select the features that better classifies AD and to understand in which regions are they situated, that situation should be avoided.

3.4. Classification

The Logistic Regression Classifier was used for classification. The probability of each class, $Pr(C|a)$,

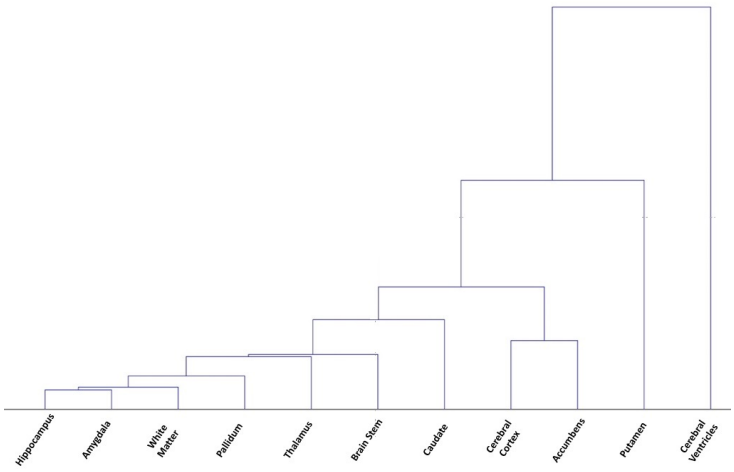

Figure 3: Tree generated from hierarchical clustering of subcortical atlas brain regions. Although this is not described in the image, there is an additional level in the bottom that corresponds to the derivation of the regions in individual voxels.

is calculated with basis on Eq. 8, where x and c are the parameters to be minimized [1].

$$\Pr(C|a) = -\sum_{i=1}^{N} log \frac{1}{1 + e^{-b_i(x^T a_i + c)}} \qquad (8)$$

Notice that elements of $b \in \{-1, 1\}$. In the present work, the label +1 was attributed to the group of subjects with a more severe condition of the two groups considered and -1 is used to designate the other group, e.g., when AD is compared to CN, +1 is assigned to the AD subjects and -1 to the CN subjects; when distinguishing AD from MCI subjects, the AD ones are labeled with +1 and the CN individuals with -1; and, finally, when differentiating AD from MCI subjects, the first ones are labeled with +1 and the others with -1. Moreover, a 10-fold cross validation was applied. In this way, the dataset was split and 90% of the samples were used for training and 10% for testing. The classification process was then repeated 10 times, one for each training set, with the main objective of avoiding any bias derived from the random partition of the dataset in cross-validation.

4. Results

All data used in this thesis was retrieved from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Moreover, all the experiments presented here, were performed using the algorithms of SLEP toolbox [6].

4.1. Participants

For the purpose of this thesis, FDG-PET scans at baseline, 12-month and 24-month were retrieved from the ADNI database. In Table 1 is presented detailed demographic information (number of subjects, mean age and percentage of males) of the population in study.

Team	AD	CN	MCI
Number of subjects	58	75	135
Age at Baseline ($\mu \pm \sigma$)	76.1 ± 6.6	76.8 ± 6.6	78.6 ± 6.6
Age at month 12 ($\mu \pm \sigma$)	76.0 ± 4.7	77.0 ± 4.7	77.8 ± 4.8
Age at month 24 ($\mu \pm \sigma$)	75.2 ± 7.3	75.9 ± 7.3	77.1 ± 7.2
Sex (% of Males)	58.6	65.3	65.2

Table 1: Demographic information of the population in study. μ and σ represent the mean and standard deviation, respectively.

4.2. Classification Results

In this section, will be presented classification results obtained for AD vs CN, AD vs MCI and MCI vs CN. Note that in Table 2, the results that achieved higher accuracy for each method, are presented. Through the analysis of Fig. 5, one can observe that all the methods achieve approximately the same maximum accuracy, i.e., all the methods achieve an accuracy between 89% and 91%. However, some methods present higher performance when almost all brain values are selected (e.g. Group Lasso big cubes and Tree atlas and Tree cubes) and others present higher accuracy when only about half of the brain voxels are selected as important features for differentiating AD and CN individuals. One should notice that there are two methods that achieve higher performance when the number of features selected is really reduced, as is the case of Multi-task Lasso (achieving higher performance when $\lambda = 0.2$, for which the number of features selected is around 8.3% of the whole number of voxels) and Tree Lasso constructed with basis

