



Published in final edited form as:

Neuroinformatics. 2014 July ; 12(3): 381–394. doi:10.1007/s12021-013-9218-x.

Identifying Informative Imaging Biomarkers via Tree Structured Sparse Learning for AD Diagnosis

Manhua Liu,

Department of Instrument Science and Engineering, SEIEE, Shanghai Jiao Tong University, Dong Chuan Rd #800, Shanghai, China. Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Daoqiang Zhang,

Department of Computer Science and Engineering, Nanjing University of Aeronautics & Astronautics, Nanjing, China. Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Dinggang Shen, and

Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

the Alzheimer's Disease Neuroimaging Initiative

Manhua Liu: mhliu@sjtu.edu.cn; Dinggang Shen: dgshen@med.unc.edu

Abstract

Neuroimaging provides a powerful tool to characterize neurodegenerative progression and therapeutic efficacy in Alzheimer's disease (AD) and its prodromal stage—mild cognitive impairment (MCI). However, since the disease pathology might cause different patterns of structural degeneration, which is not pre-known, it is still a challenging problem to identify the relevant imaging markers for facilitating disease interpretation and classification. Recently, sparse learning methods have been investigated in neuroimaging studies for selecting the relevant imaging biomarkers and have achieved very promising results on disease classification. However, in the standard sparse learning method, the spatial structure is often ignored, although it is important for identifying the informative biomarkers. In this paper, a sparse learning method with tree-structured regularization is proposed to capture patterns of pathological degeneration from fine to coarse scale, for helping identify the informative imaging biomarkers to guide the disease classification and interpretation. Specifically, we *first* develop a new tree construction method based on the hierarchical agglomerative clustering of voxel-wise imaging features in the whole brain, by taking into account their spatial adjacency, feature similarity and discriminability. In this way, the complexity of all possible multi-scale spatial configurations of imaging features can be reduced to a single tree of nested regions. *Second*, we impose the tree-structured regularization on the sparse learning to capture the imaging structures, and then use them for selecting the most relevant biomarkers. *Finally*, we train a support vector machine (SVM) classifier with the selected features to make the classification. We have evaluated our proposed method by using the baseline

MR images of 830 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, which includes 198 AD patients, 167 progressive MCI (pMCI), 236 stable MCI (sMCI), and 229 normal controls (NC). Our experimental results show that our method can achieve accuracies of 90.2 %, 87.2 %, and 70.7 % for classifications of AD vs. NC, pMCI vs. NC, and pMCI vs. sMCI, respectively, demonstrating promising performance compared with other state-of-the-art methods.

Keywords

Alzheimer's disease diagnosis; Tree-structured sparse learning; Biomarker identification; Mild cognitive impairment; Group sparse learning

Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder characterized by progressive impairment of memory and other cognitive functions. Neuroimaging measurements, including magnetic resonance image (MRI) and fluorodeoxyglucose positron emission tomography (FDG-PET), provide a powerful in vivo tool for helping diagnosis and longitudinal follow-up study of AD and MCI (Desikan et al. 2009; Klöppel et al. 2008; Stonnington et al. 2010; Zhou et al. 2011; Oliveira et al. 2010; Leung et al. 2010; Davatzikos et al. 2010; Querbes et al. 2009; Filipovych and Davatzikos 2011; Duchesne et al. 2009; Fan et al. 2007a; Wee et al. 2011, 2012; Zhang and Shen et al. 2012a, b; Li et al. 2012). In recent years, a substantial research effort has been made to investigate many machine learning and pattern recognition technologies, such as sparse learning and support vector machines (SVM), in neuroimaging analysis to assist AD diagnosis (Davatzikos et al. 2008a; Magnin et al. 2009; Hinrichs et al. 2009; Cuingnet et al. 2011; Wolz et al. 2011; Liu et al. 2012a, b). Various methods have been proposed for processing and analysis of neuroimages to investigate the pathological changes related to brain diseases (Xue et al. 2006; Wu et al. 2006; Yang et al. 2008; Magnin et al. 2009; Shen et al. 1999; Jia et al. 2010; Tang et al. 2009). One major challenge in neuroimaging analysis is its huge dimensionality of the original imaging data. For example, a typical MRI scan of brain includes several millions of measurements on the respective image voxels. Direct classification with original data is not only computationally expensive, but also inaccurate, since not all imaging information in the brain is related with the disease. Thus, extraction and selection of discriminative imaging biomarkers are necessary and important for brain disease diagnosis and interpretation (Zhu et al. 2013; Chen et al. 2009; Fan et al. 2007b; Chu et al. 2012).

To achieve dimensionality reduction and also take into account the spatial structure of the neuroimaging data, we can resort to feature agglomeration. One popular method is to group voxels into multiple anatomical regions, i.e., regions of interest (ROIs), through the warping of a pre-labeled atlas, and then extract regional features such as anatomical volumes for classification (Lao et al. 2004; Magnin et al. 2009). However, this type of anatomical parcellation may not adapt well to the disease-related pathology since the abnormal region may be part of ROI or span over multiple ROIs, which could eventually affect the classification result. To address this problem, Fan et al. (2007b) proposed to adaptively

partition the brain image into a number of the most discriminative brain regions according to the similarity of local imaging features. Then, regional features were extracted for disease classification. The advantage of all ROI-based feature extraction methods is that it can significantly reduce the feature dimensionality and make the extracted features robust to noise, inter-subject variability, and registration errors. However, the extracted regional features are generally very coarse and not sensitive to small changes in the local brain region, which will eventually degrade the classification performance.

The above-mentioned limitations with the ROI-based features could be potentially addressed by the voxel-wise analysis method (Ishii et al. 2005), i.e., extracting voxel-wise features for image classification. However, voxel-wise image analysis is sensitive to noise, registration error, and inter-subject variability. Although these limitations could be alleviated by smoothing the imaging features via a Gaussian filter, the smoothing is often performed on the whole brain for all subjects uniformly and thus not adaptive to anatomical structures, shapes and abnormal regions. More importantly, the number of voxel-wise imaging features from the whole brain is always very large (i.e., in millions), while the number of training samples is very small (i.e., in hundreds) in the neuroimaging study. This could easily make the classification model over-fitted to the training set, thus not generalized well to the test set. Since only a few number of brain regions are relevant to the disease, it is important to identify the predictive information for facilitating disease classification and interpretation. Recently, the effectiveness of four feature selection methods has been investigated and compared for AD diagnosis in (Chu et al. 2012). Their results show that the feature selection does improve the classification accuracy, but depends on the method used. The *t*-test filtering is one of the commonly used method to select the most discriminative biomarkers according to their correlations to the class labels (Davatzikos et al. 2008a; Oliveira et al. 2010). However, this method evaluates each brain voxel independently and can handle neither the spatial structure nor the multivariate nature of neuroimaging data, thus limiting its ability to detect the complex population difference. Since the informative imaging features may be distributed over the distant brain regions, the combinations of features over these regions should also be taken into account for feature selection. Recently, L1-norm sparse learning method, e.g., Lasso, have been proposed to identify a subset of features for best representing the outputs (Ghosh and Chinnaiyan 2005; Tibshirani 1996). This method enforces sparsity on the individual features while ignoring their spatial structures in feature selection. Since the disease-induced abnormal changes often happen in a few correlated regions instead of isolated voxels, the structural relationship of imaging features is an important information source for identification of informative regions.

The existence of structural relationship in spatial features can be used to build better feature selection method (Hinrichs et al. 2009). Recently, group Lasso was extended from the L1-norm Lasso to find solutions that are sparse on the group level of features (Yuan and Lin 2006). It assumes that some feature groups are inherently “good” and thus sparsity should be enforced at the group level by a L1/L2 regularization, in which the L1-norm penalty is applied over the groups of input features, while the L2 norm is applied for the input features within the same group. The basic intuition is that some feature groups will have a higher fraction of useful features than other groups, and thus features should be selected from those

“good” groups. It is worth noting that the group Lasso assumes the availability of prior knowledge on how to group the features. However, as mentioned above, the morphological changes of brain structures resulting from pathological processes usually do not occur in the pre-identified regions with certain shapes, which makes it difficult to define the meaningful feature groups for sparse learning. Recently, a sparse learning method with a tree-structured regularization was proposed as an extension of group Lasso to consider the underlying multi-level grouping structure among the inputs or also the outputs for feature selection (Kim and Xing 2009; Liu and Ye 2010). The structured regularization with a predefined hierarchical tree is defined based on the group penalty, where each tree node is for one feature group and different tree levels represent different levels of groups.

