Tree-Guided Sparse Coding for Brain Disease Classification*

Manhua Liu[1,2], Daoqiang Zhang[1,3], Pew-Thian Yap[1], and Dinggang Shen[1]

[1] IDEA Lab, Department of Radiology and BRIC,
University of North Carolina at Chapel Hill, USA
[2] Department of Instrument Science and Technology,
Shanghai Jiao Tong University, China
[3] Department of Computer Science and Engineering,
Nanjing University of Aeronautics and Astronautics, China
dgshen@med.unc.edu

Abstract. Neuroimage analysis based on machine learning technologies has been widely employed to assist the diagnosis of brain diseases such as Alzheimer's disease and its prodromal stage - mild cognitive impairment. One of the major problems in brain image analysis involves learning the most relevant features from a huge set of raw imaging features, which are far more numerous than the training samples. This makes the tasks of both disease classification and interpretation extremely challenging. Sparse coding via L1 regularization, such as Lasso, can provide an effective way to select the most relevant features for alleviating the curse of dimensionality and achieving more accurate classification. However, the selected features may distribute randomly throughout the whole brain, although in reality disease-induced abnormal changes often happen in a few contiguous regions. To address this issue, we investigate a tree-guided sparse coding method to identify grouped imaging features in the brain regions for guiding disease classification and interpretation. Spatial relationships of the image structures are imposed during sparse coding with a tree-guided regularization. Our experimental results on the ADNI dataset show that the tree-guided sparse coding method not only achieves better classification accuracy, but also allows for more meaningful diagnosis of brain diseases compared with the conventional L1-regularized Lasso.

1 Introduction

Neuroimaging data, such as magnetic resonance image (MRI) and fluorodeoxyglucose positron emission tomography (FDG-PET), provides a powerful in vivo tool for aiding diagnosis and monitoring of brain diseases, such as Alzheimer's disease (AD) and mild cognitive impairment (MCI) [1, 2]. Recently, many machine learning and pattern

* This work was partially supported by NIH grants EB006733, EB008374, EB009634, AG041721, and MH088520, Medical and Engineering Foundation of Shanghai Jiao Tong University (No. YG2010MS74), and NSFC grants (No. 61005024 and 60875030).

N. Ayache et al. (Eds.): MICCAI 2012, Part III, LNCS 7512, pp. 239–247, 2012.

recognition technologies, e.g., support vector machines (SVM), have been investigated for analysis of brain images to assist the diagnosis of brain diseases [1-6]. However, the original neuroimaging data of the whole brain is often of huge dimensionality, and their direct use for control-patient classification is not only computationally expensive, but also could lead to low performance since not all features are relevant to disease pathology. Thus, feature extraction and selection are necessary and important for identifying the most relevant and discriminative features for guiding classification.

Morphological analysis of brain images has been widely used to investigate the pathological changes related to the brain diseases. 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 guiding the classification [1, 7, 8]. However, this approach to anatomical parcellation may not adapt well to the diseased-related pathology since the abnormal region may be part of ROI or span over multiple ROIs. To address this issue, Fan *et al.* [9] proposed to adaptively partition the brain image into a number of most discriminative brain regions according to the similarity computed based on correlation of image features with respect to the class labels. Then, regional features were extracted for brain disease classification. In addition to significantly reduce the feature dimensionality, this method is also robust to noise and registration error. However, the extracted regional features are generally very coarse and not sensitive to small local changes, thus affecting classification performance. Although this limitation could be potentially solved by voxel-wise analysis method [10], i.e., using voxel-wise features for classification, the number of voxel-wise features from the whole brain is often 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 also cause a significant drop in performance for high-dimensional classification methods, such as support vector machines (SVM) [11]. Therefore, it is important to significantly reduce the number of voxel-wise features before performing classification.

So far, many feature reduction and selection techniques have been proposed to select a small number of discriminative features for brain classification. Principal Component Analysis (PCA) is one of the popular methods to reduce the feature space to the most discriminant components [12]. It performs a linear transformation of the data to a lower dimensional feature space for maximization of data variance, and thus cannot always detect features from those localized abnormal brain regions. Another popular method is to select the most discriminative features and eliminate the redundant features in terms of the correlations of the individual features to the group difference such as *t*-test [3]. However, this selection method does not consider the relationships of imaging features, thus limiting its ability to detect the complex population difference.