on hierarchical clustering (that achieves higher accuracy when $\lambda = 0.1$, for which around 12% features are selected). Moreover, although all the methods show similar accuracies, the ones that consider less than 45% of the whole brain voxels are the ones that achieve a slightly higher accuracy.
It is notorious that the multi-task Lasso method seems to be the most stable to λ changes, i.e., its accuracy does not vary a lot with the number of features selected. This can indicate that features selected by this method are really important for distinguish AD from CN. Finally, the method that showed the highest accuracy value was Lasso, that achieved an accuracy of 91% and an AUC value of around the same percentage, for a $\lambda = 0.05$, which implies that around 24.5% of the whole brain were used for classification.

Through the analysis of Fig. 6, one can observe, once again, that all the methods achieve approximately the same accuracy, between 74% and 76%, except the Multi-task Lasso that, once more shows a different behaviour, achieving an accuracy of around 67%. However, this last method shows higher Area Under Curve (AUC) values, presenting a value around 74% for $\lambda = 0.4$ (the λ for which higher accuracy was achieved). In its turn, the other methods show values of AUC between 63% and 65%, for the same λ that originated the values accuracy already mentioned.
One can observe, in Table 2, that AUC values are low because, despite presenting high specificity, the values of sensitivity are too low for all the methods except Multi-task. This situation means that there are many false negatives (AD classified as MCI) and few false positives (MCI classified as AD). This situation is the worse one, once a patient with the disease is not classified as such and thus, no treatment is administrated to him in order to slow down the progression of the disease.
However, the multi-task Lasso, although achieving lower accuracy, it presents a sensitivity around 75% and specificity around 63%. Also, this method is more stable and robust than the others, not changing to much with λ. For AD vs MCI, all the methods achieved higher accuracy when selecting less features, comparing to AD vs CN. This decrease in the number of features selected is less evident for Group Lasso big cubes and Tree atlas that selected around 96% and 71.3% of the whole brain voxels, respectively. Moreover, Tree Hierarchical cluster and Multi-task methods selected approximately the same number of features, however some changes in location and mean weight are noticed.

Once again, Figure 7 shows that the results obtained with Lasso and with the other methods that consider spatial information are similar and the same is verified for the AUC values (Table 2). When λ is lower than 0.1, all the methods achieve an accuracy around 69%, even the multi-task. Also, all methods show an AUC curve between 65% and 67% for lower λ values. In this case, the opposite to AD vs MCI happens and thus, the AUC curve is below the accuracy curve because there are many false positives (CN individuals classified as MCI), i.e., the classification shows low specificity. Moreover, one should notice that since higher accuracy is obtained for small λ values, it means that those methods distinguishing more accurately MCI from CN individuals, when almost all the brain voxels are selected.

4.3. Regions of features selected

Since multi-task was the method that showed to be more robust (there was an equilibrium verified between accuracy and AUC values obtained), this method was the one selected for further analysis. In this way, the analysis of the regions in which the features selected by this method are included will be presented in this section and can be observed in Fig. 8.
For AD vs CN this method selects almost 60% of the hippocampus and amygdala voxels and selects any or a low percentage of voxels situated in the other regions. Also, this method weights more its features in those regions.
Concerning about cortical regions, more than half of the voxels of insular cortex, middle temporal gyrus, inferior temporal gyrus, frontal medial cortex, cingulate gyrus (posterior division), precuneous cortex, parahippocampal gyrus (anterior division) and temporal fusiform cortex (anterior and posterior division) are selected. From those regions, the ones that weight more are the cingulate gyrus and parahippocampal gyrus. This results are according to the background knowledge, since AD is characterized by changes in isocortical areas (Braak and Braak stage V and VI). However, in not so advanced stages of the disease, there are alterations in the hippocampus (Braak and Braak stage II) and then, in some limbic structures (Braak and Braak stage III and IV) such as hippocampal formation, amygdala and thalamus.
In this study, although the hippocampus and amygdala have shown to be important regions to distinguish AD from CN individuals, the cortical regions were attributed higher weights.
For AD vs MCI, the features selected are still more abundant in the Hippocampus and Amygdala, but in this experiment, the weights are assigning the same importance to those features and to the ones in Cerebral White Matter and Cerebral Cortex. Regarding cortical regions, in this case around half of the voxels included in middle temporal gyrus, inferior temporal gyrus, postcentral gyrus, superior parietal lobule, supramarginal gyrus and angular

