Published in final edited form as:
Neuroimage. 2012 April 2; 60(2): 1106–1116. doi:10.1016/j.neuroimage.2012.01.055.

Ensemble Sparse Classification of Alzheimer's Disease

Manhua Liu[1,2], **Daoqiang Zhang**[1], **Dinggang Shen**[1,*], and **the Alzheimer's Disease Neuroimaging Initiative**

Manhua Liu: mhliu01@med.unc.edu; Daoqiang Zhang: zhangd@med.unc.edu; Dinggang Shen: dgshen@med.unc.edu

[1]Department of Radiology and BRIC, University of North Carolina at Chapel Hill, NC 27599, U.S.A

[2]Department of Instrument Science and Engineering, SEIEE, Shanghai Jiao Tong University, Shanghai, China

Abstract

The high-dimensional pattern classification methods, e.g., support vector machines (SVM), have been widely investigated for analysis of structural and functional brain images (such as magnetic resonance imaging (MRI)) to assist the diagnosis of Alzheimer's disease (AD) including its prodromal stage, i.e., mild cognitive impairment (MCI). Most existing classification methods extract features from neuroimaging data and then construct a single classifier to perform classification. However, due to noise and small sample size of neuroimaging data, it is challenging to train only a global classifier that can be robust enough to achieve good classification performance. In this paper, instead of building a single global classifier, we propose a local patch-based subspace ensemble method which builds multiple individual classifiers based on different subsets of local patches and then combines them for more accurate and robust classification. Specifically, to capture the local spatial consistency, each brain image is partitioned into a number of local patches and a subset of patches is randomly selected from the patch pool to build a weak classifier. Here, the sparse representation-based classification (SRC) method, which has shown effective for classification of image data (e.g., face), is used to construct each weak classifier. Then, multiple weak classifiers are combined to make the final decision. We evaluate our method on 652 subjects (including 198 AD patients, 225 MCI and 229 normal controls) from Alzheimer's Disease Neuroimaging Initiative (ADNI) database using MR images. The experimental results show that our method achieves an accuracy of 90.8% and an area under the ROC curve (AUC) of 94.86% for AD classification and an accuracy of 87.85% and an AUC of 92.90% for MCI classification, respectively, demonstrating a very promising performance of our method compared with the state-of-the-art methods for AD/MCI classification using MR images.

Keywords

AD diagnosis; sparse representation-based classifier (SRC); random subspace ensemble; local patch

*Corresponding author: Dinggang Shen (dgshen@med.unc.edu), Fax: +1 9198432641.

1. Introduction

Alzheimer's disease (AD) is an irreversible neurodegenerative disorder that leads to progressive loss of memory and cognition function. Its early and accurate diagnosis is not only challenging but also crucial for future treatments. Structural and functional brain images such as magnetic resonance imaging (MRI) and fluorodeoxyglucose positron emission tomography (FDG-PET) are powerful imaging tools in helping understand the neural changes related to the neurodegenerative disorder of AD (Chan et al., 2003; Davatzikos et al., 2010; Fan et al., 2008; Hinrichs et al., 2009; Magnin et al., 2009; Mueller et al., 2005). Recently, many pattern recognition and machine learning techniques have been widely investigated to identify the patterns of AD-related neurodegeneration by making use of neuroimaging data (Davatzikos et al., 2006, epub; Hinrichs et al., 2009; Magnin et al., 2009). Specifically, machine learning methods can provide a useful means to recognize a pattern for a new test sample based on the information learned from the training samples. For example, the morphological information about the cortical and subcortical structures of the human brain can be measured by structural MRI and used to understand the neuroanatomical differences among different populations (Desikan et al., 2009; Thompson et al., 2002). Recently, various classification methods have also been proposed to identify individuals with AD from normal control (NC) using MRI data (Cuingnet et al., 2011; Davatzikos et al., 2007; Fan et al., 2005). In most of these existing classification methods, two main steps are generally included which are: 1) Extraction and selection of discriminative features from the original neuroimaging data, and 2) Learning of an optimal separating hyperplane in a high dimensional feature space for performing AD classification.

Since the original neuroimaging data is extremely high dimensional but with small sample size, extraction of discriminative features plays an important role in classification of AD and normal control. Voxel-wise features, such as probabilities of grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF), play an important role in neuroimaging study, and have been widely used to identify regional GM loss of AD compared to the normal aging controls (Baron et al., 2001; Ishii et al., 2005). However, voxel-wise features are of huge dimensionality, and the direct use of these features for classification is computationally expensive and can often lead to low performance due to the 'curse of dimensionality' (Duda et al., 2001). To address this critical issue, different types of feature extraction, grouping and/or selection methods have been proposed to reduce the dimensionality of feature space. For example, one popular way is to group voxels into multiple anatomical regions through the warping of a labeled atlas (Lao et al., 2004; Magnin et al., 2009). On the other hand, the brain regions can also be adaptively parcellated according to the similarity of local features (Fan et al., 2007). In this way, regional features can be extracted and the total number of features can be reduced significantly.

In addition to feature extraction, feature selection is another important technique for dimensionality reduction which selects the most discriminative features and at the same time eliminates the redundant features (Davatzikos et al., 2006, epub; Davatzikos et al., 2008; Fan et al., 2005; Vemuri et al., 2008; Yoon et al., 2007). For example, principal component analysis (PCA) is often used to reduce the feature space to the most discriminant components (Davatzikos et al., 2008; Yoon et al., 2007). However, these techniques involve a careful selection of parameters (e.g., the number of components) to preserve the important subsets of feature space. On the other hand, since the disease often affects spatially contiguous regions, instead of isolated voxels, the local spatial contiguity of the selected discriminative features (voxels) should be carefully considered during the feature selection. Thus, to capture the spatial consistency, feature selection is first performed to select the discriminative voxel-wise features and the features in the neighborhood of the selected voxels are also included for classification as done in (Vemuri et al., 2008). In (Hinrichs et

al., 2009), the spatial locality information (i.e., neighboring relationship) was reserved by enforcing the spatial regularity on the learned classifiers, thus leading to improvement in classification performance. More importantly, since the feature subset selected by many algorithms is dependent on the training data set, it may be not optimal for the test data set due to the possible data overfitting problem. Thus, instead of building a single classifier with an optimal subset of features, ensemble learning was used as a general meta-learning method to aggregate the predictions of multiple classifiers for improving the generalization ability and robustness of individual classifiers (Che et al., 2011; Hinrichs et al., 2009; Ratsch et al., 2000; Valev and Asaithambi, 2001). Some well-known ensemble learning methods such as Bagging and Boosting have been applied in analyzing neuroimaging data (Fan et al., 2008; Hinrichs et al., 2009). Besides Bagging and Boosting which construct ensemble classifiers by resampling different subsets of samples, another popular ensemble method, i.e., random subspace ensemble, randomly resamples different subsets of features to build multiple weak classifiers, and has been recently applied to analyzing functional MRI data (Ho, 1998; Kuncheva et al., 2010). This ensemble method can alleviate the possible data overfitting problem and achieve good generalization ability for the balance between optimal feature selection and potential data overfitting to a specific population.

Advances in statistical learning technologies impel to develop some high-dimensional classification algorithms that are capable of dealing with neuroimaging data. Machine learning techniques are often used to design an optimal classifier that can accurately separate a set of training samples (with known class labels) based on some optimization criteria and can also be used to classify the test samples with good generalization. So far, various classification models have been constructed for classification of different patterns between AD and normal controls. Among them, support vector machine (SVM) may be the most-widely used classifier, because of its high performance for classification of high-dimensional data (Davatzikos et al., 2008; Fan et al., 2007; Kloppel et al., 2008, epub; Magnin et al., 2009; Zhang et al., 2011). SVM is a supervised learning method which searches for the optimal margin hyperplane to maximally separate different groups. It constructs a maximal margin linear classifier in a high dimensional feature space by mapping the original features using a kernel-induced mapping function. SVM classifier is not only empirically demonstrated to be one of the most powerful pattern classification methods, but also has provided many theoretical generalization bounds to estimate its capacity. However, because SVM is based on evaluation of discrimination power for classification, it has limitation in dealing with noisy data which is the case for neuroimaging data. In addition, the discriminative features from neuroimaging data may vary across different groups of subjects and thus could lie in multiple low-dimensional subspaces of a high-dimensional feature space, which makes it difficult to build a single global classifier with high classification accuracy and robustness to noises.