On the other hand, the pathology of AD and MCI might cause subtle changes in specific brain ROIs (e.g., atrophy of hippocampus) and/or the whole brain (e.g., atrophy of gray matter). To obtain high classification performance, the selected informative features should capture different patterns of structural degeneration, from local to global fashion. In this paper, we propose to investigate the tree structured sparse learning method to identify the informative imaging features from multiple levels for guiding brain disease classification and interpretation. Specifically, the spatial relationships of the imaging features are encoded into the tree regularization of group sparse learning to guide the selection of relevant features. Figure 1 shows the flowchart of the proposed method. First, the voxel-wise gray matter (GM) density map is extracted from each MR brain image as the features to decode the brain disease. Second, to capture the hierarchical spatial relationship of the imaging features in the whole brain, a tree structure is constructed by gradually agglomerating the adjacent and coherent voxels into a hierarchy of groups with a hierarchical agglomerative clustering technique. Then, the constructed tree structure is imposed on the regularization of group sparse learning to select the relevant features. Finally, SVM classifier is trained to perform classification using the selected features. We evaluate our proposed classification method with the baseline MR brain images of ADNI database. The results demonstrate that, in addition to better classify the neuroimaging data with AD and MCI, our proposed method can also identify the relevant biomarkers to facilitate the interpretation of classification results.

The rest of this paper is organized as follows. The proposed method is presented in the “Materials and Method” section. In “Experimental Results” section, extensive experiments and comparisons with other methods on ADNI dataset are presented to demonstrate the advantage of the proposed method. Finally, we conclude this paper and discuss the possible future directions in the “Conclusion” section.

Materials and Method

Imaging Data and Preprocessing

The imaging data used for preparation of this paper were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). The ADNI was launched in 2003 by the National Institute on Aging (NIA) (<http://www.nia.nih.gov/Alzheimers/ResearchInformation/ClinicalTrials/ADNI.htm>), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration

(FDA), private pharmaceutical companies and non-profit organizations, as a \$60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), Positron Emission Tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, M.D., VA Medical Center and University of California, San Francisco. ADNI was the result of efforts of many co-investigators from a broad range of academic institutions and private corporations. The study subjects was recruited from over 50 sites across the U.S. and Canada and gave written informed consent at the time of enrollment for imaging and genetic sample collection and completed questionnaires approved by each participating sites Institutional Review Board (IRB). The initial goal of ADNI was to recruit 800 adults, aged from 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years. For the up-to-date information, please see www.adni-info.org.

Although the proposed method makes no assumption on a specific neuroimaging modality, MRI is widely available, non-invasive and often used as the first biomarker in the differential diagnostics of brain diseases caused by memory problems. In this paper, the T1-weighted MRI data is tested for demonstrating the performance of our proposed method. In ADNI, the MRI datasets include the standard T1-weighted MR images acquired sagittally using volumetric 3D MPRAGE with 1.25×1.25 mm in-plane spatial resolution and 1.2 mm thick sagittal slices. Most of these images were obtained with 1.5 T scanners, while a few were acquired using 3 T scanners. Detailed information about MR acquisition procedures is available at the ADNI website. The current study in this paper involves the baseline MR brain images from 229 normal control subjects, 403 MCI subjects (including 236 stable MCI and 167 progressive MCI subjects), and 198 AD patients.

The MR brain images are first preprocessed according to the previously validated and published techniques (Liu et al. 2012a). Specifically, an intensity inhomogeneity on the T1-weighted MR images was corrected using nonparametric nonuniform intensity normalization (N3) algorithm (Sled et al. 1998). Then, a robust and automated skull stripping method was applied for brain extraction and cerebellum removal (Wang et al. 2011). Each brain image is further segmented into three types of tissue volumes, e.g., gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) volumes. All tissue volumes will be spatially normalized together onto a standard space (also called the stereotaxic space) by a mass-preserving deformable warping algorithm proposed in (Shen and Davatzikos 2002, 2003). The warped mass-preserving tissue volumes reflect the spatial distribution of tissues in the original brain. We call these warped mass-preserving tissue volumes as the tissue density maps in this paper.

Since GM is more related to AD and MCI than WM and CSF, the voxel-wise GM densities are used as the imaging features to demonstrate the classification performance of our

proposed method. Given M training images, with each represented by a feature vector and a respective class label, the brain classification involves the step of selecting the most relevant features and also the step of decoding the disease states such as the class labels, as detailed below.

Feature Selection

The voxel-wise GM densities of the whole brain are of huge dimensionality, in comparison with the small number of subjects, which makes the disease classification and interpretation difficult. In addition, the subtle changes caused by AD or MCI might reside in specific brain regions with little prior knowledge. To identify the informative imaging biomarkers, the feature learning model should capture different patterns of the brain structural degeneration, from local to global fashion. Thus, we will consider three special aspects to develop the feature learning model in this paper. *First*, since the informative imaging biomarkers may be distributed over the distant brain regions, a multivariate model is learned to consider the combinations of features over the distant brain regions for handling the long-range interactions in feature selection. *Second*, the spatial structure of imaging features need to be specially considered for selecting more meaningful and informative biomarkers. *Third*, a multi-scale approach is needed for identification of the predictive regions by respecting the underlying multi-level structures of the imaging features. Accordingly, we propose a tree-structured sparse learning method to identify informative imaging biomarkers. In the following subsections, we will first introduce the L1-regularized coding and the group sparse learning briefly, and then describe our proposed feature selection method.

L1-Regularized Sparse Learning and Group Sparse Learning—Let X denote a $M \times Q$ feature matrix with the m -th row corresponding to the m -th image's feature vector $x_m = (x_m^1, \dots, x_m^q, \dots, x_m^Q) \in R^Q$, and y be a class label (column) vector of M images with y_m denoting the class label of the m -th image. A linear regression model can be used to decode the class outputs from a set of features as follows:

$$y = X\alpha + \varepsilon \quad (1)$$

where $a = (a_1, \dots, a_q, \dots, a_Q)^T$ is a vector of coefficients assigned to the respective features, and ε is an independent error term. The least square optimization is one of the popular methods to solve the above problem. When Q is large while the number of features relevant to the class labels is small, sparsity can be imposed on the coefficients of the least square optimization via L1-norm regularization for feature selection (Tibshirani 1996; Ghosh and Chinnaiyan 2005). The L1-norm least square problem, i.e., Lasso, can be formulated as:

$$\alpha = \arg \min_{\alpha} \|y - X\alpha\|^2 + \lambda \|\alpha\|_1 \quad (2)$$

where λ is a regularization parameter that controls the amount of sparsity on the coefficients. The non-zero elements in α indicate that the corresponding features are relevant to the class labels. The L1-norm sparse learning provides an effective multivariate regression model to select a subset of relevant features by taking into account both the correlations of features to the class labels and the combinations of individual features. However, this method imposes

the L1-norm sparsity on the individual features for feature selection, which completely ignores the spatial structure of imaging features. In this situation, the associated features should be jointly selected to identify the complex population difference, since the disease-induced abnormal changes often happen in the contiguous brain regions, instead of isolated voxels.

To alleviate the above problem, the group Lasso has been proposed as an extension of L1-norm sparse learning to use the groups of features, instead of individual features, as the units of feature selection (Yuan and Lin 2006). In the regularization, group Lasso applies the L1-norm penalty over the feature groups and the L2-norm penalty for the individual features within the same group. It can be formulated as below:

$$\alpha = \arg \min_{\alpha} \|y - X\alpha\|^2 + \lambda \sum_{j=1}^N w_j \|\alpha_{G_j}\|_2 \quad (3)$$

where $G_j(j=1, \dots, N)$ is a set of predefined non-overlapping feature groups and w_j is a weight assigned to the corresponding group. Specifically, the penalty acts as the L1-norm on the vector of $\|\alpha_{G_j}\|_2$, to make some $\|\alpha_{G_j}\|_2$ zero. Thus, the sparsity is imposed at the level of feature groups by the L1/L2-norm penalty to jointly select the features, so that the group structures among the features can be considered. Obviously, group Lasso assumes that the feature groups are available with certain prior knowledge. However, in practice, the structural relationships of the imaging features are not always pre-known in the neuroimaging study.