Recently, L1-regularized sparse coding methods, e.g., Lasso, was proposed and used to sparsely identify a small subset of input features to best represent the outputs [13], and promising results were obtained. However, the selected features by L1-regularization may distribute randomly throughout the whole brain, although in reality

the disease-induced abnormal changes often happen in a few number of contiguous brain regions, instead of isolated voxels. This makes the interpretation of classification results very difficult. Actually, spatially adjacent voxels of a brain image are usually correlated, thus the coefficients assigned to them during the L1-regularization should have similar magnitudes in order to reflect their underlying correlations. Recently, a group sparse coding (Lasso) method with the hierarchical tree-guided regularization was proposed as an extension of Lasso to consider the underlying structural information among the inputs or outputs [14, 15]. In this paper, we propose to apply this tree-guided group Lasso method to identify the relevant biomarkers with the structured sparsity from MR images for brain disease classification. The hierarchical relationships of the imaging features in the whole brain are imposed in the regularization of sparse coding by a tree structure. Our experimental results on ADNI database demonstrate that, in addition to better classify the neuroimaging data of AD and MCI, the proposed classification algorithm can also identify the structured relevant biomarkers to facilitate the interpretation of classification results.

2 Method

Assume we have *M* training brain images, with each represented by a *N*-dimensional feature vector and a respective class label. The classification problem involves selection of the most relevant features and also decoding the disease states of the images, i.e., the class labels. It is observed that there are only a few brain regions affected by the disease. Thus, sparsity can be incorporated into the learning model for feature selection and disease classification.

2.1 L1-Regularized Sparse Coding (Lasso)

Let *X* denote a *N*×*M* feature matrix with the *m*-th column corresponding to the *m*-th image's feature vector $x_m = (x_m^1, \dots, x_m^n, \dots, x_m^N)^T \in R^N$ and *y* be a class label vector of *M* images with y_m denoting the class label of the *m*-th image. A linear model can be assumed to decode the class outputs from a set of features as follows:

$$y = X\alpha + \varepsilon \tag{1}$$

where $\alpha = (\alpha_1, \dots, \alpha_n, \dots, \alpha_N)^T$ 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 *N* is large and 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 [13, 16]. The L1-regularized least square problem, i.e., Lasso, can be formulated as:

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

where λ is a regularization parameter that controls the amount of sparsity in the solution. The non-zero elements of α indicate that the corresponding input features are relevant to the class labels.

The L1-regularized sparse coding provides an effective way to select a small subset of features by taking into account the correlations of individual features to the class labels. However, the structural relationships among the features, which are an important source of information, are ignored in this method. In some situations, the associated features should be jointly selected to identify the complex population difference. For example, the disease-induced abnormal changes often happen in the contiguous regions of brain image, instead of isolated voxels.

2.2 Tree-Guided Sparse Coding

To reduce the feature dimensionality while taking into account the structural relationships among the features, group Lasso was proposed as an extension of Lasso to use the groups of features instead of individual features as the units of feature selection [14]. In the regularization of sparse coding, group Lasso applies L1-norm penalty over the feature groups and L2-norm penalty for the features within each group. It assumes that the groupings of features are available as prior knowledge. However, in practice, a prior knowledge about the structures and relationships among the brain imaging features is not always available. In many applications, the features can be naturally represented using a tree structure to reflect their hierarchical spatial relationships. A tree-guided group lasso was proposed for multi-task learning where multiple related tasks follow a tree structure [14].

The brain image shows spatial correlations between the neighboring voxels, forming groups of different sizes and shapes. In this work, we propose to apply a tree structure to represent the hierarchical spatial relationships of brain image structure, with leaf nodes as the imaging features and internal nodes as the groups of features. A regularization predefined by the tree structure can be imposed on the sparse coding optimization problem to encourage a joint selection of structured relevant features. Fig. 1 shows a hierarchical tree structure imposed on a sample brain image. Assume that an index hierarchical tree T of d depth levels with $T_i = \{G_1^i, \ldots, G_j^i, \ldots, G_{n_i}^i\}$ containing n_i nodes in the ith level, $0 \le i \le d$. The different depth levels indicate the variant scales of feature groups. The index sets of the nodes in the same level have no overlapping while the index sets of a child node is a subset of its parent node. The tree-guided group Lasso (sparse coding) method can be formulated as:

$$\alpha = arg\min_{\alpha} \|y - X\alpha\|^2 + \lambda \sum_{i=0}^{d} \sum_{j=1}^{n_i} w_j^i \left\| \alpha_{G_j^i} \right\|_2 \tag{3}$$

where $\alpha_{G_j^i}$ is the set of coefficients assigned to the features within node G_j^i, w_j^i is a predefined weight for node G_j^i and is usually set to be proportional to the square root of the group size, and the number of depth levels d is set to 3 as in Fig.1. The features with non-zero coefficients are finally selected for further classification.