On the other hand, to enhance the robustness of classification to noises, sparse representation technique, which can be regarded as one of the recent major achievements in pattern classification, has been proposed and successfully used for various classification problems, e.g., robust face recognition (Huang and Aviyente, 2007; Majumdar and Ward, 2009; Wright et al., 2009). In sparse representation-based classification, the input test sample is coded as a sparse linear combination of the training samples across all classes via L1-norm minimization, and then it evaluates which class of training samples could produce the minimum reconstruction error of the input test sample with the sparse coding coefficients. Although this technique has shown high performance for classification of high-dimensional and noisy data such as faces, to the best of our knowledge, it has not been used for AD classification.

In this paper, we will investigate using sparse representation-based classifier (denoted as SRC) proposed in (Wright et al., 2009), for accurate and robust AD/MCI classification by using MRI data. Furthermore, instead of building a single global classifier, we propose a novel local patch-based subspace ensemble method which builds multiple individual classifiers based on different subsets of local patches and then combines them for more accurate and robust classification. Different from random subspace ensemble which resamples isolated voxels, our ensemble method randomly samples local patches and thus potentially preserves the local structural information which is helpful for classification.

The main contributions of this paper can be summarized as follows: (i) Sparse representation-based classification method is proposed for AD (or MCI) classification based on MRI data. To the best of our knowledge, this kind of classification method was not previously investigated for AD classification. (ii) A random patch-based subspace ensemble framework is proposed to combine multiple weak classifiers for AD classification. This framework can effectively avoid the difficulties in selection of an optimal subset of discriminative features for the single classifier, and can also enhance the robustness of classification. In addition, compared with the random (voxel-based) subspace ensemble, our proposed random patch-based subspace ensemble considers the local spatial consistency between neighboring voxels and is thus expected to achieve better classification performance.

The rest of this paper is organized as follows. The proposed random patch-based subspace ensemble classification framework is presented in detail in Section 2. Then, in Section 3, extensive experiments and comparisons with other classification methods on the ADNI dataset are presented to demonstrate the classification accuracy and advantage of the proposed method. Finally, in Section 4, we conclude this paper and discuss the possible future directions.

2. Method

In this section we will detail the proposed random patch-based subspace ensemble classification framework with the sparse representation-based classifiers (denoted as RPSE_SRC). Although the proposed classification framework makes no assumption on a specific neuroimaging modality, for demonstrating its performance, the T1-weighted MRI (Magnetic Resonance Imaging) data, which has been extensively studied for detection of AD in the past decades, is used in this work. Specifically, the T1-weighted MR brain images are skull-stripped and cerebellum-removed after a correction of intensity inhomogeneity using nonparametric nonuniform intensity normalization (N3) algorithm (Sled et al., 1998). Then, each brain image is segmented into three kinds of tissue volumes, e.g., gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) volumes. All three tissue volumes of each brain image 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, 2003). During image warping, the tissue volume within any size of region is preserved, i.e., it is increased if the region is compressed, and vice versa. We will call the warped mass-preserving tissue volumes as the tissue density maps in this paper. These tissue density maps reflect the spatial distribution of tissues in an original brain by taking into consideration the local tissue volumes prior to warping. In this work, the spatially normalized tissue density maps of T1-weighted MRI data are used as features in classification.

It is well known that the tissue density map of brain MRI is of high dimensionality, consisting of considerably more voxels (i.e., more than hundreds of thousands) than subjects (i.e., at most hundreds). When all the brain tissue densities are used as classification

features, the high feature dimensionality will likely degrade the classification capability with direct application of standard classifier models, such as linear discriminant analysis, decision trees, and SVM. Classification of the high-dimensional features is still a challenging task due to the small number of samples and the scalability problem. One common strategy is to restrict the classification feature set to only those with significant discriminating power and then construct a supervised classifier to perform classification. However, the discriminative features from the high dimensional neuroimaging data may lie in multiple low-dimensional feature subspaces, which make it difficult to find an optimal subset of discriminative features for building a single global classifier that can achieve good classification performance for all subjects. Furthermore, it has been observed that the disease-induced structural changes may not occur at isolated voxels, but in several voxels-grouped local regions (Hinrichs et al., 2009). Thus, the spatial consistency of the features should be taken into account for more accurate classification.

To address the above issues, we develop a random patch-based subspace ensemble classification framework to combine multiple individual weak classifiers for more accurate AD classification. Specifically, the sparse representation-based classifier (SRC), which shows high performance for robust classification of imaging data, will be used to design each individual weak classifier. To capture the local spatial consistency of features, a subset of local patches will be randomly resampled to construct each individual weak classifier. Since each subset of patches defines a subspace of the whole brain feature space, each individual weak classifier can be trained more easily in smaller subspace and thus the dimensionality-to-subject ratio can be substantially improved. The accuracy of final classification can be further improved by replacing a single classifier with an ensemble of multiple classifiers.

Figure 1 shows the flow chart of our random patch-based subspace ensemble classification framework, which consists of three main steps: 1) patch extraction and random patch sampling; 2) design of individual weak classifier using the sparse representation-based classifier (SRC); 3) ensemble of multiple weak classifiers to produce more accurate classification. We will detail each step in the rest of this section.

2.1. Patch Extraction

For simplicity, we uniformly divide the tissue density maps into patches of fixed size without overlapping. For accurate classification, the noisy voxels should be first excluded from the feature subspaces. On the other hand, the sampled subspaces for the individual classifiers should be diverse to give complementary information for effective ensemble. To balance the tradeoff between accuracy and diversity, we decided to carry out a preselection of individually important voxels in the hope that the relevant voxels will be contained within the sampled patch subspaces. We perform the simple t-test on each voxel of the whole brain and select the relevant voxels with the p-value smaller than 0.05. The patch extraction is carried out based on these preselected relevant voxels. The patch pool for random sampling is composed of the patches in each of which more than 50% voxels are the preselected relevant voxels. To construct a feature subspace for a weak classifier, we randomly select a subset of patches from the patch pool and all the preselected relevant voxels contained in the sampled patches are concatenated into a feature vector for classification. Each random sampling defines a feature subspace for one weak classifier. Thus, we can perform multiple random samplings to extract different feature subspaces and construct multiple individual weak classifiers for ensemble.

2.2. Sparse Representation-based Classifier (SRC)

After randomly sampling a subset of patches from the patch pool, the tissue densities from the sampled patches are concatenated into a feature vector for representation of each training subject. We construct an independent weak classifier for each randomly sampled subspace, using a sparse representation-based classification method.

Sparse representation has been successfully used in various applications where the original signal needs to be reconstructed as accurately as possible, such as denoising (Elad and Aharon, 2006), image inpainting (Fadili et al., 2009), and coding (Hazan et al., 2005). Recently, a sparse representation-based classifier (SRC) was proposed to harness the sparsity for discrimination (Wright et al., 2009). Instead of using the sparsity to identify a relevant model or relevant features that can later be used for classifying all test samples, the SRC exploits the discriminative nature of the sparse representation for classification. The basic idea of SRC is that the test data is considered as a linear composition of the training data set belonging to the same category if sufficient training samples are available for each class. Different from some conventional classifiers that optimize the discrimination power in the objective function, SRC constructs a nonparametric dictionary using all training data set across all classes and seeks for the sparse representation of a test sample in the nonparametric dictionary. SRC first codes the test sample as a sparse linear combination of all training samples by L1-norm minimization and then performs classification by evaluating which class produces the minimum reconstruction error. It combines the sparsity and reconstruction error in sparse representation for classification. With the sparsity properly harnessed, SRC achieves high performance for classification of high-dimensional data such as face images, and also high robustness to noise such as image occlusion and corruption (Wright et al., 2009). It also boosts the research on the sparsity based pattern classification (Majumdar and Ward, 2009; Yang et al., 2011).