Tree-Structured Sparse Learning—Recently, in the literature, the L1/L2-norm penalty in group sparse learning has been extended to a more general setting with various types of complex structures on the sparsity patterns by allowing the overlapping of groups. In many applications, the rich spatial relationships of features can be naturally represented using a hierarchical tree structure, with the leaf nodes as individual features and the internal nodes as groups of features. Specifically, a hierarchical feature selection has been proposed to impose the tree structural relationships on the features, by defining groups with the particular overlapping patterns (Zhao et al. 2009). Brain image shows the spatial and structural correlations between the neighboring voxels, thus naturally forming a number of brain regions with different sizes and shapes. Also, since generally only a few brain regions over the whole brain are affected by the disease, the structured sparsity can be incorporated into sparse learning for selection of informative biomarkers. Moreover, different pathological processes might affect brain regions in different ways, and thus the disease-affected brain regions might have irregular shapes and different sizes and are not known in advance. Feature selection should capture different patterns of pathological degeneration from voxel level to group level. The rich spatial relationships of the brain volume can be naturally identified as a hierarchical tree structure where the sub-trees represent local brain regions while different levels of the tree indicate multiple scales of the parcellated brain regions. Then, L1/L2-norm regularization predefined by the tree structure can be imposed on the sparse learning to encourage a joint selection of structured relevant features.

The most important problem for the tree structured sparse learning is how to construct a meaningful hierarchical tree for grouping the imaging features. A simple method for tree construction is to hierarchically partition the whole brain into 3D patches (cubic) of multiple scales, with each level of tree consisting of the non-overlapping patches of one scale (Liu et al. 2012a, b). Although this method considers the spatial adjacency of the voxel-wise features, the morphological changes of brain structures resulting from pathological processes usually do not occur in the regions necessarily with regular shapes. In fact, the disease-affected brain regions might have irregular shapes and different sizes. To capture more natural feature groups, a hierarchical agglomerative clustering was used to adaptively produce a hierarchical representation of the features by feature agglomeration, with each feature group exhibiting similar characteristics under the spatial constraint (Jenatton et al. 2011). The advantage of agglomeration is that the selected brain regions may be spatially coherent. However, this method did not take into account the target information in forming the feature groups and thus the grouped brain regions may not display similar characteristics with respect to the classification label. In the following subsection, we will introduce a new tree construction method based on the hierarchical agglomerative clustering of the voxel-wise imaging features, by taking into account the spatial adjacency, feature similarity and discriminability.

Tree Construction by Hierarchical Agglomerative Clustering—Assuming that the $M \times Q$ feature matrix X consists of M training subjects represented by Q voxel-wise imaging features, as mentioned above. We seek to group the neighboring voxels into a tree structure for representation of voxel-wise imaging features in a bottom-up way. The hierarchical agglomerative clustering is applied at this stage to encode the rich structure of voxel-wise imaging features based on a criterion defined below. It first treats each voxel as a singleton cluster and then iteratively agglomerates a pair of neighboring clusters until all clusters have been merged into a single cluster. Thus a binary tree is produced to represent a hierarchy of clusters with each node associated with the cluster obtained by merging its children clusters. The root of tree gathers all the voxels, while the leaves are the clusters, each consisting of a single voxel. To recover the true spatial support of a discriminative pattern embedded in an image, three items are taken into consideration for definition of clustering criterion.

The *first item* is the spatial adjacency which indicates that only the neighboring clusters can be merged together. Since the spatially adjacent voxels of brain images are usually correlated, it is reasonable to require that the two merged clusters should have neighboring voxels. In addition, it is well-known that the grouped voxels should have similar characteristics. Thus, the *second item* is the consistency of the local imaging features which are directly related to the feature similarity or the uniformity of the grouped voxels. Two groups of features are spatially consistent if these features are similar to each other. This characteristic can help form boundaries of clusters with similar properties and also identify the brain regions with different behaviors. In this work, the similarity of two features is simply estimated by the Euclidean distance of feature vectors.

The *third item* is the discrimination of the imaging features with respect to the classification task. The discriminative power of a feature can be quantitatively measured by its relevance to the classification task, which is usually computed as the correlation between this feature

and the class labels in a training dataset (i.e., with the normal samples labeled as -1 and the disease samples labeled as $+1$). For our continuous imaging features, linear correlation measures are easier to compute and are robust to over-fitting, and thus are widely used for feature selection in machine learning. We apply t -test to measure the relevance of each feature to the classification task using the p value which is in the range of $[0, 1]$. The smaller p -value indicates larger relevance of the feature to the classification task. Thus, the discriminative power of the i -th feature is measured by:

$$TS(i) = 1 - p(i) \quad (4)$$

Where $p(i)$ is the p value of the t -test for the i -th feature. Usually, to enhance the feature discrimination and also the robustness to noise, a parcellation (division) of the brain region should focus on partition of the strongly discriminative regions into the spatially tiny regions, while leaving the uninformative regions un-partitioned. Otherwise, grouping the informative features with others will decrease the discriminative power of the combined features.

By taking the above three items into account, two adjacent and spatially-consistent clusters with less informative features would be first merged together to increase their discriminative power and then compete with the informative parts in feature selection. Thus, at each step of the hierarchical agglomerative clustering, we group two adjacent clusters C_k and C_l that will minimize the measure defined below:

$$\Delta(k, l) = \frac{1}{|C_k| \times |C_l|} \sum_{i \in C_k, j \in C_l} \|x^i - x^j\|_2 \times (TS(i) + TS(j)) \quad (5)$$

where $|C_k|$ denotes the number of features in the cluster C_k . Different from the traditional hierarchical agglomerative clustering method that is designed only based on the spatial consistency, the above measure can incorporate the information of both feature consistency and discrimination for forming groups. Thus, the less informative imaging features can be adaptively merged into larger groups to enhance their robustness to noise and then compete with other groups. Note the informative features are less grouped in order to maintain their discriminability. Figure 2 lists the detailed steps of our proposed tree construction method by the hierarchical agglomerative clustering. Finally, a tree-structured hierarchy of the imaging features can be constructed on top of all brain voxels. Figure 3 further shows a sample hierarchical index tree consisting of 6 leaves and 5 nodes, which is constructed with 6 adjacent image voxels.

Feature Selection by Tree-Structured Sparse Learning—Based on the outputs of the tree construction presented above, we can perform feature selection by combining the hierarchical tree structure of imaging features with the sparsity-inducing penalty in the group sparse learning. The hierarchical tree constructed by the above agglomerative clustering can generate totally $Q-1$ overlapping feature groups, where Q is the number of available voxels. The largest feature group lies in the highest level of the tree, i.e., the root of tree that consists of all imaging features. However, the informative regions often occur in

the sub-regions of the whole brain. Thus, it may not be optimal to include all the nodes of the constructed tree in the sparse learning. Accordingly, we can cut the tree at a given level of hierarchy to produce a subset of tree nodes as a hierarchical parcellation of the brain volume. The tree cut will be optimized by computing a cross-validated classification performance within the training data set. Specifically, we select the tree cut that yields the highest classification performance. Then, all the descendant tree nodes are used as the overlapping groups for sparse learning.

Assume that the tree nodes produced by the tree cut are denoted as $T = \{G_1, \dots, G_j, \dots, G_{N_c}\}$, where G_j consists of all descendant leaves included in the j -th tree node. The tree leaves are the voxel-wise imaging features which is the finest level of group. Different levels of tree nodes indicate different scales of feature groups. The higher level of tree nodes means the coarser scale of feature groups. The index set of a child node is a subset of its parent node, i.e., parent node overlaps with its child nodes. Traditional L1-norm sparsity-inducing penalty, i.e., $\|\alpha\|_1$ in Eq. (2), yields sparsity at the level of individual imaging features, ignoring the potential structures existing between larger groups of features. The group sparse learning in Eq. (3) can make use of the group structures of feature set, which should be known in advance. To combine the above two methods, the tree-guided group sparse learning method with the hierarchical structured penalty can be formulated as:

$$\alpha = \arg \min_{\alpha} \|y - X\alpha\|^2 + \lambda_1 \|\alpha\|_1 + \lambda_2 \sum_{j=1}^{N_c} w_j \|\alpha_{G_j}\|_2 \quad (6)$$

where α_{G_j} is the set of coefficients assigned to all features within the tree node G_j , and w_j is a predefined positive weight for node G_j and is usually set to be proportional to the square root of the group size. The first penalty promotes sparsity at the level of individual features while the second penalty promotes sparsity at the level of overlapping groups. Since each node represents a subtree of T , if one node is selected, all its descendant child nodes in tree T will also be selected. In this way, we can make use of the hierarchical structure of feature groups to guide the selection of informative features from voxels to multi-scale groups.

