2011

Hierarchical anatomical brain networks for MCI prediction by partial least square analysis

Luping Zhou
University of Wollongong, lupingz@uow.edu.au

Yaping Wang
University of North Carolina

Yang Li
University of North Carolina

Pew-Thian Yap
University of North Carolina

Dinggang Shen
University of North Carolina at Chapel Hill

Hierarchical anatomical brain networks for MCI prediction by partial least square analysis

Abstract

Owning to its clinical accessibility, T1-weighted MRI has been extensively studied for the prediction of mild cognitive impairment (MCI) and Alzheimer's disease (AD). The tissue volumes of GM, WM and CSF are the most commonly used measures for MCI and AD prediction. We note that disease-induced structural changes may not happen at isolated spots, but in several inter-related regions. Therefore, in this paper we propose to directly extract the inter-region connectivity based features for MCI prediction. This involves constructing a brain network for each subject, with each node representing an ROI and each edge representing regional interactions. This network is also built hierarchically to improve the robustness of classification. Compared with conventional methods, our approach produces a significant larger pool of features, which if improperly dealt with, will result in intractability when used for classifier training. Therefore based on the characteristics of the network features, we employ Partial Least Square analysis to efficiently reduce the feature dimensionality to a manageable level while at the same time preserving discriminative information as much as possible. Our experiment demonstrates that without requiring any new information in addition to T1-weighted images, the prediction accuracy of MCI is statistically improved. 2011 IEEE.

Keywords

mci, networks, brain, anatomical, analysis, hierarchical, square, least, partial, prediction

Disciplines

Engineering | Science and Technology Studies

Publication Details

Zhou, L., Wang, Y., Li, Y., Yap, P. & Shen, D. (2011). Hierarchical anatomical brain networks for MCI prediction by partial least square analysis. IEEE Conference on Computer Vision and Pattern Recognition (pp. 1073-1080). Providence, United States: IEEE.

Hierarchical Anatomical Brain Networks for MCI Prediction by Partial Least Square Analysis

[1]Luping Zhou, [2]Yaping Wang, [1]Yang Li, [1]Pew-Thian Yap, [1]Dinggang Shen, and ADNI
*[1] University of North Carolina at Chapel Hill, U.S.A
[2] Northwestern Polytechnical University, China

Abstract

Owning to its clinical accessibility, T1-weighted MRI has been extensively studied for the prediction of mild cognitive impairment (MCI) and Alzheimer's disease (AD). The tissue volumes of GM, WM and CSF are the most commonly used measures for MCI and AD prediction. We note that disease-induced structural changes may not happen at isolated spots, but in several inter-related regions. Therefore, in this paper we propose to directly extract the inter-region connectivity based features for MCI prediction. This involves constructing a brain network for each subject, with each node representing an ROI and each edge representing regional interactions. This network is also built hierarchically to improve the robustness of classification. Compared with conventional methods, our approach produces a significant larger pool of features, which if improperly dealt with, will result in intractability when used for classifier training. Therefore based on the characteristics of the network features, we employ Partial Least Square analysis to efficiently reduce the feature dimensionality to a manageable level while at the same time preserving discriminative information as much as possible. Our experiment demonstrates that without requiring any new information in addition to T1-weighted images, the prediction accuracy of MCI is statistically improved.

1. Introduction

As the most common neurodegenerative disease, Alzheimer's disease (AD) is a progressive and eventually fatal disease of the brain, characterized by memory failure and degeneration of other cognitive functions. Early diagnosis of AD is not easy, because the pathology may begin long before the patient experiences any symptom and often lead to volumetric or shape changes at certain brain structures. With the aid of medical imaging techniques, it is possible to study in vivo the relationship between brain structural changes and mental disorders, and further provide a diagnosis tool for early detection of AD. Current studies focus on MCI (mild cognitive impairment) subjects who are in a transitional state between normal aging and AD. Identifying MCI subjects is important, especially for those who eventually convert to AD, because they may benefit from the therapies that could possibly slow down the progression of AD when the disease is mild.

Although T1-weighted MRI has been studied for a decade, it continues to attract researchers due to its easy access in clinical practice. The neuroimaging measurements for AD detection can be categorized into three groups: regional brain volumes, cortical thickness, and hippocampal volume and shape [3]. In this paper, we are interested in regional volume analysis of the whole brain, because the abnormalities caused by MCI may not be restricted to only cortical thickness or hippocampus. The affected regions could be the entorhinal cortex, the amygdala, the limbic system, the neocortical areas and so on.