In this work, we investigate using the sparse representation-based classifier to construct the individual weak classifier in each feature subspace. The classification problem is formulated as finding a sparse representation of the test image with respect to all the training data set. The sparse coding can be accurately and efficiently solved by the L1-norm minimization. Then the classification of a new test sample is made by checking which class produces the least coding error with the associated sparse coefficients. The class with the best approximation by sparse representation is assigned as the output class of the test sample. Unless specially noted, all feature vectors in this paper are column vectors and represents the standard Euclidean norm while $\|\cdot\|_1$ represents the standard L1 norm. Suppose that there are N training samples represented by $\mathbf{X} = [\boldsymbol{X}^1 \ldots, \boldsymbol{X}^l \ldots, \boldsymbol{X}^C] \in \Re^{M\times N}$ belonging to C categories (classes), where $N = N_1 + \cdots N_l + \cdots N_C$ and $X^l = \left[x_1^l, \ldots, x_i^l, \ldots, x_{N_l}^l\right] \in \Re^{M\times N_l}$ consists of N_l training samples of the l-th category with each feature vector x_i^l of M dimensionality. In the this study C =2, but the proposed framework can allow to include more classes such as subjects with MCI. The classification model based on sparse representation can be summarized as (Wright et al., 2009):

1. Input: A matrix of training samples $\mathbf{X} = [\boldsymbol{X}^1 \ldots, \boldsymbol{X}^l \ldots, \boldsymbol{X}^C] \in \Re^{M\times N}$ for C classes with each column being one feature vector of a training sample, a test sample represented by one column vector $\boldsymbol{y} \in \Re^M$, and an optional error tolerance $\varepsilon>0$.
2. Normalize each column of $\mathbf{X}$ and the test sample $\boldsymbol{y}$ to have unit L2 norm.
3. Compute the decomposition coefficient vector $\hat{\alpha}$ by solving the L1-norm minimization problem by sparse coding:

$$\widehat{\alpha}=arg\min_{\alpha}\left\|\alpha\right\|_1 \quad \text{subject to}\left\|\boldsymbol{X}\alpha-\boldsymbol{y}\right\|_2 \le \varepsilon \tag{1}$$

4. For each test sample $\boldsymbol{y}$, compute the residual (i.e., the sparse reconstruction error) with the sparse coefficients $\widehat{\alpha^l}$ associated to each category/class l:

$$r_l(\boldsymbol{y})=\left\|\boldsymbol{X}^l\widehat{\alpha^l}-\boldsymbol{y}\right\|_2 \text{ for } l=1,\ldots,C \tag{2}$$

5. Output: The class label for the test sample $\boldsymbol{y}$ is assigned as the class with the minimum residual over all classes:

$$\text{Label}(\boldsymbol{y})=arg\min_l r_l(\boldsymbol{y}) \tag{3}$$

The L1-norm minimization in Equation (1) can be efficiently solved by using some L1-regularized sparse coding methods such as those proposed in (Boyd and Vandenberghe, 2004; Candes and Romberg, 2005; Chen et al., 2001; Kim et al., 2007). We can see that the classification of the test sample $\boldsymbol{y}$ depends on the residuals. There are two important terms in the above classification model. One is to characterize the signal sparsity by the L1-norm constraint $\| \boldsymbol{\alpha} \|_1$. Another one is to characterize the signal fidelity by the L2-norm term $\|\boldsymbol{X\alpha} - \boldsymbol{y}\|_2 \le \varepsilon$ especially when the test sample $\boldsymbol{y}$ is noisy. Ideally, the sparse coefficients of $\| \boldsymbol{\alpha} \|_1$ are associated with the training samples from a single class so that the test sample can be easily assigned to that class. However, noise and modeling error may also lead to small nonzero sparse coefficients associated with multiple classes. Instead of classifying test sample only based on the sparse coefficients, the classification made by Equation (3) is based on how well the sparse coefficients associated with the training data in each class can reconstruct the test sample, which can better harness the subspace structure of each class. Thus, SRC is able to effectively combine the discriminative nature of sparsity and the reconstruction power for classification. For each randomly sampled patch-based subspace, we construct a nonparametric dictionary composed of all training data samples to build a sparse representation-based classifier. Finally we can get multiple SRC classifiers based on the multiple random samplings of the feature space.

2.3. Ensemble of Weak Classifiers

The classifier ensemble is usually considered to be more accurate and robust than individual classifier. The simple majority voting is one of widely used methods for fusion of multiple classifiers. However, this method puts equal weight on the outputs of all individual weak classifiers to ensemble. In fact, the classifiers might have different classification confidences for a test sample, e.g., the test sample may be located near the decision boundary of some classifiers (low classification confidence) or far from the decision boundary of other classifiers (high classification confidence). From Equation (3), we know that the classification of a test sample is performed in terms of the residuals with respective to the C classes, which also measures the similarity between the test sample and the training data of each individual class. Smaller residual also indicates that the test sample is better approximated by the sparse representation of the training samples belonging to the corresponding class. We combine the multiple weak classifiers by using the residuals of sparse representation instead of the class label output. In this way, if the residuals of a classifier corresponding to the C classes are close to each other, the classifier will have low contribution to the final ensemble, and vice versa. Suppose that we have K weak classifiers for final ensemble. Defining $r_l^k(\boldsymbol{y})$ as the residual of the test sample $\boldsymbol{y}$ obtained from the k-th

weak classifier corresponding to the *l*-th class, then the empirical average of the *l*-th residuals over the *K* weak classifiers can be calculated as follows:

$$E_l(\boldsymbol{y}) = \frac{\sum_{k=1}^{K} r_l^k(\boldsymbol{y})}{K} \tag{4}$$

Finally, the class label of the test sample $\boldsymbol{y}$ can be assigned to the class with the minimum average residual as:

$$\text{Label}(\boldsymbol{y}) = \arg\min_l E_l(\boldsymbol{y}) \tag{5}$$

3. Results

3.1. Data set and image pre-processing

Data used for evaluation of our developed classification algorithm were taken 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), 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 the 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).

Our experimental evaluations utilized a portion of the ADNI database. We used the T1-weighted Magnetic Resonance (MR) imaging data from the baseline visit. MRI acquisition had been done according to the ADNI acquisition protocol in (Jack et al., 2008). T1-weighted MR image data from 652 ADNI participants are used for evaluation. These 652 subjects include 198 AD, 225 MCI (112 stable MCI (sMCI) and 113 progressive MCI (pMCI)) and 229 NC. Table 1 presents a summary of the demographic characteristics of the studied population from the ADNI database in this paper.

The image processing of the T1-weighted MR brain images was performed as described in Section 2, which included the correction of intensity inhomogeneity, skull-stripping, and cerebellum-removing. Furthermore, each MR brain image was segmented into three tissue types: GM, WM and CSF, and was further spatially normalized into a template space by a mass-preserving registration framework (Shen and Davatzikos, 2003). After spatial normalization, the tissue density maps were smoothed using a Gaussian kernel (its sigma was set to the default value 1.0) to improve signal to noise ratio. Since the gray matter (GM) probability maps were more related to AD than white matter and CSF, we used only the GM probability maps for classification in the experiments. To further reduce the size of image

data, we downsampled the GM tissue density maps from 256×256×256 to 64×64×64 voxels. The downsampled GM tissue density maps were directly used as the representation of features for classification.

In the experiments, 10-fold cross-validation is performed to evaluate the classification performance. For each time, one-fold data set is used for testing while the other folds are used for training. The training set can be split further into training part and validation part for parameter tuning. The final classification accuracy is the average of the classification accuracies across all 10 cross-validation folds.

3.2. AD classification results

(1) Results using single classifier—Before we evaluate the performance of the proposed RPSE_SRC classification framework, we first test the classification performance using a single sparse representation-based classifier (SRC) on the MRI data, in comparison with the standard SVM classifier which has been widely used for AD classification. In this experiment, the SVM classifier is implemented using LIBSVM toolbox (Chang and Lin, 2001), with a linear kernel and a default value set for the parameter C (i.e., C=1). Both the SRC and SVM classifiers are tested on the selected voxel-wise features. To test the classification performances on varying number of relevant features, we perform the t-test on each voxel of the GM tissue density maps. Then all voxels are ranked in ascending order according to their p-values of the t-test. Smaller p-value indicates larger group difference for the voxel-wise feature with more discriminative information for classification. We select different numbers of top ranked voxels in terms of p-values to construct feature vector as the inputs to SRC and SVM classifiers, respectively, for classification. The number of top ranked features varies from 200 to 24000. Figure 2 shows the classification accuracies of SVM and SRC classification methods with respect to different numbers of top ranked features selected for AD classification.