The tree penalized convex minimization problem in Eq. (6) is more challenging to solve than both Lasso and group Lasso, since the tree-structured regularization is much more complex. An efficient optimization has been proposed for tree-structured group Lasso, where the structure over the features is represented as a hierarchical tree that is composed of multi-scale image patches (Liu and Ye 2010). In light of the results from (Liu and Ye 2010), a large class of the convex minimization problem penalized by the grouped tree structure as defined in Eq. (6) can be solved efficiently. This work showed that the associated Moreau-Yosida regularization admits an analytical solution and an efficient algorithm was proposed for determining the effective interval for the regularization parameter λ . This efficient algorithm has a time complexity comparable to both Lasso and group Lasso and is employed in this paper to solve the above convex minimization problem with a tree-guided sparse regularization. Finally, the features with non-zero coefficients are selected for subsequent classification.

Classification

Based on the selected imaging features by the tree-guided sparse learning method, a classifier model will be trained to make the final classification. There are various classifier models investigated for classification of brain images. Among them, SVM is one of the widely used classifiers because of its high classification performance (Fan et al. 2007b; Davatzikos et al. 2008b; Zhang et al. 2011; Magnin et al. 2009). SVM constructs a maximal margin classifier in a high-dimensional feature space by mapping the original features using a kernel-induced mapping function. For simplicity, we choose the SVM model with a linear kernel as the classifier, which is implemented by the SVM functions in MATLAB software (Kecman 2001). The value of the box constraint C for the soft margin is optimized by the training data set in the cross validation. (Note that this C is different from the parameter C used in Fig. 2.)

Experimental Results

We evaluate the proposed method with the T1-weighted baseline MR brain images of 830 subjects from ADNI database, which include 198 AD patients, 236 stable MCI (sMCI), 167 progressive MCI (pMCI), and 229 normal controls (NC). Table 1 provides a summary of the demographic characteristics of the studied subjects (denoted as mean \pm standard deviation).

Experimental Design

Before performing disease classification, the image preprocessing was performed as described in “Imaging Data and Preprocessing” section. The spatially normalized tissue volumes called as tissue densities are used for classification in this paper. To reduce the effects of noise, registration inaccuracy, and inter-individual anatomical variations, tissue density maps were further smoothed using a Gaussian filter and then down-sampled by a factor of 4 for the purpose of saving the computational time and memory cost. In this experiment, we use only the GM density map as the imaging features because of its more relevancy to AD and MCI.

The tree-structured sparse learning method is implemented using the SLEP package downloaded at <http://www.public.asu.edu/~jye02/Software/SLEP>. Since some recent publications (Wolz et al. 2011; Cuingnet et al. 2011) were based on three pairs of classifications, our proposed method is performed to test three classification problems related to AD and MCI, which are, respectively, AD vs. NC, pMCI vs. NC, and sMCI vs. pMCI classifications. To statistically evaluate the classification performance, we conduct standard 10-folds cross-validation to compute the classification accuracy which evaluates the proportion of correctly classified subjects among the test dataset. In addition, we also compute the sensitivity (SEN), i.e., the proportion of AD (or MCI) patients correctly classified, and the specificity (SPE), i.e., the proportion of correctly classified normal controls for further evaluation. In each time, one fold of the data set was used for testing, while the other remaining 9 folds were used as training data. The construction of the hierarchical tree is performed on the training data set, and then both the feature selection by the tree-structured sparse learning and the SVM classification model are conducted with the training data. The training set can be further divided into 10 folds to fine-tune the parameters

in our method when needed. In the tree-structured sparse learning method, the parameters λ_1 and λ_2 in Eq. (6) are proportional to each other and thus can be controlled by one parameter such as λ_2 which is optimized in the range of [0 1] with 10 folds of the training data set.

Results on Disease Classification

The first experiment is to test the effect of structural constraints on feature selection and classification. We compare our proposed method against other two feature selection methods without using the feature structure information. The first feature selection method is the *t*-test, which is one of the commonly-used approaches in the literature. The *t*-test is performed at each voxel of the training data, and then we set different thresholds for the absolute *p*-value based on the number of voxels (that we want to select) at each level. The second feature selection method is based on the L1-norm sparse learning (Lasso) which applies the L1-norm on the imaging features, i.e., the first regularization of Eq. (6). In this method, we can adjust the regularization parameter λ_1 to control the sparsity and select various numbers of features.

For fair comparison, we test the classification performance with respect to different level of selected features in all three methods. In the proposed tree-structured sparse learning method, we change the value of regularization parameter λ_2 in the range of [0 1] to adjust the sparsity of sparse learning and thus obtain different numbers of selected features. All these methods are tested on the same 10-fold partition of the data set. Given a regularization parameter, we can obtain a classification accuracy and the corresponding averaging number of selected features over the 10 folds. For better showing the results, we use polynomial models to fit all the data obtained by different regularization parameters. Figure 4a, b and c show the classification results (the fitted data plots of classification accuracies vs. the number of selected features) for AD vs. NC, pMCI vs. NC, and sMCI vs. pMCI, respectively. From these results, we can see that the proposed tree-guided method can achieve better classification accuracy than both *t*-test and Lasso based methods, especially when the number of selected features is small (i.e., $<1 \times 10^4$, less than half of available features). This shows that, imposing the structural relationships of features on the regularization of sparse learning can help select the discriminative features for better classification. When further increasing the number of selected features, the improvement of classification accuracy by the proposed method will decrease, compared with other two methods. These results further demonstrate the importance of the tree-structured regularization because further increasing the number of selected features will reduce the effect of structure constraint on selection of discriminative features. All these results show that the proposed tree-structured sparse learning method can make full use of the hierarchical structure of imaging features to select the informative biomarkers and then improve the disease classification.

The second experiment is to test the effectiveness of the proposed tree construction method in the tree-structured sparse learning on the disease classification performance. In this experiment, we compare the proposed method with other four methods. The first method is a baseline method that uses all the imaging features of the whole brain, without feature selection, for disease classification. (The other three methods are corresponding to different

strategies used to define the tree structure for sparse learning, named as the 2nd, 3rd, and 4th methods next.) The 2nd method (namely ‘Patch-based’) is obtained from our previous work (Liu et al. 2012b), which defined the hierarchical tree by dividing brain image into multi-scale 3D patches and using the multi-scale patches as the multi-level tree nodes for guiding feature selection. The 3rd method (namely ‘Cluster-based’) defines the tree nodes by using the agglomerative clustering and also taking into account their spatial adjacency and feature similarity. On the other hand, the specific regions of interest (ROIs) can also be generated by grouping the voxels of brain images into anatomical regions through the registration of a labeled atlas (Kabani et al. 1998). This anatomical parcellation can take the local features and their similarity into account for feature grouping. Thus, in the 4th method (namely ‘ROI-based’), we define the feature groups, i.e., the tree nodes, with the 93 manual ROIs as in (Kabani et al. 1998), instead of adaptively defining the feature groups using the hierarchical clustering as described above. The optimization of these methods is same as our proposed method. All the methods in this experiment are tested on the same 10 runs of 10-fold cross-validation, and the averaged results are reported.

The classification results by the above 4 methods are listed together with our proposed method in Table 2. From these results, we can see that the feature selection by the sparse learning with the patch-based tree structure obtains slight improvement (1 %) for classification of AD vs. NC when compared with the baseline method that uses the whole-brain features. Although the patch-based tree structure takes the local adjacency of imaging features into consideration by multi-scale patches, it cannot make full use of the information in imaging features and class labels for describing both local feature similarity and discrimination during the tree construction. This limits its ability to use the effective spatial structures for identifying the informative features. On the other hand, the Cluster-based method defines multi-level feature groups by taking into account the local similarities, but uses no discrimination information for grouping features, thus leading to possibly the non-informative features. Similarly, the feature selection by the ROI-based tree structure cannot obtain obvious improvement for classifications when compared with the baseline method using the whole-brain features, except some slight improvement for classification of AD vs. NC. Although the feature groups defined by ROIs take the local similarities into consideration, the respective tree structure consists of just one-level non-overlapping groups and, more importantly, no discrimination information with respect to the classification task is used to define the feature groups. In fact, the informative region of abnormality might be part of ROI or span over multiple ROIs, thus potentially reducing the statistical power of ROI-based feature selection. On the other hand, our proposed tree construction by hierarchical agglomerative clustering can adaptively generate the feature groups, and take into account both the feature similarity and the discrimination of adjacent voxels. Thus our proposed method achieves better classification performance than other methods.