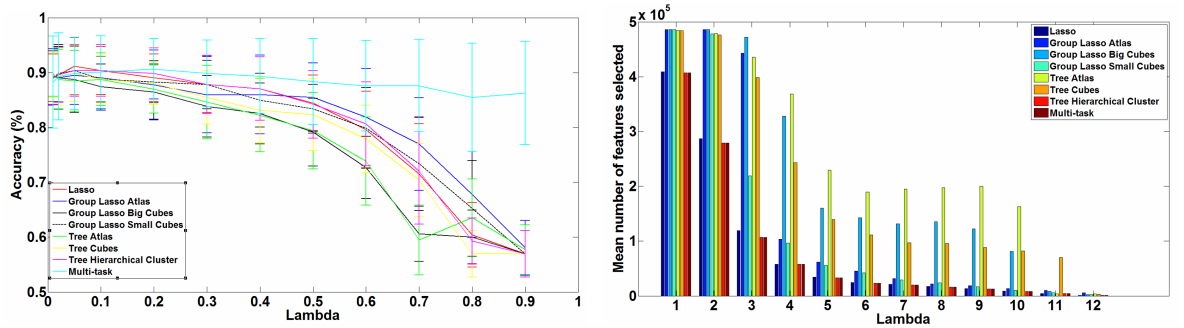

Figure 5: Accuracy(%) for each λ value (left) and Mean Number of Features Selected per λ, for AD vs CN.

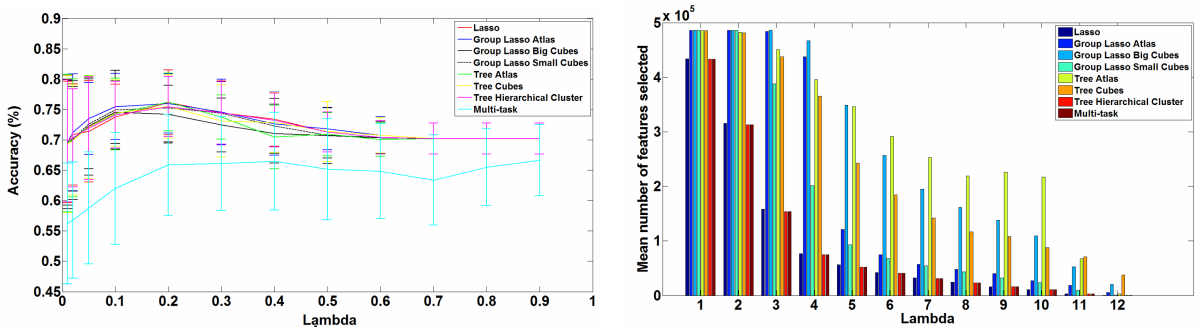

Figure 6: Accuracy(%) for each λ value (left) and Mean Number of Features Selected per λ, for AD vs MCI.

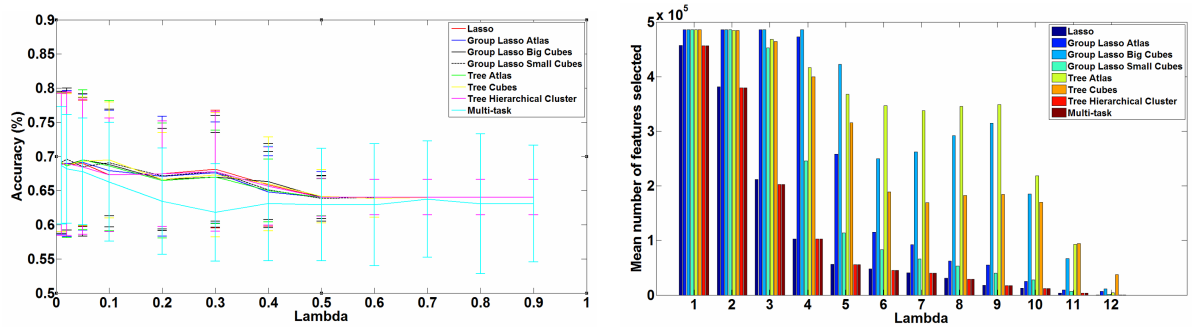

Figure 7: Accuracy(%) for each λ value (left) and Mean Number of Features Selected per λ, for MCI vs CN.