As can be seen from Figure 2, SVM classifier performs better than the SRC method when the number of features is smaller than 1500, but its performance degrades gradually and is inferior to SRC when the number of features is further increased beyond a certain number. In contrast, SRC can achieve much better classification performance than SVM when more features are used. Since SVM classifier aims to maximize the discriminative power on the training data, the features with larger p-values will provide more irrelevant or noisy information which will reduce the discrimination capability and degrade the classification performance. This explains why SVM achieves better performance with a relatively small number of top ranked features. On the other hand, SRC is based on combining the sparsity and reconstruction via sparse representation and thus can achieve high robustness to noisy features due to its reconstruction property (i.e., The classification of a test sample is made by checking which class produces the least reconstruction error with the associated sparse coefficients). In general, to make the L1-norm sparse coding computationally feasible, the dimensionality of the training and testing features should be reduced by extracting a subset of features from the original image. However, our experimental results show that SRC method continues to perform well when the feature dimensionality increases.

(2) Effects of classification parameters—Next, we have performed a number of experiments to test the effects of classification parameters on the performance of the random patch based subspace ensemble classification framework. The SRC is used to construct each weak classifier. In general, there are three important parameters that are required to determine and affect the ensemble performance, which are, respectively, the patch size, the sampling rate (i.e., the ratio of sampled patches to the cardinality of patch pool), and the ensemble size (i.e., the number of weak classifiers for the final ensemble). Since ensemble

of different weak classifiers may produce different classification results, the ensemble classification accuracy is computed by averaging the accuracies of multiple independent runs (20 in our experiments) for each ensemble size.

It is worth repeating that the main purpose for randomly sampling local patches, instead of voxels, is to capture the information of local spatial consistency for AD classification. Thus, in this experiment, we vary the patch size from 1×1×1 (i.e., voxel-wise) to 3×3×3, 5×5×5, 7×7×7, 9×9×9 and 11×11×11 voxels to test their classification performances. In total, there are about 22880, 847, 185, 64, 35 and 16 patches of size 1×1×1, 3×3×3, 5×5×5, 7×7×7, 9×9×9 and 11×11×11, respectively. We randomly select 20%, 40%, 60% and 80% patches from the patch pool to be concatenated into a feature vector for construction of weak classifier. If the sampling rate is small, i.e., less patches are selected for construction of each classifier, more diverse classifiers are usually required to obtain the coverage of feature space and achieve the stable classification performance. Based on our experiments, about 15 weak classifiers can obtain stable ensemble classification performance. Thus, the ensemble size, i.e., the number of weak classifier used for the final ensemble, is varied from 1 to 20. Figure 3(a)–(d) show the SRC ensemble classification results (Classification accuracy vs. Number of weak classifiers) by using different patch sizes at the sampling rates of 20%, 40%, 60% and 80% patches, respectively. From Figure 3, we can see that the classification performances are improved by increasing the patch size from 1×1×1 to 7×7×7 voxels, but further increasing the patch size to 9×9×9 and 11×11×11 voxels will degrade the classification performance. The classifiers ensemble by using the patches of 7×7×7 voxels consistently outperforms those using other patch sizes at four different sampling rates. This may indicate that the patch of size 7×7×7 voxels is able to capture the local spatial consistency of AD-related patterns on this imaging data. When the sampling rate is small, e.g., 20%, the ensemble performance can be gradually improved by increasing the number of weak classifiers until most of the feature space is covered. On the other hand, when the sampling rate is large, e.g., 80%, small number of weak classifiers can get good ensemble performance and further increasing the number of weak classifiers will likely result in redundancy. Nevertheless, about 15 weak classifiers can obtain stable ensemble performance and the ensemble of multiple classifiers often performs better than the individual classifier.

Another important parameter that affects the ensemble performance is the sampling rate, which determines the dimensionality of sampled subspace (i.e., how many local patches should be selected in each random sampling) to construct a weak classifier. Smaller sampling rate can be used to better address the ‘curse of dimensionality’ problem and obtain high diverse subspaces, but the subspace covers less information which limits the performance of each individual weak classifier. On the other hand, high sampling rate can capture more information of the feature space and achieve better performance for each individual weak classifier, but the sampled subspaces will be of high dimensionality and have less diversity. Figure 4 demonstrates the RPSE_SRC classification results (Classification accuracy vs. Number of weak classifiers) by using five different sampling rates which are 20%, 40%, 60%, 80% and 100% with the patch size fixed to 7×7×7 voxels. The 100% sampling rate indicates that all patches are selected and thus only one classifier is constructed to make the classification. From this figure, we can see that the ensemble classification accuracy is gradually improved by increasing the number of weak classifiers with the lower sampling rate, while classifiers ensemble is less sensitive to the number of weak classifier with higher sampling rate. To balance the tradeoff between the accuracy and diversity of weak classifiers, sampling 40% and 60% patches can achieve better classification performance than others. At the sampling rate of 60%, we can see that the combined classification accuracy is stable when the number of weak classifiers is 12 while more weak classifiers are required to achieve better combined classification accuracy at the

sampling rate of 40%. The results also show that the ensemble of multiple weak classifiers consistently performs better than that using only one classifier (i.e., sampling 100% patches), when a sufficient number of weak classifiers are used.

In addition, for analysis of the performance variance, we compute the standard deviations (std) of the ensemble classification accuracies of multiple independent runs with the patch size set to 7×7×7. At a certain ensemble size, the std often decreases by increasing the sampling rate, e.g., the std decreases from 0.021, 0.015, 0.014 to 0.011 when the sampling rate is increased from 20%, 40%, 60% to 80%, respectively, with the ensemble size set to 7. On the other hand, at a certain sampling rate, the std will gradually decrease by increasing the ensemble size (i.e., the number of weak classifiers for ensemble), e.g., the std decreases from 0.022 to 0.004 when increasing the ensemble size from 1 to 20 at the sampling rate of 60%. Thus, our method can achieve stable classification performance by ensemble of multiple classifiers.

(3) Results comparison—To compare with the widely used SVM classifier, we perform the random patch based subspace ensemble method by replacing SRC with SVM on the same data set. Similar to the previous experiments, the SVM classifier is implemented using LIBSVM toolbox (Chang and Lin, 2001), with a linear kernel and a default value for the classifier parameter C (i.e., C=1). We have tested the random patch based subspace ensemble method with the SVM classifiers (RPSE_SVM) on the same feature space in patch extraction as that with the SRC by varying the patch size and sampling rate. The patch size and sampling rate that achieve best ensemble classification accuracy are 7×7×7 voxels and 60%, respectively, which are same as those in ensemble of SRC. However, the ensemble classification accuracy may not be optimal for SVM. From Figure 2, we know that the SVM classification performance will degrade when the number of top ranked features is increased beyond 1500, since more irrelevant features are included in the feature set.

To make more fair comparison, we preselect a smaller number of relevant voxels for patch extraction by using a threshold of p-values smaller than 0.05 to improve the accuracy of individual SVM classifiers. However, by using a smaller threshold, less number of relevant voxels can be obtained for random sampling, which will result in low diversity of individual weak classifiers and also degradation of ensemble classification performance. We have tested different thresholds of p-value from 0.05 to 0.0001 to preselect the relevant voxels and determine the optimal threshold (e.g., 0.001) with the best ensemble classification accuracy. Thus, the voxels with p-values smaller than 0.001 are preselected for patch extraction. All other processing steps are the same as those used to implement the RPSE_SRC. Similar to the SRC based experiments, we also vary the patch size from 1×1×1 to 3×3×3 5×5×5, 7×7×7, 9×9×9 and 11×11×11 voxels by sampling 20%, 40%, 60%, 80% patches. The optimal patch size with which the ensemble SVM achieves the best classification performance is 9×9×9 voxels. Figure 5 shows the comparison of the ensemble classification performance with SVM and SRC classifiers (Classification accuracy vs. Number of weak classifiers) at the four different sampling rates: 20%, 40%, 60% and 80%. From this figure, we can see that RPSE_SRC consistently performs better than the RPSE_SVM at different sampling rates.