In conventional volume-based methods, the mean tissue volumes of gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) are usually calculated locally within Region of Interest (ROI), and used as features for classification. Nevertheless, disease-induced brain structural changes may not happen at isolated spots, but in several inter-related regions. The measurement of the correlations between ROIs may give possible biomarkers associated with pathology, and hence is of great research interest. However, in the conventional methods, such correlations are not explicitly modelled in the feature extraction procedure, but only implicitly considered by some classifiers, such as some nonlinear SVMs, in the classification process. The interpretation of these implicitly encoded correlations in nonlinear SVMs is a challenging problem. Based on this observation, we hypothesize that *representing the brain as a system of interconnected regions is a more effective way of characterizing subtle changes than by using local isolated measures*, and directly model the pairwise ROI interactions within a subject as features for classification. Any criterion that mea-

*Data used in this article were obtained from the Alzheimer Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI).

sures the correlations between two regions can be employed for this purpose, such as correlation coefficients and mutual information. We use the correlation coefficients in our study to simplify the problem. In particular, each ROI is characterized by a volumetric vector that consists of the volumetric ratios of GM, WM and CSF in this ROI. The interaction between two ROIs within the same subject is computed as the Pearson correlation of the corresponding volumetric elements. This gives us an anatomical brain network using the T1-weighted MRI, with each node denoting an ROI and each edge characterizing the pairwise connection. Note that the correlation value measures the similarity of the tissue compositions between a pair of brain regions. When a patient is affected by MCI, the correlation values of some brain regions with other regions will be affected, due possibly to the factors such as tissue atrophy.

By computing the pairwise correlation between ROIs, our approach provides a second order measure of the ROI volume, while the conventional approaches only employ the first order measure of the volume. As higher order measures, our new features may be more descriptive, but also more sensitive to noise, such as registration errors. Therefore, a hierarchy of multi-resolution ROIs is introduced to increase the robustness of classification. Effectively, the correlations are considered at different scales of regions, thus giving different levels of noise suppression and discriminant information, which can be sieved by the classification scheme as discussed below. This approach considers the correlations both within and between different resolution scales, because a certain "optimal" scale often cannot be known a priori.

However, the dimensionality of the network features is much higher than that of the volumetric features. Without identifying a small set of the most discriminative features, it may be intractable to train an efficient classifier. Therefore a classification scheme is proposed by employing Partial Least Square analysis to embed the original features into a much lower dimensional space as well as optimally maintaining the discrimination power of features. This approach outperforms some commonly used unsupervised and supervised methods as shown in our experiment. The most important advantage of our proposed hierarchical anatomical brain network is: without requiring any new information in addition to the T1-weighted images, the prediction accuracy of MCI is statistically improved as evaluated by the data sets randomly drawn from the ADNI dataset [7]. Our study shows that this improvement comes from the use of both regional interactions and the hierarchical structure.

The merits of our proposed method are summarized as follows. Firstly, the proposed method utilizes a second-order volumetric measure that is more descriptive than the conventional first-order volumetric measure. Secondly, while the conventional approaches only consider local volume changes, our proposed method considers global information by pairing ROIs that may be spatially far away. Thirdly, our proposed method seamlessly incorporates both the local volume features and the proposed global network features into the classification by introducing a whole-brain ROI at the top of the hierarchy. By correlating with the whole-brain ROI, each ROI can provide a first order measurement of local volume. Fourthly, the proposed method involves only linear methods, leading to easy interpretations of the classification results. Note that the nterpretation is equally important as classification in neuro-imaging analysis. Finally, for the first time, the proposed method investigates the *relative* disease progression speeds in different regions, providing a complementary perspective of the spatial atrophy patterns to conventional methods.

2. Method

The overview of our proposed method is illustrated in Fig. 1. Each brain image is parcellated in multi-resolution according to our predefined hierarchical ROIs. The local volumes of GM, WM, and CSF are measured within these ROIs and used to construct an anatomical brain network. The edge weights of the network are used for the classification. This gives rise to a large amount of features. Without efficiently removing many noisy features, the training of classifier may be intractable. Therefore, both feature selection and feature embedding algorithms are used to identify those essentially discriminative features for training classifiers which can be well generalized to predict previously unseen subjects.

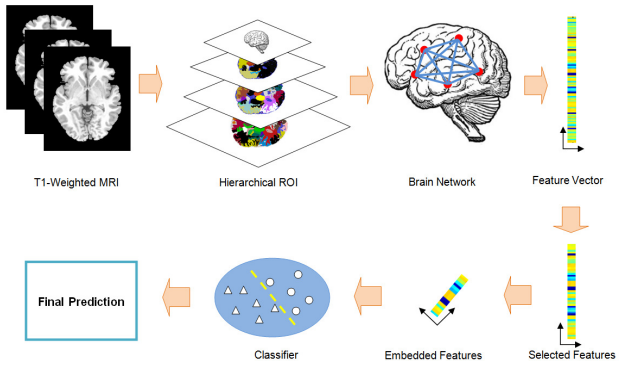

Figure 1. *Overview of our proposed method.*

2.1. Image Preprocessing

The T1-weighted MR brain images are skull-stripped and cerebellum-removed after a correction of intensity inhomogeneity. Then each MR brain image is further segmented into three tissues, namely GM, WM, and CSF. To compare structural patterns across subjects, the tissue-

segmented brain images are spatially normalized into a template space by a mass-preserving registration framework proposed in [13]. During the image warping, the tissue density within a region is increased if the region is compressed, and vice versa. After mass-preserving spatial normalization of each subject into a template space, we can measure the volumes of GM, WM, and CSF of each ROI in this subject. The definitions of hierarchical ROIs are detailed as follows.