gyrus are selected. Moreover more than a half of the voxels of cingulate gyrus (posterior division), parahippocampal gyrus (anterior division) and precuneous are selected.
Since MCI patients are usually included in Braak and Braak III-IV stages, it means that some limbic structures (hippocampal formation, amygdala and thalamus) can be already affected in those patients. In this way, the obtained results are according to the expected, since it makes sense to take advantage from those limbic regions (that can be affected in both AD and MCI) and from isocortical regions, when differentiating AD from MCI individuals. Once again, the cingulate gyrus (posterior division) and the parahippocampal gyrus (anterior division) are the regions that are considered with higher weight.

Finally, for MCI vs CN it is possible to observe that almost all brain voxels are selected and the method recognizes, approximately, the same order of importance to all subcortical regions.
Concerning about the cortical regions used for distinguish MCI from CN individuals, since the anterior division of the parahippocampal gyrus includes

Samples	Methods	Lambda	Accuracy	AUC curve	Sens	Spec	Features
AD vs CN	Lasso	0.05	91.1 ± 5.2	90.7	87.8	93.5	24.5
	Group Atlas	0.05	89.5 ± 5.4	89.0	85.8	92.16	91.2
	Group Big cubes	0.01	89.5 ± 4.1	88.8	84.5	93.1	100
	Group Small cubes	0.05	90.4 ± 4.7	89.7	85.9	93.5	45
	Tree Atlas	0.01	89.8 ± 4.1	89.1	85.1	93.1	99.7
	Tree cubes	0.05	89.8 ± 5.1	89.2	85.4	93.1	82
	Tree CN	0.1	90.4 ± 4.7	89.8	86.0	93.5	12
	Multi-task	0.2	90.6 ± 5.5	92.9	91.7	89.9	8.3
AD vs MCI	Lasso	0.2	76.2 ± 5.4	64.4	35.7	93.2	11.6
	Groups Atlas	0.2	76.0 ± 4.9	64.6	36.4	92.7	24.9
	Groups Big cubes	0.1	74.6 ± 5.2	65.3	42.6	88.0	96.0
	Groups Small cubes	0.2	75.3 ± 5.7	63.2	33.4	92.9	19.2
	Tree Atlas	0.2	76.2 ± 4.8	64.9	36.8	93.0	71.3
	Tree cubes	0.2	75.5 ± 5.2	64.1	36.0	92.1	50.0
	Tree CN	0.2	75.5 ± 4.9	63.6	34.5	92.6	10.7
	Multi-task	0.4	66.5 ± 8.1	74.0	62.7	75.5	8
MCI vs CN	Lasso	0.05	69.0 ± 9.3	66.3	75.9	56.6	43.6
	Group Atlas	0.05	69.2 ± 10.0	66.2	76.8	55.6	100
	Group Big cubes	0.05	69.5 ± 9.7	66.8	76.5	57.0	100
	Group Small cubes	0.02	69.6 ± 10.4	67.8	74.4	61.3	100
	Tree Atlas	0.05	69.5 ± 10.2	67.0	76.0	57.9	100
	Tree cubes	0.1	69.5 ± 8.5	65.4	79.8	50.9	82.3
	Tree CN	0.02	69.0 ± 10.5	67.1	73.9	60.3	78.1
	Multi-task	0.01	68.7 ± 8.6	65.5	74.1	58.9	99.7

Table 2: Results obtained for the different methods used, including the λ that originated higher accuracy for each one, as well as the accuracy (mean $\pm$ standard deviation), AUC curve, sensitivity, specificity and percentage of brain voxels selected. All that results are represented for AD vs CN, AD vs MCI and MCI vs CN.