In addition, we evaluate the ensemble of individual SVM and SRC classifiers using the kappa-accuracy diagram which evaluates the level of agreement between two classifier outputs while correcting for chance (Rodriguez et al., 2006). Figure 6 shows the diversity-accuracy diagrams of the pairs of individual SVM and SRC classifiers, the centroids of the kappa-accuracy cloud points and their ensemble classification accuracies. From this figure, we can see that there is no great difference in the kappa diversity of individual SVM and SRC classifiers but SRC classifiers achieve much higher accuracy than SVM classifiers.

In practice, the three classification parameters (i.e., patch size, sampling rate and ensemble size) can be optimized in each fold with the training data set to run the random patch-based subspace ensemble classification algorithm. However, from the above analysis, the effect of ensemble performance by ensemble size is stable if the number of weak classifiers is larger than 15 for ensemble. Thus, for simplicity, the ensemble size is fixed to 17 in this experiment. The other two parameters are optimized with the training data set. We also run the COMPARE algorithm (Fan et al., 2007) on the same data set for comparison, by using its suggested parameters. COMPARE algorithm divides the image space into homogeneously discriminative regions and the voxel values in these regions are aggregated to form the features for classification. SVM classifier is employed to perform classification. The classification accuracy by COMPARE algorithm is 81.07%, with 78.84% sensitivity and 82.94% specificity. The main reason that COMPARE algorithm gets lower classification performance is likely because the adaptively extracted regions may not be discriminative enough for different populations. Actually, it is not easy to identify a set of discriminative regions for large population. Table 2 gives the comparison of AD classification in five different classification methods, which are COMPARE, single SVM classifier (SVM), single SRC classifier (SRC), random patch-based SVM ensemble (RPSE_SVM), and random patch-based SRC ensemble (RPSE_SRC), respectively. For single SVM and SRC classifiers, we report their best classification results in Table 2 among those on the different numbers of top ranked features selected by t-test as shown in Figure 2. The ROC curves of these methods are shown in Figure 7. We can see that single SRC method performs better than COMPARE and both single and ensemble SVM methods. RPSE_SRC can further improve the classification accuracy by ensemble of multiple weak classifiers.

3.3. MCI classification results

In addition to classification of AD and NC, we perform the single SRC and the RPSE_SRC algorithms as well as the SVM and RPSE_SVM algorithms for classification of MCI and NC, where 225 MCI and 229 NC subjects are used for test. Similar to AD classification, we select different numbers of top ranked voxels in terms of p-values to construct feature vector as the input to the SRC and SVM classifiers. The number of top ranked features varies from 200 to 24000. Figure 8 shows the classification accuracies of single SVM and SRC methods with respect to different numbers of top ranked features selected for MCI classification. The random patch-based subspace ensemble classification framework with the SVM and SRC are also performed on the same data set. The classifier parameters are optimally determined based on the cross validation of training data set. Table 3 lists the MCI classification results of the SVM, SRC, RPSE_SVM, and RPSE_SRC methods, respectively. For single SVM and SRC methods, we report their best classification results in Table 3 among those on the different numbers of top ranked features selected by t-test as shown in Figure 8. These results also show that the single SRC method performs better than the SVM-based methods. RPSE_SRC can further improve the classification accuracy by ensemble of multiple weak classifiers, which indicates the efficacy of the proposed classification method in classifying MCI from NC.

3.4. Comparison with existing classification methods

Furthermore, we have compared the results of the proposed RPSE_SRC method with some recent results reported in the literature that are also based on MRI data of ADNI subjects for AD and MCI classification. In particular, three recent methods are compared in Table 4, as briefly described next. In (Hinrichs et al., 2009), the linear Program (LP) boosting method with novel additional regularization have been proposed to incorporate the spatial smoothness of 3D MR imaging space into the learning process and improve the classification accuracy. In (Cuingnet et al., 2011), ten methods, which include five voxel-

based methods, three cortical thickness based methods, and two hippocampus based methods, are compared with the linear SVM classifier on MRI data of ADNI subjects. The best result using voxel-wise gray matter (GM) features is provided in Table 4. In (Zhang et al., 2011), 93 volumetric features extracted from the 93 regions of interest (ROI) in GM density maps of MRI data have been used as classification features and SVM classifier is used to make classification. The results of these three methods, along with our proposed method, reported in Table 4 further show the efficacy of our proposed method in AD and MCI classification.

3.5. Discussion

Since the AD is affected in certain pathological patterns, only a subset of voxels may be related to AD changes. Thus, the simple thresholding based on t-test was performed to preselect the relevant voxels for patch extraction and random sampling in the above experiments. We have also performed the RPSE_SRC on the whole brain tissue density maps and the classification accuracy is 87.61% with 80.87% sensitivity and 93.42% specificity which is worse than that on the preselected brain regions. This shows the importance of preselection of relevant voxels in our proposed RPSE_SRC method. For accurate classification, the sampled feature space should include more relevant voxels. If the whole brain is used for random sampling, it is more likely to select the unimportant and noisy voxels which will degrade the accuracies of individual classifiers. On the other hand, the threshold on the p-value cannot be too strict to get diverse subspaces on individual classifiers for effective ensemble. This threshold can be also regarded as an additional parameter of the algorithm, which can be optimized by using cross validation on the training data set. For simplicity, we use a fixed value, i.e., 0.05, in our experiments. It is worth noting that the use of the simple t-test to select the relevant features may eliminate features that are unimportant in isolation but discriminative when combined with others. Using more complex feature selection methods (e.g., wrapper-based) may overcome this problem. However, in our method, t-test is used only for a coarse feature selection and another random feature (patch) selection process will be performed in subsequent steps. Our experiments show that even with the simple t-test, our method can achieve better performance than conventional methods. More complex feature selection methods could be used to further improve the performance which will be our future work.

The sampled feature subspace has direct effect on the performance of individual classifier. If the sampled patches are from the AD-affected regions, the corresponding individual classifier will achieve high classification accuracy. On the other hand, the classification accuracy will be low if the sampled patches are from the less relevant regions. For each fold of the training data set, we have constructed 30 individual classifiers by randomly sampling the original feature space 30 times with the optimized patch size. The classification accuracies of the individual classifiers are sorted in ascending order. The average sampling frequencies of the available patches from the five most accurate classifiers are computed across the ten-fold cross-validation to evaluate the importance of patches. We choose the patches which are more frequently sampled in high-accuracy classifiers. We found these patches are located at the brain regions such as the hippocampus, the parahippocampal gyrus, the entorhinal cortex and the amygdala which are consistent with those reported in the literature (Cuingnet et al., 2011; Zhang et al., 2011).

In addition to reporting the classification accuracy, visualizing the learned decision process is also important to understand the classification algorithm and gain clinical insights. However, as in most AD classification methods (e.g., SVM), visualizing the learned decision process of our method is not informative which is a limitation that is expected to address in the future.

4. Conclusion

In this paper, we have investigated using the SRC for classification of high dimensional MRI data. Furthermore, we have presented a random patch based subspace ensemble classification framework with the SRC. Instead of randomly sampling the voxels, the local patches are extracted from the relevant regions to capture the local spatial consistency and are randomly sampled to construct a feature subspace for design of individual weak classifier. Then, multiple classifiers are combined to make more accurate and robust classification. The experimental results on ADNI database show that SRC continues to perform well when the dimensionality increases. It achieves better classification performance than SVM classifier when more features are used for classification. The random patch based subspace ensemble classification can further improve the classification accuracy by combining multiple weak classifiers and using local patches to capture spatial consistency.

In the current paper, we validate our method using MRI data from ADNI. However, other modality of data can also be used in our method. In the future work, we will validate our method on other imaging data, e.g., PET. Moreover, since recent studies have shown that different modalities of neuroimaging data can provide complementary information for AD diagnosis, we will extend our method to multiple modalities of biomarker to further improve the accuracy of AD classification..