2.2. Hierarchical ROI Construction

In this paper, a four-layer ROI hierarchy is proposed to improve the robustness of classification. Each layer corresponds to a brain atlas with different sizes of ROIs. Let us denote the bottommost layer that contains the finest ROIs as $\mathcal{L}^4$, while the other three layers are denoted as $\mathcal{L}^l$, where $l = 1, 2, 3$. A smaller l denotes a coarser ROI which is in a layer closer to the top of the hierarchy. In our approach, the bottommost layer $\mathcal{L}^4$ contains 100 ROIs obtained according to [8]. These ROIs include fine cortical and subcortical structures, ventricle system, etc. The number of ROIs reduces to 44 and 20, respectively, in the layers $\mathcal{L}^3$ and $\mathcal{L}^2$ by agglomerative merging of the 100 ROIs in the layer $\mathcal{L}^4$. In the layer $\mathcal{L}^3$, the cortical structures are grouped into frontal, parietal, occipital, temporal, limbic, and insula lobe in both left and right brain hemispheres. Each cortical ROI has three sub-ROIs, namely the superolateral, medial and white matter ROIs. The subcortical structures are merged into three groups in each hemishphere of the brain, namely, the basal ganglia, hippocampus and amygdala, and diencephalon. In the layer $\mathcal{L}^2$, the sub-groups within each cortical ROI are merged together. All the subcortical ROIs are grouped into one ROI. The topmost layer $\mathcal{L}^1$ contains only one ROI, i.e., the whole brain. This layer $\mathcal{L}^1$ is included because when correlated with the ROIs in other layers, it gives us a measurement of local volumes. In this way, the proposed method can seamlessly incorporate both the local information (obtained by correlating local ROIs with the whole brain) and the global information (obtained by correlating local ROIs with each other) for classification. The ROIs for different layers are shown in Fig. 2 (a).

2.3. Feature Extraction

With the ROI hierarchy defined above, an anatomical brain network $\mathcal{G}(\mathcal{V}, \mathcal{E})$ can be constructed for each subject. Its nodes $\mathcal{V}$ correspond to the brain ROIs, and its undirected edges $\mathcal{E}$ correspond to the interactions between two ROIs. There are two types of nodes in our model (Fig. 3-left): the simple ROI in the bottommost layer $\mathcal{L}^4$, and the compound ROI in the other layers. Similarly, we have two types of edges, each modelling the within-layer and between-layer ROI interactions, respectively (Fig. 3-right).

The brain network may be quite complicated. For instance, Fig. 2 (b) partially shows the network connections

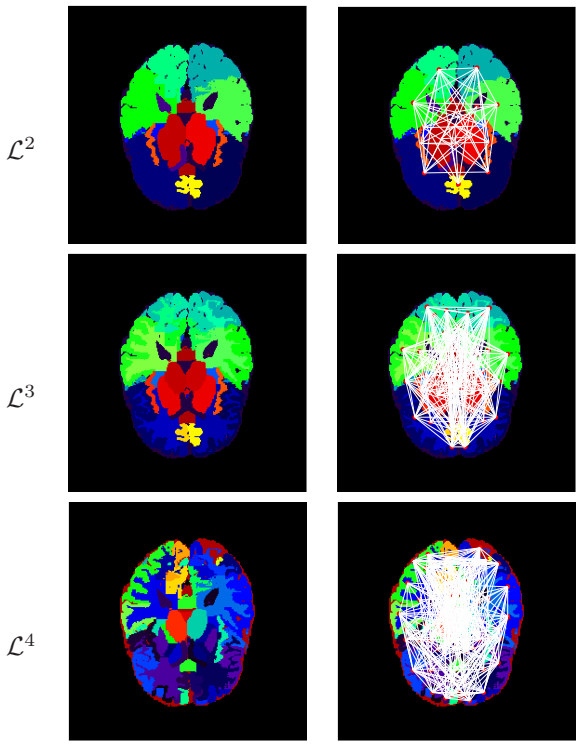

Figure 2. *(a) Hierarchical ROIs in three different layers; (b) Network connections between ROIs within different layers.*

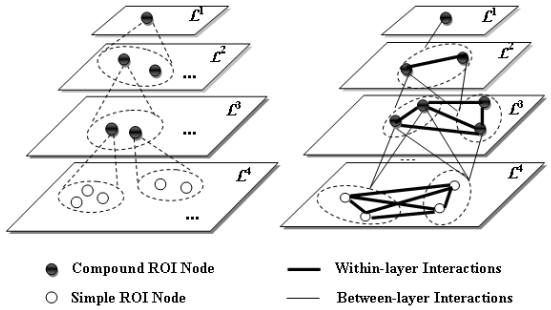

Figure 3. *Left: Two types of nodes are included in the hierarchical network: the simple node in $\mathcal{L}^4$, and the compound node in $\mathcal{L}^l (l = 1, 2, 3)$. Right: Two types of edges are included in the hierarchical network, each modeling the within-layer and between-layer interactions, respectively.*

between ROIs in the layers of $\mathcal{L}^2$, $\mathcal{L}^3$ and $\mathcal{L}^4$, respectively. To efficiently obtaining the informative network features, a membership matrix is created to indicate the relationship of ROIs from different layers. The membership matrix is computed offline: it is fixed once the hierarchical structure has been determined. For a new brain image, we only need to compute the ROI interactions on the bottommost layer $\mathcal{L}^4$, and then propagate the correlations to other layers effec-

tively via this membership matrix as shown in (1) and (2). The process is detailed as follows.