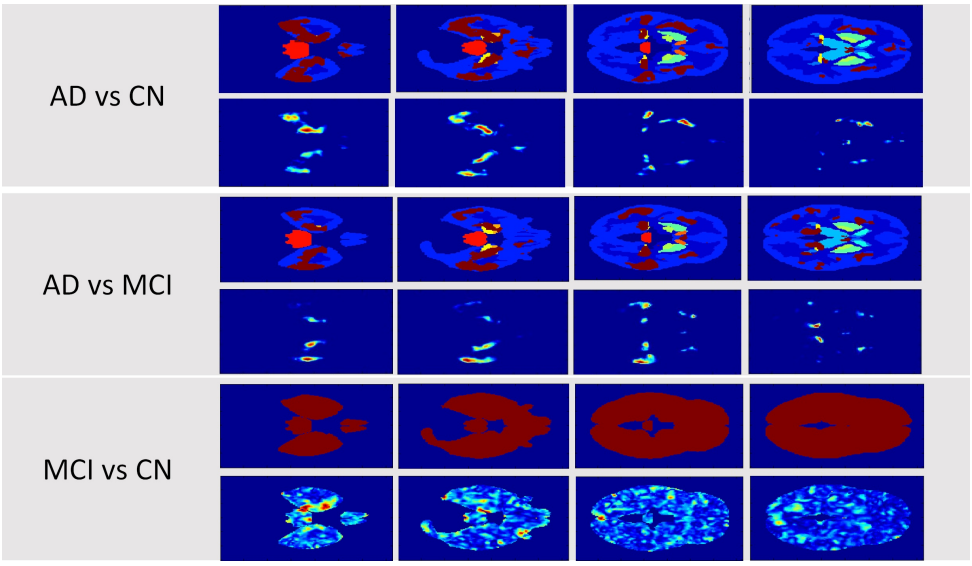

Figure 8: Features selected and mean weights of those features for the different methods in study for AD vs CN, AD vs MCI and MCI vs CN and for different slices of the brain - $Z = 30$, $Z = 35$, $Z = 45$, $Z = 50$, respectively

the perihinal and entorhinal cortices and since, in its turn, the entorhinal cortex is affected in Braak and Braak stage I, it was expectable to obtain an higher weight for this region. The same happens for

the other limbic regions (changed during Braak and Braak stage III and IV) that can be affected in the MCI individuals. In this way, the results obtained and are according to the expected.

5. Conclusions

At first glance, although being the simplest method, i.e. the only one that does not incorporates spatial or temporal information, Lasso did not present lower performance than the methods that incorporate spatial information.

In terms of performance, Lasso, Group Lasso and Tree Group Lasso (i.e., the methods that considered some spatial information) have shown a similar behavior, presenting satisfactory accuracy values for the population in study, but the sensitivity and specificity values calculated were not balanced. It is evident from the results expressed in table 2 that the use of much more MCI samples (the number of MCI sucjects is almost the double of the number of AD and CN individuals, as shown in Table 1) is decreasing the performance of Lasso, Group Lasso and Tree Group Lasso.

It should be noticed that although higher λ implies less number of features selected, the extent of the penalization introduced by a given λ is highly dependent on the structure considered for a method (except for the Lasso, where no spatial or temporal groups are considered) and on the weights attributed to the groups considered, as can be concluded by the number of features selected for different approaches based on the same method. In its turn, the Multi-task has shown to be the most robust and stable method, because it was the only one that did not changed abruptly with λ and that have shown similar values of sensitivity and specificity, what has conducted to an AUC value similar to the accuracy value. This method achieved an accuracy of 90.6%, 66.5% and 68% for AD vs CN, AD vs MCI and MCI vs CN, respectively and AUC value of 92.9%, 74% and 65.5%, for the same samples.

However, when classifying MCI vs CN, all the methods present approximately the same results and thus, it seems that multi-task is a more appropriated method for classifying more advanced stages of the disease.

Since AD is a progressive neurological disorder, metabolic decline occurs along time and thus, temporal information was expected to be useful for early AD diagnosis and prediction of the course of the disease.

Although being more robust, the Multi-task is also the method associated with higher computational costs once it takes much more time (around one day) running than the other methods (that run in less than half an hour).

In general, the features selected in the amygdala and hippocampus have shown to be important subcortical regions for distinguish AD, MCI and CN subjects.

Furthermore, the parahippocampal gyrus (cortical region) has shown to be one of the most important cortical regions (and the most weightened one). Those regions are also mentioned in the literature as important regions to AD, assuming special relevance to the doctors.