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by NIH grants EB006733, EB008374, EB009634 and MH088520, and by NBRPC 973 Program grant (No. 2010CB732505), NSFC grants (No. 61075010, No. 61005024) and Medical and Engineering Foundation of Shanghai Jiao Tong University (No. YG2010MS74). 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

http://www.nia.nih.gov/Alzheimers/ResearchInformation/ClinicalTrials/ADNI.htm.

Baron JC, Chetelat G, Desgranges B, Perchey G, Landeau B, de la Sayette V, Eustache F. In vivo mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease. NeuroImage. 2001; 14:298–309. [PubMed: 11467904]

Boyd, SP.; Vandenberghe, L. Convex optimization. Cambridge Univ Pr; 2004.

Candes, E.; Romberg, J. l1-magic: Recovery of sparse signals via convex programming. 2005. URL: www.acm.caltech.edu/l1magic/downloads/l1magic.pdf

Chan D, Janssen J, Whitwell J, Watt H, Jenkins R, Frost C, Rossor M, Fox N. Change in rates of cerebral atrophy over time in early-onset Alzheimer's disease: longitudinal MRI study. Lancet. 2003; 362:1121–1122. [PubMed: 14550701]

Chang, CC.; Lin, CJ. LIBSVM: a library for support vector machines. 2001.

Che DS, Liu Q, Rasheed K, Tao XP. Decision Tree and Ensemble Learning Algorithms with Their Applications in Bioinformatics. Software Tools and Algorithms for Biological Systems. 2011; 696:191–199.

Chen SS, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. SIAM review. 2001; 43:129.

Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehericy S, Habert MO, 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. 2011; 56:766–781. [PubMed: 20542124]

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

Davatzikos C, Fan Y, Wu X, Shen D, Resnick SM. Detection of Prodromal Alzheimer's Disease via Pattern Classification of MRI. Neurobiology of aging. 2006 epub.

Davatzikos, C.; Resnick, SM.; Wu, X.; Parmpi, P.; Clark, CM. Individual Patient Diagnosis of Alzheimer's and Frontotemporal dementias via High-Dimensional Pattern Classification of MRI. Alzheimer's Disease conference; 2007.

Davatzikos C, Resnick SM, Wu X, Parmpi P, Clark CM. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. NeuroImage. 2008; 41:1220–1227. [PubMed: 18474436]

Desikan RS, Cabral HJ, Hess CP, Dillon WP, Glastonbury CM, Weiner MW, Schmansky NJ, Greve DN, Salat DH, Buckner RL, Fischl B. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. Brain. 2009; 132:2048–2057. [PubMed: 19460794]

Duda, RO.; Hart, PE.; Stork, DG. Pattern Classification. John Wiley and Sons, Inc; 2001.

Elad M, Aharon M. Image denoising via sparse and redundant representations over learned dictionaries. Image Processing, IEEE Transactions on. 2006; 15:3736–3745.

Fadili M, Starck JL, Murtagh F. Inpainting and zooming using sparse representations. The Computer Journal. 2009; 52:64.

Fan Y, Resnick SM, Wu X, Davatzikos C. Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. NeuroImage. 2008; 41:277–285. [PubMed: 18400519]

Fan, Y.; Shen, D.; Davatzikos, C. Classification of Structural Images via High-Dimensional Image Warping, Robust Feature Extraction, and SVM. In: Duncan, JS.; Gerig, G., editors. MICCAI. Springer; Berlin/Heidelberg, Palm Springs, California, USA: 2005. p. 1-8.

Fan Y, Shen D, Gur RC, Gur RE, Davatzikos C. COMPARE: Classification Of Morphological Patterns using Adaptive Regional Elements. IEEE transactions on medical imaging. 2007; 26:93–105. [PubMed: 17243588]

Hazan, T.; Polak, S.; Shashua, A. Sparse image coding using a 3D non-negative tensor factorization. The tenth IEEE International Conference on Computer Vision; 2005; Beijing, China: IEEE; 2005. p. 50-57.

Hinrichs C, Singh V, Mukherjee L, Xu G, Chung MK, Johnson SC. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. NeuroImage. 2009; 48:138–149. [PubMed: 19481161]

Ho TK. The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence. 1998; 20:832–844.

Huang K, Aviyente S. Sparse representation for signal classification. Advances in neural information processing systems. 2007; 19:609.

Ishii K, Kawachi T, Sasaki H, Kono AK, 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. 2005; 26:333–340. [PubMed: 15709131]

Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, JLW, Ward C, Dale AM, Felmlee JP, Gunter JL, Hill DL, Killiany R, Schuff N, Fox-Bosetti S, Lin C, Studholme C, DeCarli CS, Krueger G, Ward HA, Metzger GJ, Scott KT, Mallozzi R, Blezek D,

Levy J, Debbins JP, Fleisher AS, Albert M, Green R, Bartzokis G, Glover G, Mugler J, Weiner MW. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. Journal of magnetic resonance imaging : JMRI. 2008; 27:685–691. [PubMed: 18302232]

Kim SJ, Koh K, Lustig M, Boyd S, Gorinevsky D. An interior-point method for large-scale l1-regularized least squares. Selected Topics in Signal Processing, IEEE Journal of. 2007; 1:606–617.

Kloppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, Fox NC Jr, CRJ, Ashburner J, Frackowiak RSJ. Automatic classification of MR scans in Alzheimer's disease. Brain. 2008 epub.

Kuncheva LI, Rodriguez JJ, Plumpton CO, Linden DE, Johnston SJ. Random subspace ensembles for FMRI classification. IEEE transactions on medical imaging. 2010; 29:531–542. [PubMed: 20129853]

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. NeuroImage. 2004; 21:46–57. [PubMed: 14741641]

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. 2009; 51:73–83. [PubMed: 18846369]

Majumdar A, Ward RK. Fast group sparse classification. Canadian Journal of Electrical and Computer Engineering-Revue Canadienne De Genie Electrique Et Informatique. 2009; 34:136–144.

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's & dementia : the journal of the Alzheimer's Association. 2005; 1:55–66.

Ratsch G, Scholkopf B, Smola AJ, Mika S, Onoda T, Muller KR. Robust ensemble learning for data mining. Knowledge Discovery and Data Mining, Proceedings. 2000; 1805:341–344.

Rodriguez JJ, Kuncheva LI, Alonso CJ. Rotation forest: A new classifier ensemble method. IEEE transactions on pattern analysis and machine intelligence. 2006:1619–1630. [PubMed: 16986543]

Shen D, Davatzikos C. Very high resolution morphometry using mass-preserving deformations and HAMMER elastic registration. NeuroImage. 2003; 18:28–41. [PubMed: 12507441]

Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE transactions on medical imaging. 1998; 17:87–97. [PubMed: 9617910]

Thompson P, Cannon TD, Toga AW. Mapping genetic influences on human brain structure. Annals of medicine. 2002; 34:523–536. [PubMed: 12553492]

Valev V, Asaithambi A. Multidimensional pattern recognition problems and combining classifiers. Pattern Recognition Letters. 2001; 22:1291–1297.

Vemuri P, Gunter JL, Senjem ML, Whitwell JL, Kantarci K, Knopman DS, Boeve BF, Petersen RC Jr, CRJ. Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies. NeuroImage. 2008; 39:1186–1197. [PubMed: 18054253]

Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust Face Recognition via Sparse Representation. IEEE transactions on pattern analysis and machine intelligence. 2009; 31:210–227. [PubMed: 19110489]

Yang, M.; Zhang, L.; Yang, J.; Zhang, D. Robust Sparse Coding for Face Recognition. CVPR; 2011.