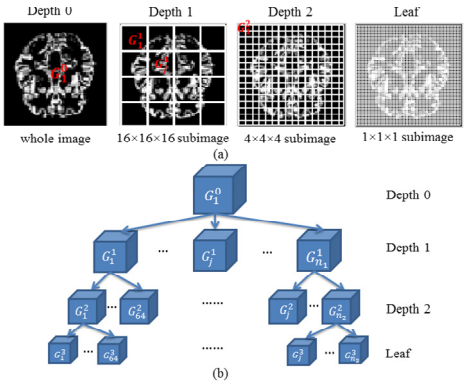


Fig. 1. Illustration of the tree structure using 2D slice as an example: (a) the subimages in different levels of tree and (b) the hierarchical tree nodes and leaves

2.3 Classification

Based on the selected imaging features by the tree-guided sparse coding method, a classifier model will be trained to make the final classification. There are various classifier models investigated for classification of brain diseases. Among them, SVM is one of the widely used classifiers because of its high classification performance [1, 7, 9, 12]. SVM constructs a maximal margin classifier in a high-dimensional feature space by mapping the original features using a kernel-induced mapping function. We choose the SVM model with a linear kernel and implement it using MATLAB SVM toolbox and the default parameters to train a classifier with the selected features for classification.

3 Experiments

We evaluate the proposed classification algorithm with the T1-weighted baseline MR brain images of 643 subjects, which include 196 AD patients, 220 MCI subjects, and 227 normal controls (NC), randomly selected from Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Table 1 provides a summary of the demographic characteristics of the studied subjects (denoted as mean ± standard deviation). Before performing classification, the image preprocessing was performed as follows. All MR brain images were first skull-stripped and cerebellum-removed after a correction of intensity inhomogeneity [17]. Then, each image was segmented into three brain tissues, i.e., gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), which were spatially normalized onto a standard space by a mass-preserving deformable registration algorithm [18]. The spatially normalized tissues are called as tissue densities 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 for the purpose of saving the computational time.

Table 1. Demographic characteristics of the studied subjects from ADNI database

Diagnosis	Number	Age	Gender (M/F)	MMSE
AD	196	75.6±7.7	103/93	23.2±2.0
MCI	220	75.0±7.4	150/70	26.8±1.8
NC	227	76.1±5.0	117/110	29.0±1.0

In the experiments, we only use the GM density map as the imaging features because it is more relevant to AD and MCI. The tree-guided group Lasso is implemented using SLEP (http://www.public.asu.edu/~jye02/Software/SLEP). The proposed algorithm is performed to classify AD vs NC and MCI vs NC. To evaluate the classification performance, 10-folds cross-validations were performed to compute the classification accuracy, i.e., the proportion of correctly classified subjects among the test dataset. In addition, we also test the L1-regularized sparse coding (Lasso) on the same dataset for comparison. Specifically, the classification accuracies are compared with respect to different number of the selected features by sparse coding. In the experiments, we change the regularization parameter $\lambda \in [0\ 1]$ to adjust the sparsity and obtain different number of selected features, with the increasing of λ leading to a smaller number of features. The classification results for AD vs NC and MCI vs NC are shown in Fig. 2. From these results, we can see that the tree-guided group Lasso can achieve better classification accuracy than the L1-regularized Lasso when the number of selected features is small ($<1.4 \times 10^4$, i.e., less than half of available features). The classification results have no large difference for both methods when the number of selected features is large. This indicates that further increasing the number of selected features will reduce the effect of structure constraint of sparse coding in classification. Also, the low computation time (i.e., needing less than 3 seconds for feature selection) makes it feasible for grouping of effective voxel-wise features.