Results on Biomarker Identification

The proposed method also aims to identify the informative biomarkers that are associated with the AD status to improve the disease interpretation. We examine the selected imaging features by the proposed tree-structured sparse learning method with the best regularization parameters. It is worth noting that the feature selection is performed on the training data

only. Thus, the selected imaging biomarkers at each cross-validation fold may be different. For example, we checked the selected features from all cross-validation folds, and found that some selected features do vary across different folds of different partitions. Thus, we compute the frequencies of the voxels included in the selected features over the cross-validation folds of all partitions for disease classification. Then we further identify the imaging features with the frequency values (normalized in the range of [0 1]) larger than a threshold (set to 0.4 in our experiments) as the selected imaging biomarkers. Similarly, we also provide the selected features by Lasso for comparison. Figures 5, 6, and 7 show the identified biomarkers from the GM density map by both Lasso and tree-structured sparse learning methods for AD vs. NC, pMCI vs. NC, and pMCI vs. sMCI classifications, respectively. It can be observed that the features selected by the proposed method are usually grouped at the relevant regions which helps interpretation of the obtained results. The regions identified by the proposed method include hippocampus, parahippocampal gyrus, entorhinal cortex, and amygdala, which are consistent with those reported in the literature for AD and MCI studies (Cuingnet et al. 2011; Zhang et al. 2011; Hinrichs et al. 2009). These results verify the effectiveness of the proposed method for guiding the identification of relevant biomarkers.

Discussion

In this study, we have evaluated the classification performance of the proposed method with 830 baseline MR brain images, acquired in the ADNI study. Our results demonstrate that the feature selection method by the tree-structured sparse learning can improve the classification performance when compared to both the *t*-test and the L1-norm *Lasso* methods, which do not consider the data structure during the feature selection. In addition, the proposed tree construction method can result in a more powerful classifier than other tree construction methods, including the Patch-based, Cluster-based and ROIs-based methods.

Furthermore, in Table 3, the results of the proposed classification method are compared with five recent classification methods, also using the baseline T1-weighted MRI data of ADNI dataset, as briefly described:

- In (Hinrichs et al. 2009), the linear program (LP) boosting method with a novel additional regularization was proposed to incorporate the spatial smoothness of MR imaging space into the learning process and improve the classification accuracy. Only classification results for AD vs. NC were provided in that paper.
- In (Zhang et al. 2011), 93 volumetric features were extracted from the 93 regions of interest (ROI) in GM densities for both classifications of AD vs. NC and MCI vs. NC. A single SVM classifier was constructed to make the final classification.
- In (Cuingnet et al. 2011), ten methods on different types of structural MRI-based features, which included five voxel-wise imaging features based methods, three cortical thickness based methods, and two hippocampus based methods, were tested and compared with a linear SVM classifier. For classifications of AD vs. NC and MCI vs. NC, the best classification results, which were obtained using voxel-wise GM densities, are provided for comparison in our paper. For prediction of

MCI conversion, i.e., classification of pMCI vs. sMCI, the best results that used hippocampal volume are provided for comparison in our paper.

- In (Wolz et al. 2011), instead of using single MRI feature, four types of MRI-based features, i.e., hippocampal volume, tensor-based morphometry, cortical thickness, and manifold-learning based features, were combined to achieve improved classification accuracies. Both linear discriminant analysis (LDA) and SVM classification approaches are tested for classifications of AD vs. NC, pMCI vs. NC, and pMCI vs. sMCI. For comparison, we present their best results that were obtained with the LDA classification approach.
- More recently, in (Chu et al. 2012), the impact of feature selection and sample size on brain disease classification was extensively studied by using the GM features and SVM classifier. In particular, they compared four different feature selection methods, which are the pre-selected ROIs based on prior knowledge, univariate *t*-test filtering, recursive feature elimination, and *t*-test filtering constrained by ROIs. Their experimental results showed that the most accurate classification was achieved by the feature selection using prior knowledge about the regions of brain atrophy found in previous studies, i.e., using all GM voxels in the hippocampal and parahippocampal masks. Therefore, their best results reported for classifications of AD vs. NC, MCI vs. NC, and pMCI vs. sMCI are used here for comparison.

Table 3 summarizes the classification results of the above five methods, along with our proposed method. It can be observed that our results compare favorably to all other existing methods for brain disease classifications. It is worth noting that the variations of the reported results may be due to the use of different MRI feature extraction and classification methods, and also the use of different ADNI subjects. All these make the comparison of the results complicated, since it is difficult to implement all other methods on the same conditions for fair comparison. In addition, the variations in the size of test samples, the use of cross-validation, and separating the training and testing sets can also make the fair comparison difficult to achieve. Nevertheless, our results were obtained using the largest data set, consisting of almost all subjects in the ADNI database.

Conclusion

The pathology of AD and MCI might cause the changes of brain regions in different ways, and thus the disease-affected regions might have various sizes and irregular shapes with little prior knowledge. To identify the informative biomarkers, feature selection should capture different patterns of pathological degeneration from fine to coarse scales. In this paper, a tree structured sparse learning method is proposed to identify the informative biomarkers for classifications of AD and MCI. Specifically, a hierarchical tree is constructed to capture the rich structural relationships among the imaging features by using the agglomerative clustering and taking into account their spatial adjacency, feature similarity and discriminability, and then a tree structured regularization is imposed on sparse learning for feature selection. The tree structured sparse learning can provide an effective way to identify more meaningful biomarkers to facilitate brain disease classification and interpretation. Experimental results on the ADNI dataset show that the proposed method can

not only identify the grouped relevant biomarkers, but also improve the performance of brain disease classification.

In the current paper, we validated our method using MRI data from ADNI database. However, our method can also be extended to use other modality of data for AD or other brain disease classification. In the future work, we will evaluate our method on other imaging data, e.g., PET. Moreover, since recent studies have shown that different modalities of neuroimaging data can be combined to provide complementary information and achieve better classification performance, we will extend our method into the use of multi-modality biomarkers for further improving the accuracy of brain disease classification.

Information Sharing Statement

The MRI brain image dataset used in this paper was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) which is available at <http://www.adni-info.org>. In this paper, the proposed method was implemented based on the SLEP package, which is also publicly available at <http://www.public.asu.edu/~jye02/Software/SLEP>. Some other source codes and binary programs used and developed in this paper are available in our website (<http://bric.unc.edu/ideagroup/>).

Acknowledgments

This work was supported in part by NIH grants EB006733, EB008374, EB009634 and AG041721, MH100217, and AG042599, and by National Natural Science Foundation of China (NSFC) grants (No. 61375112, No. 61005024) and Medical and Engineering Foundation of Shanghai Jiao Tong University (No. YG2012MS12). This work was also partially supported by the National Research Foundation grant (No. 2012-005741) funded by the Korean government, and supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), and by Jiangsu Natural Science Foundation for Distinguished Young Scholar (No. BK20130034), and NUA Fundamental Research Funds under grant (No. NE2013105). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc., F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles.

References

- Chen Y, An H, Zhu H, Stone T, Smith JK, Hall C, et al. White matter abnormalities revealed by diffusion tensor imaging in non-demented and demented HIV+ patients. *Neuro Image*. 2009; 47(4): 1154–1162. [PubMed: 19376246]
- Chu C, Hsu AL, Chou KH, Bandettini P, Lin C. for the Alzheimer's Disease Neuroimaging Initiative. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuro Image*. 2012; 60(1):59–70. [PubMed: 22166797]
- Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehericy S, Habert MO, et al. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *Neuro Image*. 2011; 56(2):766–781.10.1016/j.neuroimage.2010.06.013 [PubMed: 20542124]