Yoon U, Lee JM, Im K, Shin YW, Cho BH, Kim IY, Kwon JS, Kim SI. Pattern classification using principal components of cortical thickness and its discriminative pattern in schizophrenia. NeuroImage. 2007; 34:1405–1415. [PubMed: 17188902]

Zhang D, Wang Y, Zhou L, Yuan H, Shen D. Multimodal classification of Alzheimer's disease and mild cognitive impairment. NeuroImage. 2011; 55:856–867. [PubMed: 21236349]

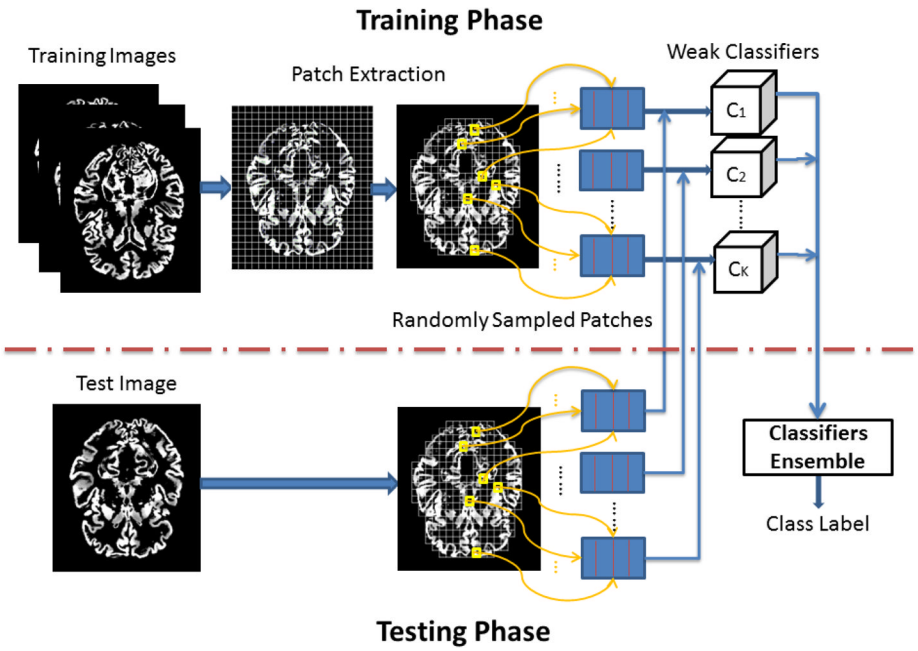


Figure 1.
The framework of the random patch-based subspace ensemble classification method.

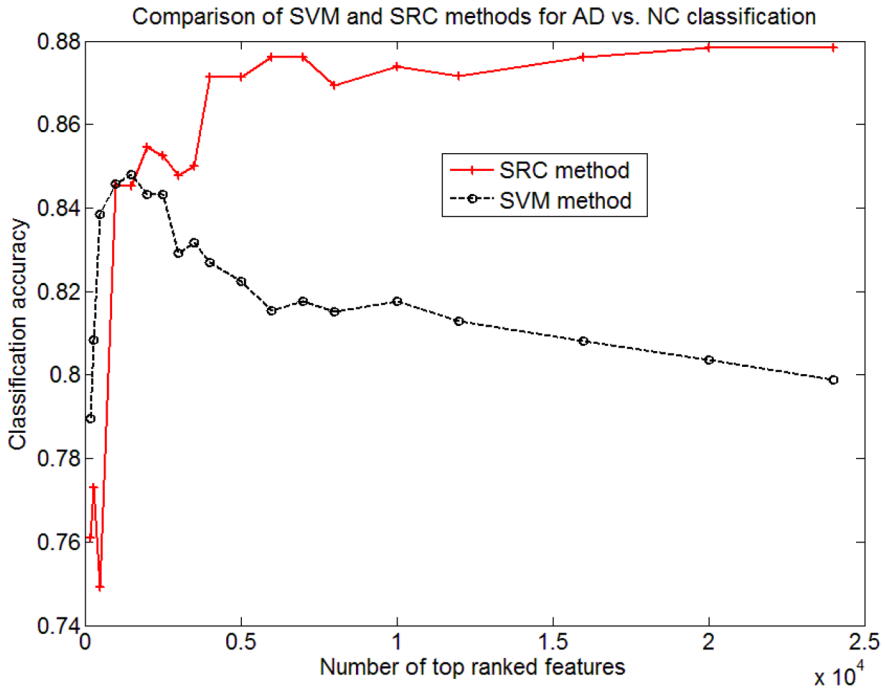
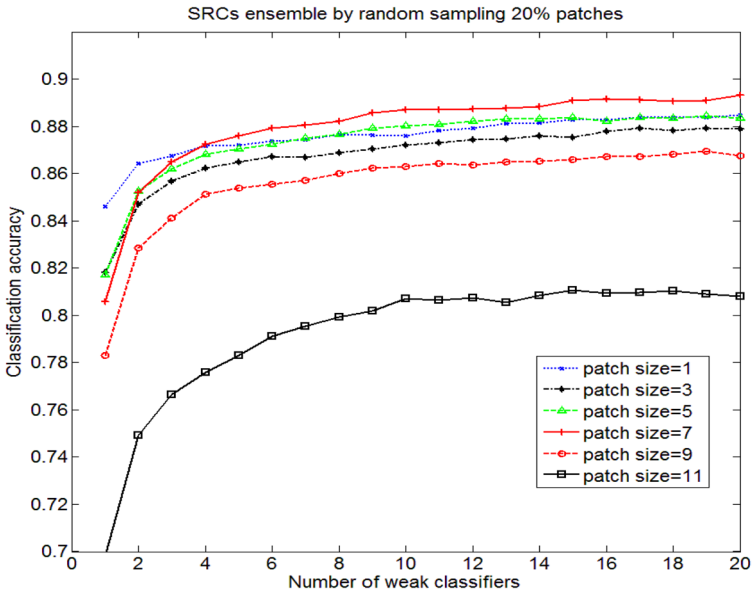
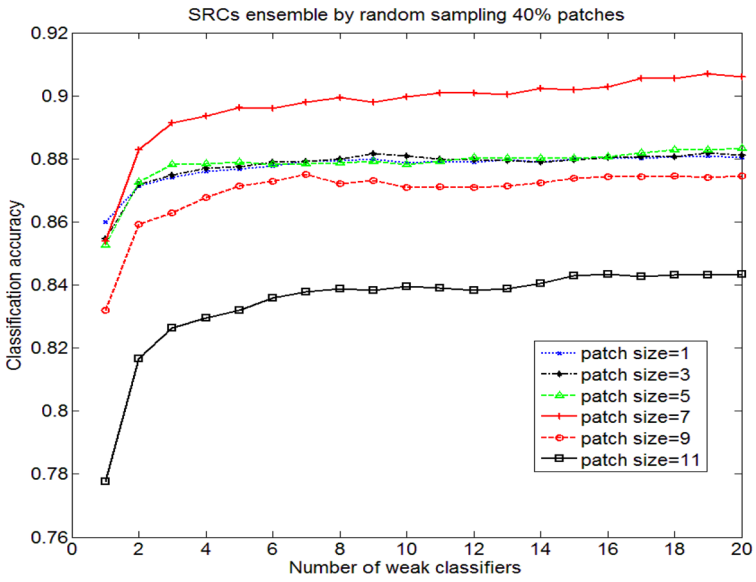
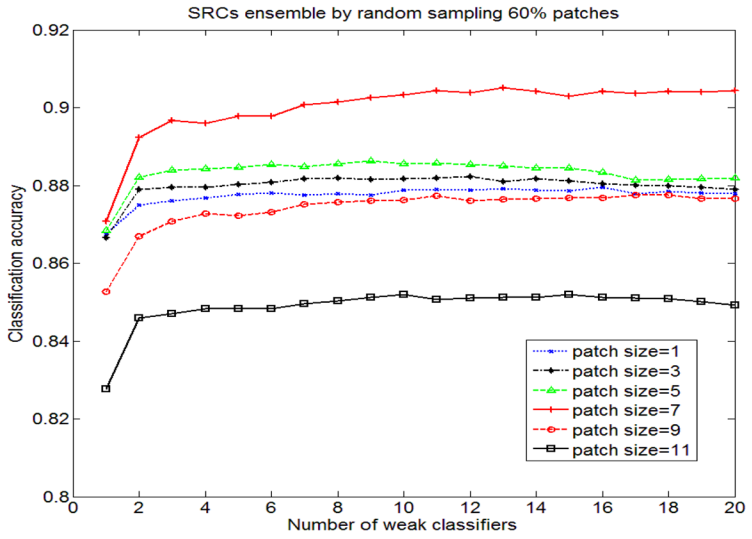


Figure 2.
Classification accuracies of SVM and SRC with respect to different numbers of top ranked features selected for AD classification.