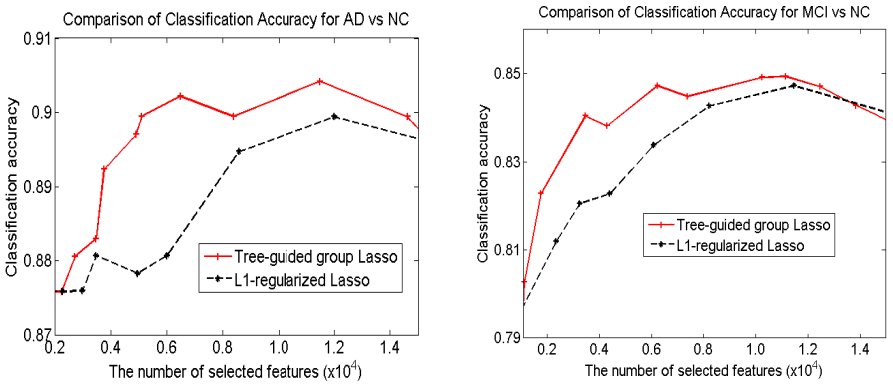
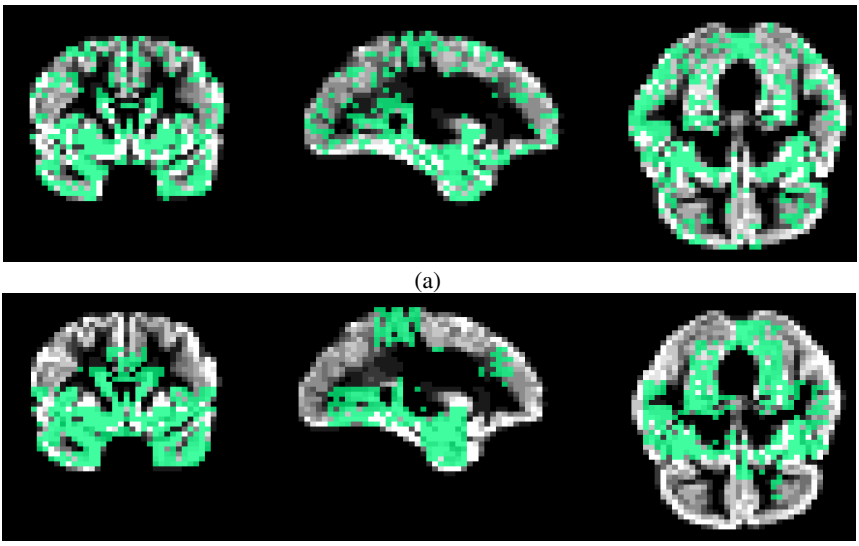


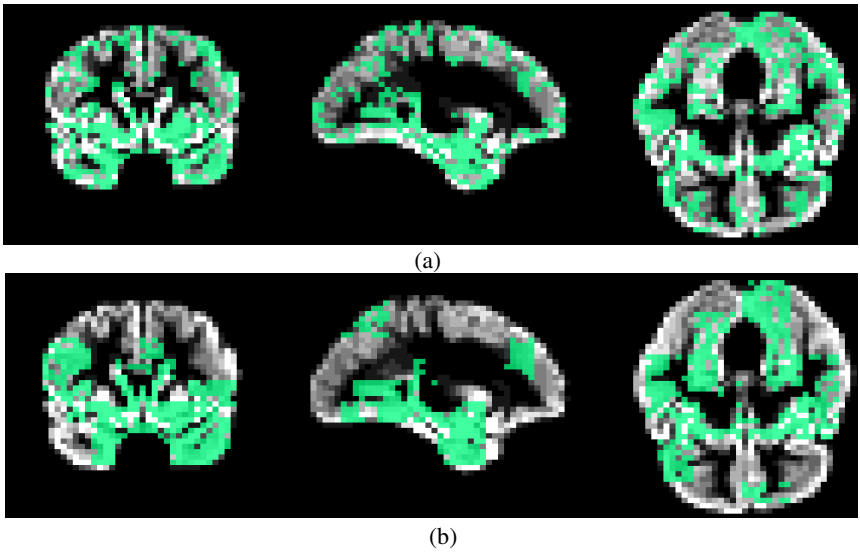
Fig. 2. Comparison of classification accuracies by the tree-guided group Lasso and L1-regularized Lasso, with respect to different number of selected features



(a)

(b)

Fig. 3. The imaging features identified from the GM density map by (a) L1-regularized Lasso and (b) tree-guided group Lasso, for the case of classifying AD vs NC



(a)

(b)