Firstly, let us consider the bottommost layer $\mathcal{L}^4$, which consists of 100 ROIs. Let $\mathbf{f}_i$ denote the 3×1 vector of the i-th ROI in $\mathcal{L}^4$, consisting of the volumetric ratios of GM, WM, and CSF in that ROI. We can obtain an $N^4 \times N^4$ matrix $\mathbf{C}^4$, where N^4 is the number of ROIs in $\mathcal{L}^4$. The (i, j)-th component in $\mathbf{C}^4$ corresponds to the weight of the edge between the i-th node and the j-th node in $\mathcal{L}^4$. We define $\mathbf{C}^4(i,j) = corr(\mathbf{f}_i, \mathbf{f}_j)$, i.e., the Pearson correlation between feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$.

For any other layer $\mathcal{L}^l$, let R_i^l represent the i-th ROI in the layer $\mathcal{L}^l$. The number of ROIs in the layer $\mathcal{L}^l$ is denoted as N^l. A membership matrix $\mathbf{M}^l$ is used to define the composition of the compound ROI R_i^l in $\mathcal{L}^l$. The matrix $\mathbf{M}^l$ has N^l rows and N^4 columns. Each row corresponds to a single compound ROI in $\mathcal{L}^l$. Each column corresponds to a single simple ROI in $\mathcal{L}^4$. The (i, j)-th component of $\mathbf{M}^l$ takes the value of either 1 or 0, indicating whether the j-th ROI in $\mathcal{L}^4$ is included in the i-th ROI in $\mathcal{L}^l$. For example, if the ROI R_i^l is composed of the simple nodes R_m^4, R_n^4 and R_t^4 in $\mathcal{L}^4$, the elements of (i, m), (i, n) and (i, t) in $\mathbf{M}^l$ are set to 1, while the others in the i-th row are set to 0. In particular, for the whole brain in $\mathcal{L}^1$, the membership matrix $\mathbf{M}^1$ is a row vector with all N^4 elements set to 1.

Within-layer ROI interaction

Given the ROI interactions in the bottommost layer $\mathcal{L}^4$, the ROI interactions within each of the higher layers are computed as follows. Let R_i^l and R_j^l represent the i-th and j-th ROIs in a certain layer $\mathcal{L}^l$. Again, a matrix $\mathbf{C}^l$ is defined similar to $\mathbf{C}^4$, but its (i, j)-th component now indicates the correlation between the compound ROIs R_i^l and R_j^l. Suppose R_i^l and R_j^l contain a and b simple ROIs, respectively. The correlation between R_i^l and R_j^l is computed as the mean value of all the correlations between a simple ROI node from R_i^l and a simple ROI node from R_j^l, that is,

$$corr(R_i^l, R_j^l) = \frac{1}{a \times b} \sum_{R_m^4 \in S_i^l} \sum_{R_n^4 \in S_j^l} corr(R_m^4, R_n^4),$$

where R_m^4 and R_n^4 represent the simple ROIs in $\mathcal{L}^4$, and S_i^l and S_j^l are two sets containing the simple nodes that comprise R_i^l and R_j^l, respectively.

Represented in the form of matrix, the correlation matrix $\mathbf{C}^l$ can be computed as follows:

$$\mathbf{C}^l(i,j) = corr(R_i^l, R_j^l) = \frac{\mathbf{1}^\top \mathbf{K}_{i,j} * . \mathbf{C}^4 \mathbf{1}}{a \times b}, \quad (1)$$

where $\mathbf{C}^l(i,j)$ denotes the (i,j)-th element in the matrix $\mathbf{C}^l$, the vector $\mathbf{1}$ is the $N^l \times 1$ vector with all elements equal to 1, the symbol $*.$ represents component-wise product of two matrices, and the $N^4 \times N^4$ matrix $\mathbf{K}_{i,j} = \mathbf{M}^l(i,\cdot)^\top \otimes \mathbf{M}^l(j,\cdot)$ is the Kronecker product of the i-th and the j-th rows in the membership matrix $\mathbf{M}^l$.