- Davatzikos C, Fan Y, Wu X, Shen D, Resnick SM. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of Aging*. 2008a; 29(4):514–523.10.1016/j.neurobiolaging.2006.11.010 [PubMed: 17174012]
- Davatzikos C, Resnick SM, Wu X, Parnpi P, Clark CM. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *Neuro Image*. 2008b; 41(4):1220–1227.10.1016/j.neuroimage.2008.03.050 [PubMed: 18474436]
- Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging*. 2010; 32(12):2322.e2319–2322.e2327. [PubMed: 20594615]
- Desikan RS, Cabral HJ, Hess CP, Dillon WP, Glastonbury CM, Weiner MW, et al. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain*. 2009; 132(Pt 8):2048–2057. [PubMed: 19460794]
- Duchesne S, Caroli A, Geroldi C, Collins DL, Frisoni GB. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. *Neuro Image*. 2009; 47(4):1363–1370. [PubMed: 19371783]
- Fan Y, Rao H, Hurt H, Giannetta J, Korczykowski M, Shera D, et al. Multivariate examination of brain abnormality using both structural and functional MRI. *Neuro Image*. 2007a; 36(4):1189–1199. [PubMed: 17512218]
- Fan Y, Shen D, Gur RC, Gur RE, Davatzikos C. COMPARE: Classification Of Morphological Patterns using Adaptive Regional Elements. *IEEE Transactions on Medical Imaging*. 2007b; 26(1):93–105. [PubMed: 17243588]
- Filipovych R, Davatzikos C. Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI). *Neuro Image*. 2011; 55(3):1109–1119.10.1016/j.neuroimage.2010.12.066 [PubMed: 21195776]
- Ghosh D, Chinnaiyan AM. Classification and selection of biomarkers in genomic data using LASSO. *Journal of Biomedicine and Biotechnology*. 2005; 2005(2):147–154. [PubMed: 16046820]
- Hinrichs C, Singh V, Mukherjee L, Xu G, Chung MK, Johnson SC. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. *Neuro Image*. 2009; 48(1):138–149. [PubMed: 19481161]
- Ishii K, Kawachi T, Sasaki H, Kono AK, Fukuda T, Kojima Y, et al. Voxel-based morphometric comparison between early-and late-onset mild Alzheimer's disease and assessment of diagnostic performance of z score images. *American Journal of Neuroradiology*. 2005; 26(2):333–340. [PubMed: 15709131]
- Jenatton, R.; Gramfort, A.; Michel, V.; Obozinski, G.; Bach, F.; Thirion, B. Multi-scale mining of fMRI data with hierarchical structured sparsity. *IEEE International Workshop on Pattern Recognition in Neuro Imaging*; Seoul, Korea. May 16–May 18 2011; p. 69-72.
- Jia H, Wu G, Wang Q, Shen D. ABSORB: Atlas building by self-organized registration and bundling. *Neuro Image*. 2010; 51(3):1057–1070. [PubMed: 20226255]
- Kabani N, MacDonald D, Holmes CJ, Evans A. A 3D atlas of the human brain. *Neuro Image*. 1998; 7(4):S717.
- Kecman, V. *Learning and soft computing-support vector machines, neural networks, fuzzy logic systems*. Cambridge: The MIT Press; 2001.
- Kim, S.; Xing, EP. Tree-guided group lasso for multitask regression with structured sparsity. 2009. ArxivpreprintarXiv:0909.1373
- Klöppel S, Stonnington CM, Chu C, Draganski B, Schill RI, Rohrer JD, et al. Automatic classification of MR scans in Alzheimer's disease. *Brain*. 2008; 131(3):681–689. [PubMed: 18202106]
- Lao Z, Shen D, Xue Z, Karacali B, Resnick SM, Davatzikos C. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. *Neuro Image*. 2004; 21(1):46–57. [PubMed: 14741641]
- Leung K, Shen KK, Barnes J, Ridgway G, Clarkson M, Frupp J, et al. Increasing power to predict mild cognitive impairment conversion to Alzheimer's disease using hippocampal atrophy rate and statistical shape models. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*. 2010; 13:125–132.

- Li Y, Wang Y, Wu G, Shi F, Zhou L, Lin W, Shen D. Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. *Neurobiology of aging*. 2012; 33(2):427.e15–427. e30. [PubMed: 21272960]
- Liu J, Ye J. Moreau-Yosida regularization for grouped tree structure learning. *Advances in Neural Information Processing Systems*. 2010; 23:1459–1467.
- Liu M, Zhang D, Shen D. Ensemble sparse classification of Alzheimer's disease. *Neuro Image*. 2012a; 60(2):1106–1116.10.1016/j.neuroimage.2012.01.055 [PubMed: 22270352]
- Liu, M.; Zhang, D.; Yap, P-T.; Shen, D. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012. Vol. 7512. Berlin Heidelberg: Springer; 2012b. Tree-Guided Sparse Coding for Brain Disease Classification; p. 239-247. *Lecture Notes in Computer Science*
- Magnin B, Mesrob L, Kinkingnehun S, Pelegrini-Issac M, Colliot O, Sarazin M, et al. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology*. 2009; 51(2):73–83. [PubMed: 18846369]
- Oliveira PJ, Nitrini R, Busatto G, Buchpiguel C, Sato J, Amaro EJ. Use of SVM methods with surface-based cortical and volumetric subcortical measurements to detect Alzheimer's disease. *Journal of Alzheimer's Disease*. 2010; 19(4):1263–1272.
- Querbes O, Aubry F, Pariente J, Lotterie JA, Demonet JF, Duret V, et al. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain*. 2009; 132(Pt 8): 2036–2047. [PubMed: 19439419]
- Shen D, Davatzikos C. HAMMER: hierarchical attribute matching mechanism for elastic registration. *Medical Imaging, IEEE Transactions on*. 2002; 21(11):1421–1439.
- Shen D, Davatzikos C. Very high resolution morphometry using mass-preserving deformations and HAMMER elastic registration. *Neuro Image*. 2003; 18(1):28–41. [PubMed: 12507441]
- Shen D, Wong W, Ip HHS. Affine-invariant image retrieval by correspondence matching of shapes. *Image and Vision Computing*. 1999; 17(7):489–499.
- Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *Medical Imaging, IEEE Transactions on*. 1998; 17(1):87–97.10.1109/42.668698
- Stonnington CM, Chu C, Kloppel S, Jack CR Jr, Ashburner J, Frackowiak RS. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *Neuro Image*. 2010; 51(4):1405–1413. [PubMed: 20347044]
- Tang S, Fan Y, Wu G, Kim M, Shen D. RABBIT: rapid alignment of brains by building intermediate templates. *Neuro Image*. 2009; 47(4):1277–1287. [PubMed: 19285145]
- Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B: Methodological*. 1996; 58(1):267–288.
- Wang, Y.; Nie, J.; Yap, P-T.; Shi, F.; Guo, L.; Shen, D. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011. Springer; 2011. Robust deformable-surface-based skull-stripping for large-scale studies; p. 635-642.
- Wee CY, Yap PT, Li W, Denny K, Browndyke JN, Potter GG, et al. Enriched white matter connectivity networks for accurate identification of MCI patients. *Neuro Image*. 2011; 54(3): 1812–1822. [PubMed: 20970508]
- Wee CY, Yap PT, Zhang D, Denny K, Browndyke JN, Potter GG, Welsh-Bohmer KA. Identification of MCI individuals using structural and functional connectivity networks. *Neuroimage*. 2012; 59(3):2045–2056. [PubMed: 22019883]
- Wolz R, Julkunen V, Koikkalainen J, Niskanen E, Zhang DP, Rueckert D, et al. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. *PLoS ONE*. 2011; 6(10):e25446. [PubMed: 22022397]
- Wu G, Qi F, Shen D. Learning-based deformable registration of MR brain images. *Medical Imaging, IEEE Transactions on*. 2006; 25(9):1145–1157.
- Xue Z, Shen D, Karacali B, Stern J, Rottenberg D, Davatzikos C. Simulating deformations of MR brain images for validation of atlas-based segmentation and registration algorithms. *Neuro Image*. 2006; 33(3):855–866. [PubMed: 16997578]

- Yang, J.; Shen, D.; Davatzikos, C.; Verma, R. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008. Springer; 2008. Diffusion tensor image registration using tensor geometry and orientation features; p. 905-913.
- Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*. 2006; 68(1):49–67.10.1111/j.1467-9868.2005.00532.x
- Zhang D, Shen D. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer’s disease. *Neuroimage*. 2012a; 59(2):895–907. [PubMed: 21992749]
- Zhang D, Shen D. Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers. *PloS one*. 2012b; 7(3):e33182. 2012. [PubMed: 22457741]
- Zhang D, Wang Y, Zhou L, Yuan H, Shen D. Multimodal classification of Alzheimer’s disease and mild cognitive impairment. *Neuro Image*. 2011; 55(3):856–867. [PubMed: 21236349]
- Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*. 2009; 37(6A):3468–3497.
- Zhou L, Wang Y, Li Y, Yap PT, Shen D. Hierarchical anatomical brain networks for MCI prediction: revisiting volumetric measures. *PLoS ONE*. 2011; 6(7):e21935. [PubMed: 21818280]
- Zhu D, Li K, Guo L, Jiang X, Zhang T, Zhang D, et al. DICCOL: dense individualized and common connectivity-based cortical landmarks. *Cerebral Cortex*. 2013; 23(4):786–800. [PubMed: 22490548]