Further investigation should be conducted in order to improve the performance of the sparse logistic methods for AD classification. For that purpose, some methods should be considered in order to compensate the higher number of MCI individuals. The results obtained suggest that the compensation of that unbalanced number can be enough to obtain promising results using Lasso, Group Lasso and Tree Lasso. Moreover, further spatial structures should be explored and tested. For example, the development of spatial structures not based on cortical and subcortical atlases regions would be interesting. Furthermore, it would be also interesting to design a method able to conciliate temporal and spatial information at the same time because a method with those characteristics was not already explored in the literature. Finally, the same experiences can be performed for feature selection, but other classifiers can be used for classification, such as the widely used SVM. However, note that this classifier is not perfect and thus, some classification errors can arise, based on the C paramenter.

Acknowledgements

The author would like to thank Professor Margarida Silveira for all the guidance during the development of the present thesis.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] R. Casanova, C. T. Whitlow, B. Wagner, J. Williamson, S. A. Shumaker, J. A. Maldjian, and M. A. Espeland. High dimensional classification of structural mri alzheimers disease data based on large scale regularization. *Front Neuroinform*, 5:22, 2011.

[3] L. K. Ferreira and G. F. Busatto. Neuroimaging in alzheimer's disease: current role in clinical practice and potential future applications. *Clinics*, 66:19–24, 2011.

[4] C. Hennig and M. Kutlukaya. Some thoughts about the design of loss functions. *REVSTAT–Statistical Journal*, 5(1):19–39, 2007.

[5] F. Liu, L. Zhou, C. Shen, and J. Yin. Multiple kernel learning in the primal for multimodal

alzheimers disease classification. *Biomedical and Health Informatics, IEEE Journal of*, 18(3):984–990, 2014.

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

[7] M. Liu, D. Zhang, D. Shen, A. D. N. Initiative, et al. Ensemble sparse classification of alzheimer's disease. *NeuroImage*, 60(2):1106–1116, 2012.

[8] M. Liu, D. Zhang, D. Shen, A. D. N. Initiative, et al. Identifying informative imaging biomarkers via tree structured sparse learning for ad diagnosis. *Neuroinformatics*, 12(3):381–394, 2014.

[9] M. Liu, D. Zhang, P.-T. Yap, and D. Shen. Tree-guided sparse coding for brain disease classification. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 239–247. Springer, 2012.

[10] M. Maalouf, J. Ringman, and J. Shi. An update on the diagnosis and management of dementing conditions. *Reviews in neurological diseases*, 8:e68, 2011.

[11] L. P. R. MD and T. A. P. M. (12th Edition), editors. *Merritt's Neurology*. LWW, 2009.

[12] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. Ways toward an early diagnosis in alzheimers disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

[13] M. Stanford. Hierarchical cluster analysis @ONLINE.

[14] E. E. Tripoliti, D. I. Fotiadis, and M. Argyropoulou. A supervised method to assist the diagnosis and monitor progression of alzheimer's disease using data from an fmri experiment. *Artificial intelligence in medicine*, 53(1):35–45, 2011.

[15] F. Wang, P. Zhang, X. Wang, and J. Hu. Clinical risk prediction by exploring high-order feature correlations. In *AMIA Annual Symposium Proceedings*, volume 2014, page 1170. American Medical Informatics Association, 2014.

[16] I. Yakushev, A. Hammers, A. Fellgiebel, I. Schmidtmann, A. Scheurich, H.-G. Buchholz, J. Peters, P. Bartenstein, K. Lieb, and M. Schreckenberger. Spm-based count normalization provides excellent discrimination of mild alzheimer's disease and amnestic mild cognitive impairment from healthy aging. *Neuroimage*, 44(1):43–50, 2009.

[17] J. Ye, M. Farnum, E. Yang, R. Verbeeck, V. Lobanov, N. Raghavan, G. Novak, A. DiBernardo, V. A. Narayan, et al. Sparse learning and stability selection for predicting mci to ad conversion using baseline adni data. *BMC neurology*, 12(1):46, 2012.

[18] D. Zhang, D. Shen, A. D. N. Initiative, et al. Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers. *PloS one*, 7(3):e33182, 2012.