Between-layer ROI interaction

The benefits to model between-layer interactions are demonstrated by our experiment in Table 1. The correlation matrix that reflects between-layer interactions can be defined similarly to that of within-layer interactions. First, let us consider the correlation matrix for two different layers $\mathcal{L}^{l_1}$ and $\mathcal{L}^{l_2}$ (where $l_1 = 1, 2, 3$; $l_2 = 1, 2, 3$; and $l_1 \neq l_2$). It is defined as:

$$\mathbf{C}^{l_1,l_2}(i,j) = corr(R_i^{l_1}, R_j^{l_2}) = \frac{\mathbf{1}^\top \mathbf{K}_{(l_1,i),(l_2,j)} * . \mathbf{C}^4 \mathbf{1}}{a \times b}, \quad (2)$$

where $\mathbf{K}_{(l_1,i),(l_2,j)} = \mathbf{M}^{l_1}(i,\cdot)^\top \otimes \mathbf{M}^{l_2}(j,\cdot)$ is the Kronecker product of the i-th row in $\mathbf{M}^{l_1}$ and the j-th row in $\mathbf{M}^{l_2}$.

Feature vector construction

Note that the proposed brain network may not have the property of small-worldness (sparseness) as shown in DTI and fMRI networks [1], because the connections in our case are not based on functions or real neuron-connections. The dense adjacency matrix resulting from the correlation of tissue compositions implies that WM, GM and CSF fractions of many different brain regions are consistently similar. Note that the far-away region pairs can have meaningful tissue composition similarity, since distance information is not included in our framework. Some prior knowledge could be used to prune the edges if it is believed that two ROIs are independent of each other conditioned on the disease. However, we keep all the connections so that new relationships between structural changes and the disease are not left unexplored. But on the other side, some commonly used network features, such as local clustering coefficients, do not work efficiently as they do for sparse networks in DTI and fMRI. Therefore, we directly use the weights of edges as features, that is, we concatenate the elements in the upper triangle matrices of correlation matrices computed above.

2.4. Classification

When the number of predefined ROIs is large, the traditional approaches encounter the high feature dimensionality problem. Either feature selection or feature embedding has to be used to reduce data dimensionality. For example, in [4, 5], a small subset of features are selected by SVM-Recursive Feature Elimination (SVM-RFE) proposed in [6] and then fed into a nonlinear SVM with a Gaussian kernel. In [9], the volumetric features are nonlinearly embedded into a lower dimensional feature space by Laplacian Eigenmap, and then a clustering method is used to predict the AD from the normal control.

The dimensionality of network features is much larger than that of the volumetric features. For example, given only 10 discriminative ROIs, there are 45 pairwise interactions to model for just the bottommost level. So even after feature selection, there still might be many informative features left. On the other hand, since our study considers a hierarchical fully-connected brain network, each subject is represented by more than 10,000 features. Feature embedding directly on this large number of features becomes unreliable. Therefore, either feature selection or feature embedding *alone* may not be sufficient to identify the discriminative network features. In this paper, we optimally incorporate feature dimensionality reduction and classification, and propose to combine both feature selection and feature embedding in the same framework to efficiently reduce the feature dimensionality. The key point of the proposed scheme is Partial Least Square (PLS) analysis [12], which both considers the classification labels and respects the underlying data structure during dimensionality reduction. PLS especially has advantages to deal with the characteristics of our network features, where the size of the samples is much smaller than the size of the features.

Let the $n \times d$ matrix $\mathbf{X}$ represent the d-dimensional feature vectors for the n subjects, and $\mathbf{Y}$ represent the corresponding 1-dimensional label vector. PLS models the relations between $\mathbf{X}$ and $\mathbf{Y}$ by maximizing the covariance of their projections onto some latent structures. In particular, PLS decomposes the zero-mean matrix $\mathbf{X}$ and the zero-mean vector $\mathbf{Y}$ into

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^\top + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^\top + \mathbf{F} \end{aligned} \tag{3}$$

where $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_p)$ and $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_p)$ are $n \times p$ matrices containing p extracted latent vectors, the $d \times p$ matrix $\mathbf{P}$ and the $1 \times p$ vector $\mathbf{Q}$ represent the loadings, and the $n \times d$ matrix $\mathbf{E}$ and the $n \times 1$ vector $\mathbf{F}$ are the residuals. The latent matrices $\mathbf{T}$ and $\mathbf{U}$ have the following properties: each column of them, called a latent vector, is a linear combination of the original variables $\mathbf{X}$ and $\mathbf{Y}$, respectively; and the covariance of two latent vectors $\mathbf{t}_i$ and $\mathbf{u}_i$ is maximized. PLS can be solved by an iterative deflation scheme. In each iteration, the following optimization problem is solved:

$$[\mathrm{c}ov(\mathbf{t}_i, \mathbf{u}_i)]^2 = \max_{||\mathbf{w}_i||=1} [\mathrm{c}ov(\mathbf{X}\mathbf{w}_i, \mathbf{Y})]^2,$$

where $\mathbf{X}$ and $\mathbf{Y}$ are deflated by subtracting their rank-one approximations based on $\mathbf{t}_{i-1}$ and $\mathbf{u}_{i-1}$. Once the optimal weight vector $\mathbf{w}_i$ is obtained, the corresponding latent vector $\mathbf{t}_i$ can be computed by $\mathbf{t}_i = \mathbf{X}\mathbf{w}_i$.

Based on PLS analysis, our proposed method achieves good classification and generalization in four steps. The number of features selected in each step is determined by cross-validation on the training data.