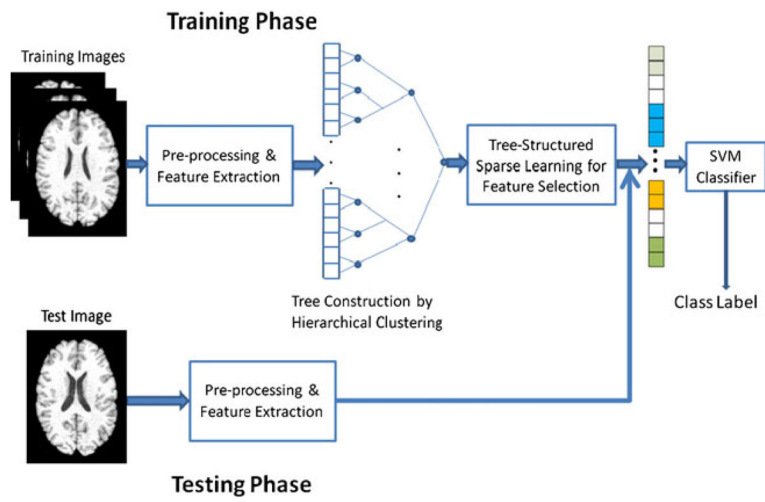


Fig. 1.
The flowchart of the proposed method

Tree Construction by Hierarchical Agglomerative Clustering

Input: $X = [x^1, \dots, x^Q] \in R^{M \times Q}$: M training images, each with Q voxels; Y : the class label vector of M images.

Output: Group sets G produced by the hierarchical clustering.

Initialization: Each voxel is initialized as one cluster: $C = \{\{q\}; q \in \{1, 2, \dots, Q\}\}$;

$G = \emptyset$;

While $|C| > 1$ do

Step 1: Find a pair of clusters $c_i, c_j \in C$ which share neighboring voxels, and minimize the measure $\Delta(i, j)$ defined in Equation (5).

Step 2: Merge these two clusters into one group set $G \leftarrow G \cup (c_i \cup c_j)$.

Step 3: Remove the clusters c_i, c_j from C : $C \leftarrow C \setminus \{c_i, c_j\}$, and add the new merged cluster into C : $C \leftarrow C \cup \{c_i \cup c_j\}$.

End while

Fig. 2.

The proposed tree construction method by hierarchical agglomerative clustering

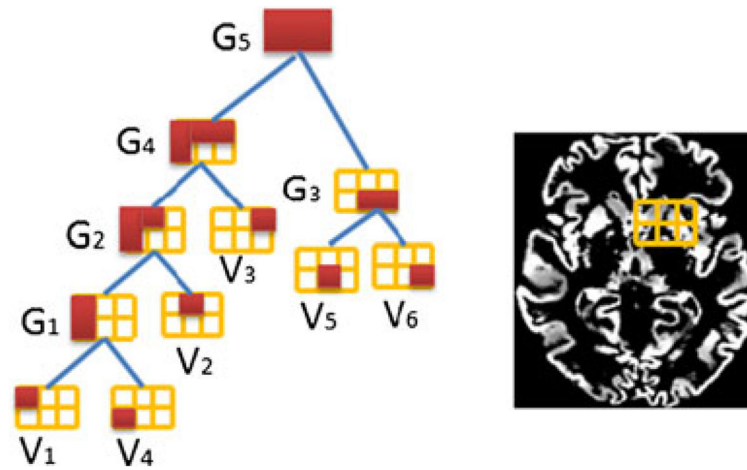


Fig. 3. A sample tree (constructed with 6 adjacent image voxels) for illustration of 6 leaves: $\{V1, V2, V3, V4, V5, V6\}$ and 5 nodes: $\{G1, G2, G3, G4, G5\}$

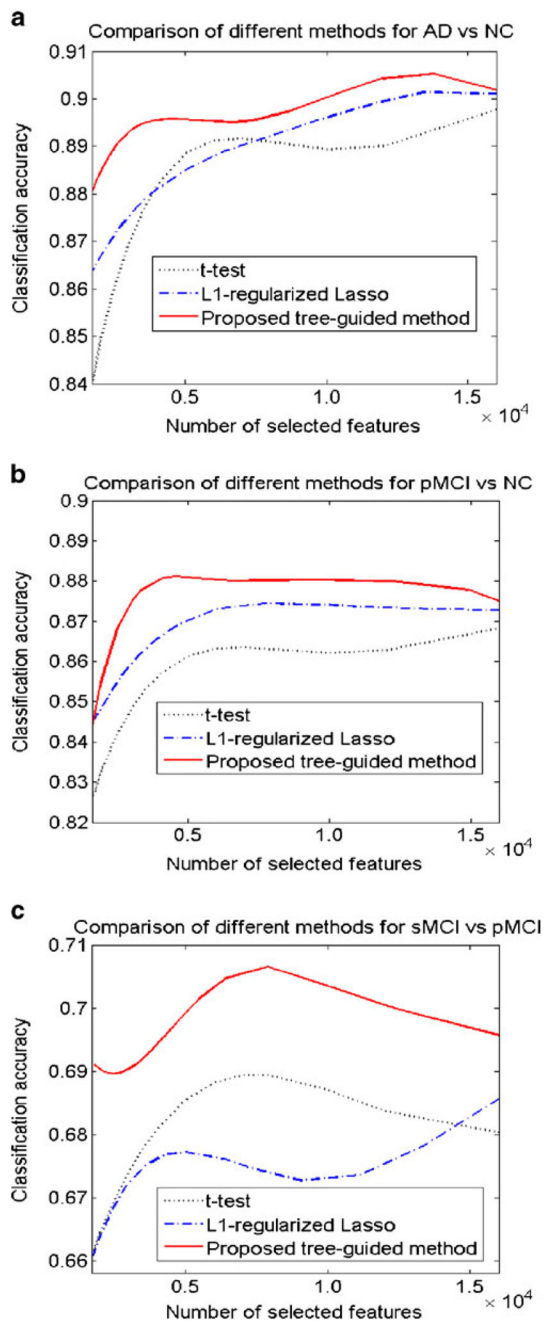


Fig. 4. Comparison of classification accuracy with respect to different number of selected features by three feature-selection methods, *t*-test, Lasso, and the proposed tree-guided method, in classification of **a** AD vs. NC, **b** pMCI vs. NC, and **c** sMCI vs. pMCI

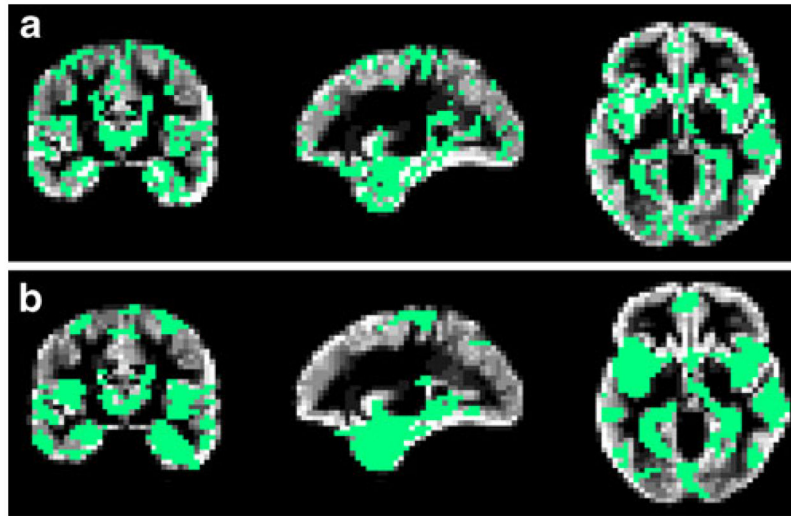


Fig. 5. The biomarkers identified from the GM density map by **a** L1-norm Lasso and **b** our proposed tree-structured sparse learning method for AD vs. NC classification