Fig. 4. The imaging features identified from the GM density map by (a) L1-regularized Lasso and (b) tree-guided group Lasso, for the case of classifying MCI vs NC

For interpretations, we show the selected image features by both the L1-regularized and tree-guided Lasso methods, with their own best regularization parameters, in Fig. 3 and 4 for AD vs NC and MCI vs NC classifications, respectively. We can see that the spatial overlaps between the L1-regularized and tree-guided Lasso methods are usually at the most relevant regions such as hippocampus, entorhinal cortex, and parahippocampal gyrus. But the features selected by L1-regularized Lasso are irregularly distributed throughout the whole brain, while the features selected by tree-guided Lasso are usually grouped at the relevant regions which are able to facilitate the interpretation of the obtained results. We evaluated that the resulting regions identified by tree-guided Lasso include hippocampus, entorhinal cortex, parahippocampal gyrus, and amygdala, which are consistent with those reported in the literature for AD and MCI studies [4, 5, 7]. These results verify the effectiveness of the tree-guided sparse coding method in incorporating the spatial structure and relationships of imaging features for guiding the disease classification and also identification of grouped relevant features.

4 Conclusion

In this paper, a sparse coding method with a tree-guided regularization is investigated to sparsely identify the grouped relevant biomarkers for brain disease classification. The tree-guided regularization is used to capture the hierarchical spatial relationships among the imaging features. Thus, the tree-guided sparse coding can provide an effective way to identify the meaningful biomarkers for brain disease classification and interpretation. Experimental results on ADNI dataset show that the proposed method not only identifies the grouped relevant biomarkers but also achieves better classification performance than the conventional L1-regularized Lasso method. Although we test this method only on classification of MR images for AD and MCI diagnosis, the similar idea can be extended and applied to other neuroimaging modalities for diagnosis of AD or other brain diseases.

References

1. Magnin, B., Mesrob, L., Kinkingnehun, S., Pelegrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehericy, S., Benali, H.: Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 51, 73–83 (2009)
2. Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D.P., Rueckert, D., Soininen, H., Lötjönen, J.: Multi-Method Analysis of MRI Images in Early Diagnostics of Alzheimer's Disease. PloS One 6, e25446 (2011)
3. Davatzikos, C., Fan, Y., Wu, X., Shen, D., Resnick, S.M.: Detection of Prodromal Alzheimer's Disease via Pattern Classification of MRI. Neurobiol. Aging (2006, epub)
4. Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M.K., Johnson, S.C.: Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. Neuroimage 48, 138–149 (2009)

5. Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehericy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O.: Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 56, 766–781 (2011)
6. Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack Jr., C.R.: Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies. Neuroimage 39, 1186–1197 (2008)
7. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55, 856–867 (2011)
8. Lao, Z., Shen, D., Xue, Z., Karacali, B., Resnick, S.M., Davatzikos, C.: Morphological classification of brains via high-dimensional shape transformations and machine learning methods. Neuroimage 21, 46–57 (2004)
9. Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C.: COMPARE: Classification of Morphological Patterns using Adaptive Regional Elements. IEEE Trans. Med. Imaging 26, 93–105 (2007)
10. Ishii, K., Kawachi, T., Sasaki, H., Kono, A.K., Fukuda, T., Kojima, Y., Mori, E.: 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 26, 333–340 (2005)
11. Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S.: Choosing multiple parameters for support vector machines. Mach. Learn. 46, 131–159 (2002)
12. Davatzikos, C., Resnick, S.M., Wu, X., Parmpi, P., Clark, C.M.: Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. Neuroimage 41, 1220–1227 (2008)
13. Ghosh, D., Chinnaiyan, A.M.: Classification and selection of biomarkers in genomic data using LASSO. J. Biomed. Biotechnol., 147–154 (2005)
14. Kim, S., Xing, E.P.: Tree-guided group lasso for multi-task regression with structured sparsity. Arxiv preprint arXiv:0909.1373 (2009)
15. Liu, J., Ye, J.: Moreau-Yosida regularization for grouped tree structure learning. In: Advances in Neural Information Processing Systems (2010)
16. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. Roy. Stat. Soc. B Met. 58, 267–288 (1996)
17. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17, 87–97 (1998)
18. Shen, D., Davatzikos, C.: Very high resolution morphometry using mass-preserving deformations and HAMMER elastic registration. Neuroimage 18, 28–41 (2003)