In Step 1, the discriminative power of a feature is measured by its relevance to classification. The relevance is computed by the Pearson correlation between each original feature and the classification label. The larger the absolute value of the correlation, the more discriminative the feature. Roughly $200 \sim 300$ features with correlation values higher than a threshold are kept.

In Step 2, a subset of features are further selected from the result of Step 1 in order to optimize the performance of PLS embedding in Step 3. In particular, a PLS model is trained using the selected features from Step 1. Then a method called Variable Importance on Projection (VIP) [14] is used to rank these features according to their discriminative power in the learned PLS model. The discriminative power is measured by a VIP score. The higher the score, the more discriminative the feature. A VIP score for the j-th feature is

$$VIP_j = \sqrt{\frac{d \sum_{k=1}^{p} \rho_k^2 w_{jk}^2}{\sum_{k=1}^{p} \rho_k^2}},$$

where d is the number of features, p is the number of the latent vectors as defined above, w_{jk} is the j-th element in the vector $\mathbf{w}_k$, and ρ_k is the regression weight for the k-th latent variable, that is, $\rho_k = \mathbf{u}_k^\top \mathbf{t}_k$. About $60 \sim 80$ features with the top VIP scores are selected for feature embedding in the next step.

In Step 3, using the features selected in Step 2, a new PLS model is trained to find an embedding space which best preserves the discrimination of features. The embedding is performed by projecting the feature vectors in the matrix $\mathbf{X}$ onto the new weight vectors $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_p)$ learned by PLS analysis. In other words, the representation of each subject changes from a row in the feature matrix $\mathbf{X}$ to a row in the latent matrix $\mathbf{T}$. The feature dimensionality is therefore reduced from d to p ($p \ll d$).

In Step 4, after PLS embedding, a small number of features ($4 \sim 5$ components) in the new space are able to capture the majority of the class discrimination. This greatly reduces the complexity of relationships between data. Therefore, a linear SVM can achieve better or at least comparable classification accuracies as a non-linear SVM, as shown in the experiment in Section 3.2.

The advantages of PLS for our network features over some commonly used unsupervised and supervised nonlinear methods, such as Laplacian eigenmap embedding and Kernel Fisher Discriminant Analysis (KFDA), have been evidently shown in our experiment in Section 3.2.

3. Results

Our experiment involves 125 normal control subjects and 100 MCI subjects randomly drawn from the ADNI dataset. Two kinds of comparisons are conducted, that is, to compare the discrimination power of the network and the volumetric features, and to compare the performance of different classifiers for the network features. The discussion of the classification results are given at the end of this section.

3.1. Comparison of Features

Firstly, we compare the efficacy of different features with respect to classification. The data set is randomly partitioned into 20 training and test groups, each with 150 samples for training and 75 samples for test. For a fair comparison, our proposed classification process is applied similarly to both the volumetric and the network features.

As aforementioned, our network features differ from the conventional volumetric features in two aspects: i) the network features model the regional interactions; ii) the network features are obtained from a four-layer hierarchy of brain atlases. To investigate the contribution of these two aspects, five methods are tested in the experiment: i) FN: the proposed method in this paper, using the four-layer hierarchical network features; ii) SN: using only the network features from the bottommost layer $\mathcal{L}^4$; iii) FN-NC: using the network features from all the four layers, but removing the edges across different layers; iv) SV: using the volumetric features from the bottommost layer $\mathcal{L}^4$; v) FV: using volumetric measures from all four layers.

The results are summarized in Table 1. The classification accuracy is averaged across the 20 randomly partitioned training and test groups. A paired t-test is conducted between the proposed method (FN) and the other four methods, respectively, to demonstrate the advantage of our proposed method. The p-value of the paired t-test is also reported. It can be seen from Table 1 that the proposed method (FN) is always statistically better (at the significance level of 0.05) than any of the other four methods.

Table 1. Comparison of discrimination efficacy of features

	Mean Test Accuracy (%)	Paired t-test p-value
FN	85.07	-
SN	83.00	0.00272
FN-NC	83.13	0.00367
SV	81.93	0.00166
FV	81.47	0.00015

From Table 1, we observe the following:

- Our proposed hierarchical network features in FN outperform the conventional volumetric features in SV. The advantage may come from using both regional interactions and the hierarchical structure.
- The better performance of SN over SV, and FN over FV demonstrate the benefits purely from using the regional interactions. It can be seen from Table 1 that the hierarchical structure does not improve the discrimination of volumetric features in FV.
- The better performance of FN over SN demonstrates the benefit purely from the hierarchy. The advantage of the four-layer structure is statistically significant over the single-layer. Moreover, the result that FN statistically outperforms FN-NC indicates the necessity of using the cross-layer edges in the network.