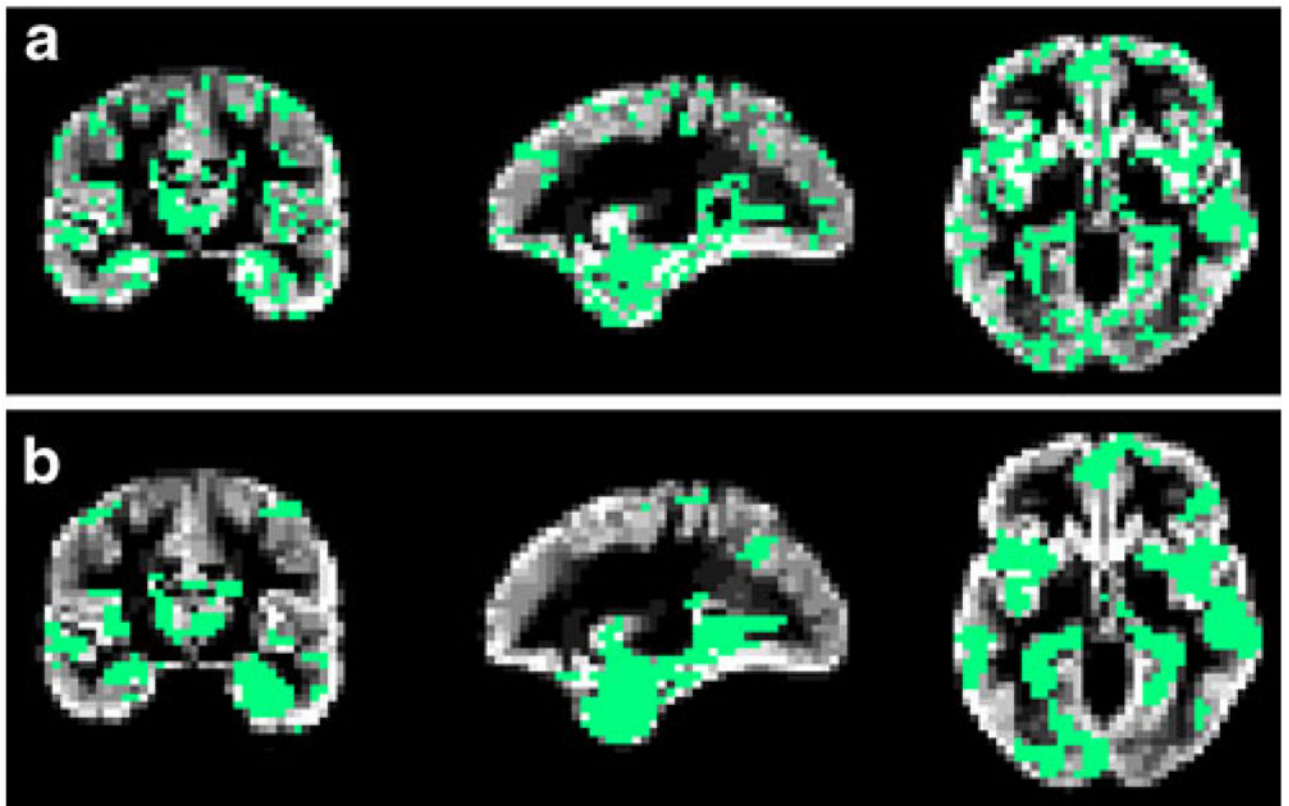


Fig. 6. The biomarkers identified from the GM density map by **a** L1-norm Lasso and **b** our proposed tree-structured sparse learning method for pMCI vs. NC classification

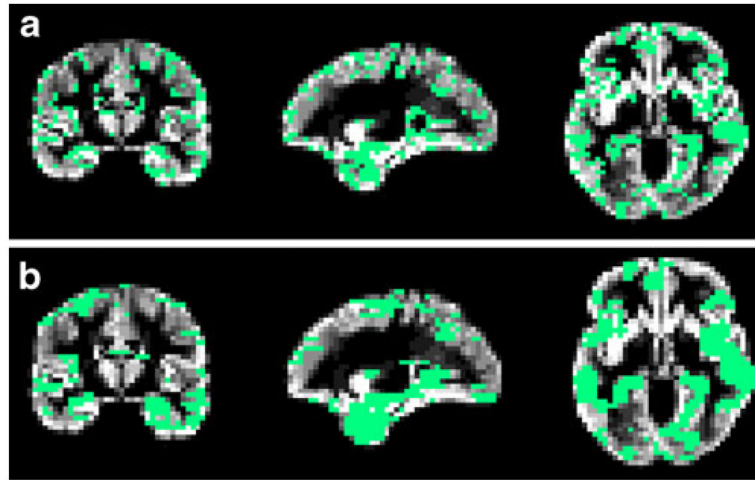


Fig. 7. The biomarkers identified from the GM density map by **a** L1-norm Lasso and **b** our proposed tree-structured sparse learning method for pMCI vs. sMCI classification

Table 1

Demographic characteristics of the studied subjects from ADNI database

Diagnosis	Number	Age	Gender (M/F)	MMSE (Mini Mental State Examination)
AD	198	75.7±7.7	103/95	23.3±2.0
pMCI	167	74.9±6.8	102/65	26.6±1.7
sMCI	236	74.9±7.7	158/78	27.3±1.8
NC	229	76.0±5.0	119/110	29.1±1.0

Table 2

Comparison of classification results by different tree construction methods

Method	Tree construction	Classes	ACC (%)	SEN (%)	SPE (%)
Baseline (whole brain)	Without feature selection	AD/NC	88.5±0.5	84.6±0.6	92.0±0.6
		pMCI/NC	86.5±0.7	80.0±1.5	91.2±1.0
		pMCI/sMCI	69.8±1.5	56.4±2.4	79.4±1.5
Patch-based	Defining tree nodes by multi-scale patches	AD/NC	89.6±0.5	84.9±0.9	93.5±0.8
		pMCI/NC	86.5±0.7	80.0±1.1	91.3±1.0
		pMCI/sMCI	69.6±1.0	56.5±1.5	78.8±1.4
Cluster-based	Hierarchical clustering on feature similarity	AD/NC	89.7±0.4	85.2±0.7	93.6±0.5
		pMCI/NC	86.3±0.6	79.4±1.2	91.3±0.8
		pMCI/sMCI	69.1±1.3	55.3±2.0	78.9±1.5
ROI-based	Defining tree nodes by 93 ROIs	AD/NC	89.5±0.7	84.9±1.1	93.6±0.5
		pMCI/NC	86.1±0.6	79.5±1.1	90.9±0.9
		pMCI/sMCI	69.5±1.4	56.4±2.2	79.5±1.8
Our proposed method	Defining tree nodes by hierarchical clustering	AD/NC	90.2±0.5	85.3±0.6	94.3±0.4
		pMCI/NC	87.2±0.6	80.1±1.1	92.2±0.3
		pMCI/sMCI	70.7±0.5	56.2±1.1	80.9±0.8

Table 3

Comparison of classification results on the NC, sMCI, pMCI, and AD subjects, with the baseline MRI data in ADNI database

Methods	Features	Classifier	Subjects	ACC (%)	SEN (%)	SPE (%)
(Hinrichs et al. 2009)	Voxel-wise GM features	(LP) boosting	183 (AD + NC)	82.0	85.0	80.0
(Zhang et al. 2011)	GM volumes of 93 ROIs	SVM	51 AD + 52 NC	86.2	86.0	86.3
(Cuingnet et al. 2011)	Voxel-wise GM features	SVM	99 MCI + 52 NC	72.0	78.5	59.6
			137 AD + 162 NC	88.6	81.0	95.0
			76 pMCI + 162 NC	81.2	73.0	85.0
			76 pMCI + 134 sMCI	66.5	62.0	69.0
(Chu et al. 2012)	Voxel-wise GM features	SVM	131 AD + 188 NC	85.0	–	–
			261 MCI + 188 NC	70.0	–	–
			261 (pMCI + sMCI)	65.0	–	–
(Wolz et al. 2011)	4 types of MRI features ^a	LDA	198 AD + 231 NC	89.0	85.0	93.0
			167 pMCI + 231 NC	84.0	82.0	86.0
			167 pMCI + 238 sMCI	68.0	67.0	69.0
Our proposed method	Voxel-wise GM features	SVM	198 AD + 229 NC	90.2	85.3	94.3
			167 pMCI + 229 NC	87.2	80.1	92.2
			167 pMCI + 236 sMCI	70.7	56.2	80.9

^aFour types of MRI features include hippocampal volume, tensor-based morphometry, cortical thickness, and manifold-learning based features