It is noticed that different ratios of training and test partitions may lead to a variation in the classification accuracy. To reflect the influence of this factor, we test seven different numbers of training samples, occupying 50% to 80% of the total data size. For each number of training samples, 20 training and test groups are randomly generated and the averaged classification accuracy is summarized in Fig. 4. The classification accuracy goes up slightly in general when the number of the training samples increases, because the larger the number of training samples, the more the learned information. It can be seen that the network features show a consistent improvement in classification accuracy of approximately 3% in all cases, compared to those by using the conventional volumetric features. Averaged across different numbers of training samples, the classification accuracy becomes 84.35% for the network features, and 80.83% for the volumetric features, which represents an overall classification performance of these two different types of features. A paired t-test is performed on the seven different ratios of training-test partitions using both features. The obtained p-value of 0.000024 indicates that the improvement of the proposed features is statistically significant.

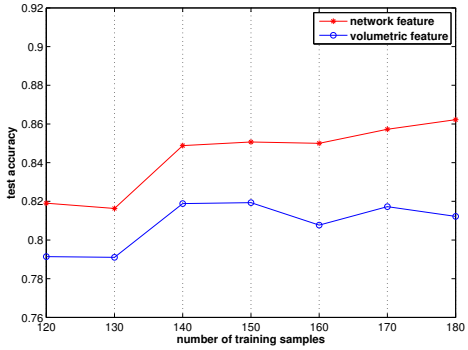

Figure 4. *Classification comparison using network features and volumetric features with different numbers of training samples.*

3.2. Comparison of Classifiers

The classification performance of our proposed classification scheme is compared with other six possible schemes shown in Table 2. To simplify the description, our proposed scheme is denoted as P1, while the other six schemes in comparison are denoted as P2 $\sim$ P7. To keep consistent with P1, each of the six schemes P2 $\sim$ P7 is also divided into four steps: rough feature selection, refined feature selection, feature embedding and classification, corresponding to Step 1 $\sim$ Step 4 in P1. Please note that the first step, rough feature selection, is the same for all schemes P1 $\sim$ P7. In this step, the discriminative features are selected by their correlations with respect to the classification labels. From the second step onwards, different schemes utilize different configurations of strategies as shown in the second column of Table 2.

To clarify the settings of our experiment, the Laplacian Eigenmap (LE) embedding used in P6 is described as follows. The embedding is applied on a connection graph that shows the neighboring relationship of the subjects. Based on the connection graph, the distance between two subjects is computed as the shortest distance between the corresponding two nodes in the graph. This distance is used to construct the adjacent matrix and Laplacian matrix used in the LE embedding.

The classification results are summarized in Fig. 5 and Table 2. Note that the classification accuracy at each number of training samples in Fig. 5 is an average over 20 random training and test partitions as mentioned in Section 3.1. Also, the overall classification accuracy in Table 2 is an average of accuracies at different numbers of training samples in Fig. 5. The best overall classification accuracy of 84.35% is obtained by our proposed scheme P1: VIP selection + PLS embedding + a linear SVM. This is slightly better than P2, where a nonlinear SVM is used. It can be seen that the classification schemes with PLS embedding (P1 $\sim$ P4) achieve an overall accuracy above 84%, better than those without PLS embedding (P5 $\sim$ P7). The supervised embedding methods, i.e., PLS (P1 $\sim$ P4) and KFDA (P7), perform better than the unsupervised Laplacian Eigenmap embedding (P6). Moreover, PLS embedding (P1 $\sim$ P4) preserves more discrimination than the nonlinear supervised embedding of KFDA (P7).

3.3. Spatial Patterns

To get understanding on the regions affected by the disease, we investigate the network features selected by the two-step feature selection process in the proposed method. Note that each network feature characterizes the relationship between two ROIs, instead of an individual ROI as in the conventional approaches. Therefore, for the first time, we study the *relative* progression speed of the disease in different ROIs of the same subject, which eliminates the impact of personal variations. On the contrary, the conventional methods study the *absolute* progression speeds of ROIs among different subjects. Normalizing subjects by the whole brain volume in conventional methods may not completely remove the personal variations.

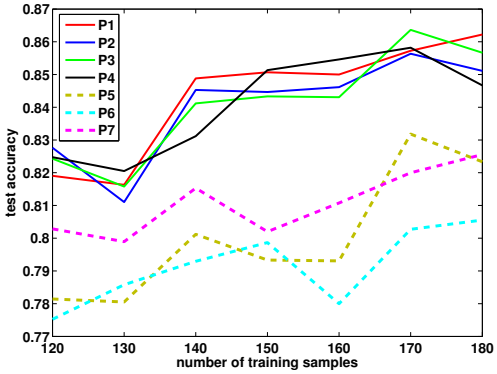

Figure 5. *Comparison of seven classification schemes on network features. The classification accuracy at each number of training samples is averaged over 20 randomly partitioned training and test groups. The scheme configurations are shown in Table 2.*

To be a essentially discriminative network feature, the two associated ROIs may satisfy one of the two following conditions: i) One ROI shows significant difference between the MCI group and the normal control group, while the other ROI is relatively constant with respect to the disease; or ii) both ROIs change with the disease, but their change speeds are different over two different groups.

Table 3 shows the most discriminative features selected by more than half of the training and test groups. It can be clearly seen that hippocampus remains the most discriminative ROI in differentiating the normal controls and MCI patients. Table 3 is separated into two parts. On the upper portion of the table, the two ROIs of a network feature may be both associated with the MCI diagnosis, such as hippocampus, entorhinal cortex, fornix, cingulate etc, as reported in the literature [11, 5, 3]. A typical example is the correlation between hippocampus and ventricle. It is known that the enlargement of ventricle is a biomarker for the diagnosis of the AD [10]. However, different from the hippocampus volume loss that often occurs at the very early stage of the dementia, the ventricle enlargement often appears in the middle and late stages. Therefore, the different progression patterns makes the correlation between the two regions the discriminative feature. On the lower portion of the table, the first ROI is associated with the disease, while the second ROI is not. For example, it has been reported that the anterior and posterior limbs of internal capsule and the occipital lobe white matter are not significantly different between MCI and normal controls in a DTI study [2].

Table 2. Configurations of classification Schemes

Schemes	Configurations	classification accuracy overall (%)
P1	VIP selection + PLS embedding + linear SVM	84.35
P2	VIP selection + PLS embedding + nonlinear SVM	84.03
P3	no selection + PLS embedding + linear SVM	84.11
P4	no selection + PLS embedding + nonlinear SVM	84.10
P5	SVM-RFE selection + no embedding + nonlinear SVM	80.07
P6	no selection + Laplacian Eigenmap embedding + nonlinear SVM	79.16
P7	no selection + KFDA embedding + linear SVM	81.08

Table 3. Selected discriminative features

hippocampus - amygdala
hippocampus - lingual gyrus
hippocampus - uncus
hippocampus - prefrontal/superolateral frontal lobe
hippocampus - globus palladus
hippocampus - entorhinal cortex
hippocampus - cingulate region
hippocampus - ventricle
hippocampus and amygdala and fornix - ventricle
uncus - fornix
hippocampus - posterior limb of internal capsule
globus palladus - anterior limb of internal capsule
hippocampus - occipital lobe WM

4. Conclusion

In this paper, we have presented how hierarchical anatomical brain networks based on T1-weighted MRI can be used to model brain regional interactions. Features extracted from these networks are employed to improve the prediction of MCI from the conventional volumetric measures. The discrimination of the network features is effectively learned by our proposed framework that addresses the properties of these new features. Without requiring new sources of information, our experiments show that the improvement of our proposed approach is statistically significant compared with the conventional volumetric measures. Such an improvement comes from both the network features and the hierarchical structure. Moreover, the selected network features provide us a new perspective of inspecting the discriminative regions of the dementia by revealing the relationship of two ROIs, which is different from the conventional approaches.

References

[1] D. Bassett and E. Bullmore. Small-world brain networks. *Neuroscientist*, 12:512–523, 2006.

[2] M. Bozzali, A. Falini, M. Franceschi, and et.,al. White matter damage in alzheimer's disease assessed in vivo using diffusion tensor magnetic resonance imaging. *Journal of Neurol Neurosurg Psychiatry*, 72:742–746, 2002.

[3] R. Cuingnet, E. Gerardin, J. Tessieras, and et.,al. Automatic classification of patients with alzheimer's disease from structural mri: A comparison of ten methods using the adni database. *Neuroimage*, 2010.

[4] Y. Fan, D. Shen, and C. Davatzikos. Classification of structural images via highdimensional image warping, robust feature extraction, and svm. In *Proceedings of MICCAI*, pages 1–8, 2005.

[5] Y. Fan, D. Shen, R. Gur, and C. Davatzikosa. Compare: classification of morphological patterns using adaptive regional elements. *IEEE Trans. Med. Imaging*, 26(1):93–105, 2007.

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–442, 2002.

[7] C. Jack Jr., M. Bernstein, N. Fox, and et.,al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *J. Magn. Reson. Imaging*, 27(4):685–691, 2008.

[8] N. Kabani, J. MacDonald, C. Holmes, and A. Evans. A 3d atlas of the human brain. *NeuroImage*, 7:S7–S17, 1998.

[9] X. Long and C. Wyatt. An automatic unsupervised classification of mr images in alzheimers disease. In *Proceedings of CVPR*, 2010.

[10] S. Nestor, R. Rupsingh, M. Borrie, and et.,al. Ventricular enlargement as a possible measure of alzheimer's disease progression validated using the alzheimer's disease neuroimaging initiative database. *Brain*, 131(9):2443–2454, 2008.

[11] G. Pengas, J. Hodges, P. Watson, and P. Nestor. Focal posterior cingulate atrophy in incipient alzheimer's disease. *Neurobiol Aging.*, 31(1):25–33, 2010.

[12] R. Rosipal and N. Kramer. Overview and recent advances in partial least squares. *Lecture Notes in Computer Science*, 3940:34–51, 2006.

[13] D. Shen and C. Davatzikos. Very high resolution morphometry using mass-preserving deformations and hammer elastic registration. *NeuroImage*, 18(1):28–41, 2003.

[14] S. Wold, W. Johansson, and M. Cocchi. Pls - partial least- squares projections to latent structures. *3D QSAR in Drug Design: Volume 1: Theory Methods and Applications*, 1:523–550, 1993.