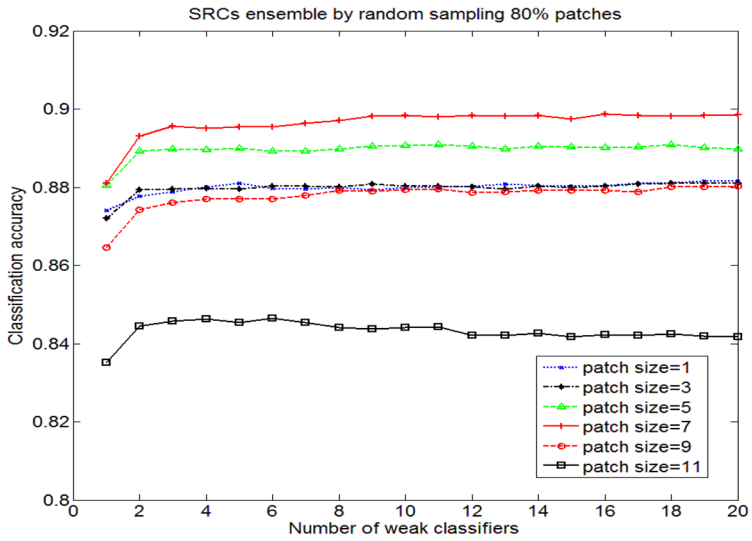


(a)





(c)



(d)

Figure 3.
Classification results using different patch sizes at four sampling rates: (a) 20%, (b) 40%, (c) 60%, and (d) 80%.

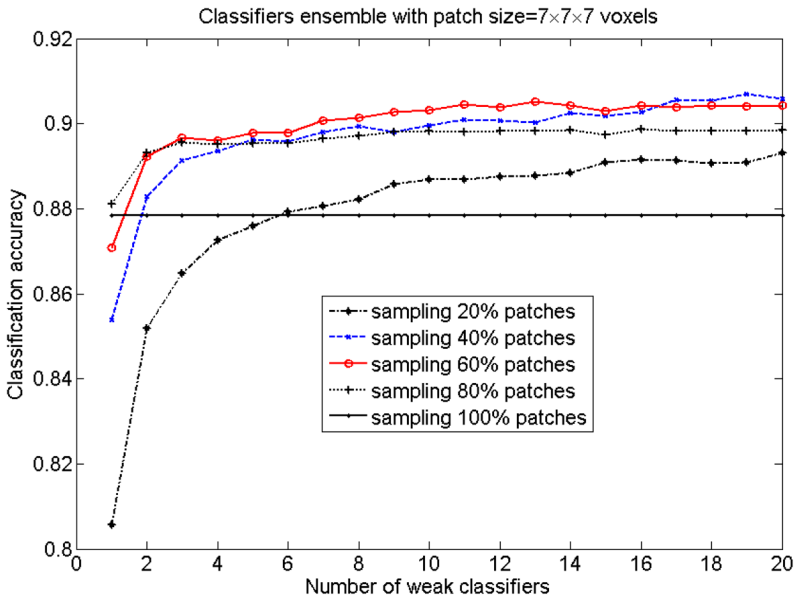


Figure 4.

SRCs ensemble classification results using five different sampling rates with the patch size set to 7×7×7 voxels.

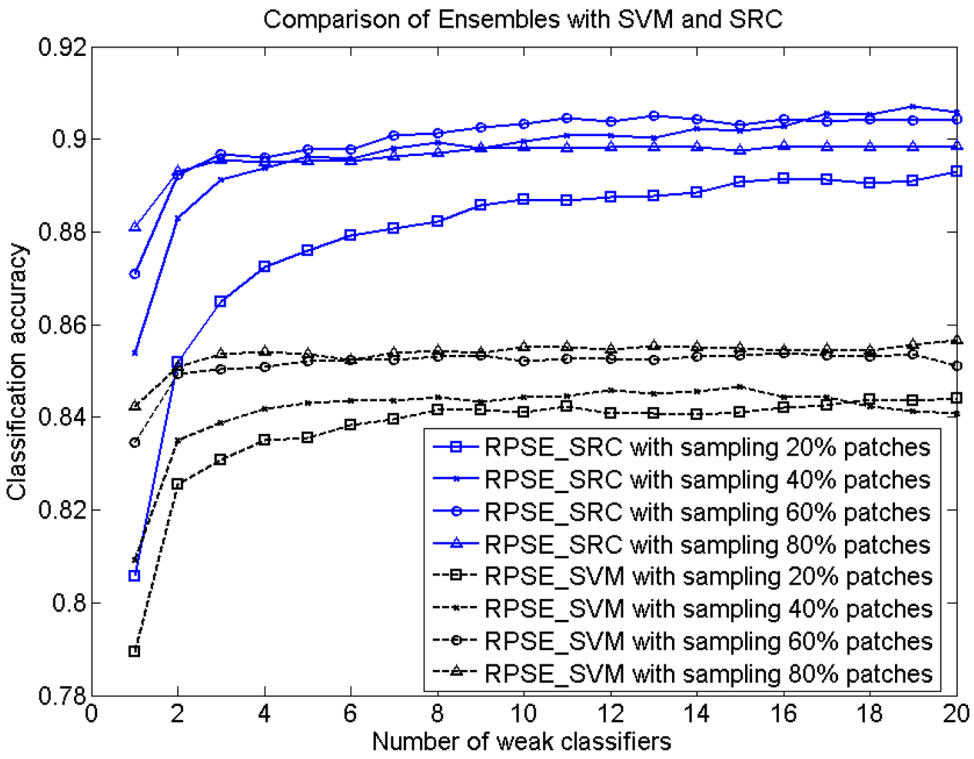


Figure 5.
Comparison of ensemble classification results using SVM with 9×9×9 patch size and SRC with 7×7×7 patch size at four sampling rates.

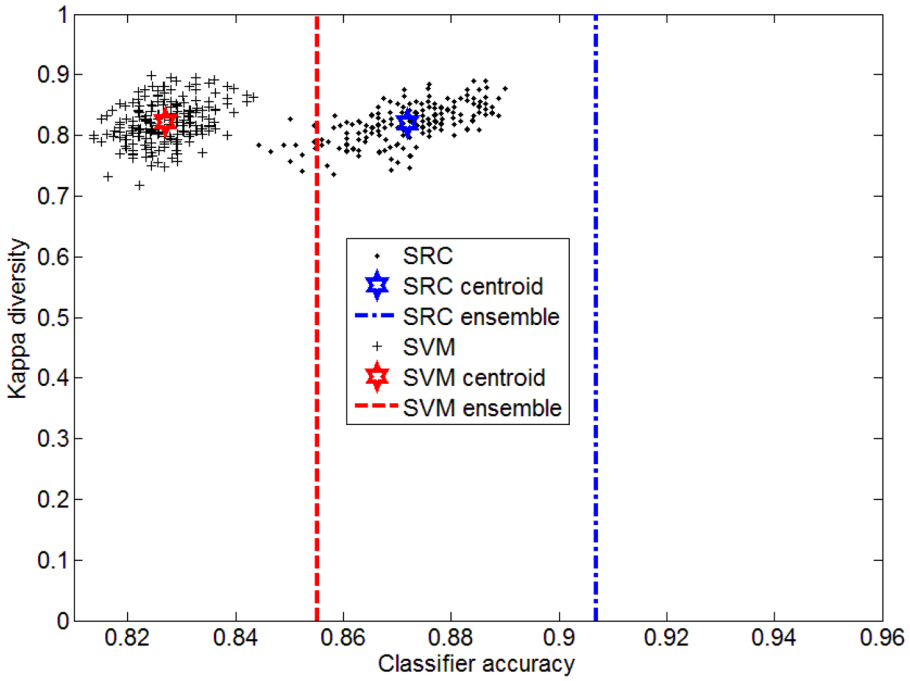


Figure 6.

The diversity-accuracy diagrams of SVM and SRC classifiers. The patch size, sampling rate and ensemble size are set to 7×7×7, 60% and 20, respectively. The x-axis represents average accuracy of a pair of classifiers, and y-axis represents diversity of a pair of classifiers evaluated by the kappa measure. The blue and red dashed vertical lines show the ensemble accuracies of SRC and SVM, respectively. The blue and red hexagrams denote the centroids of SRC and SVM classifier clouds, respectively.

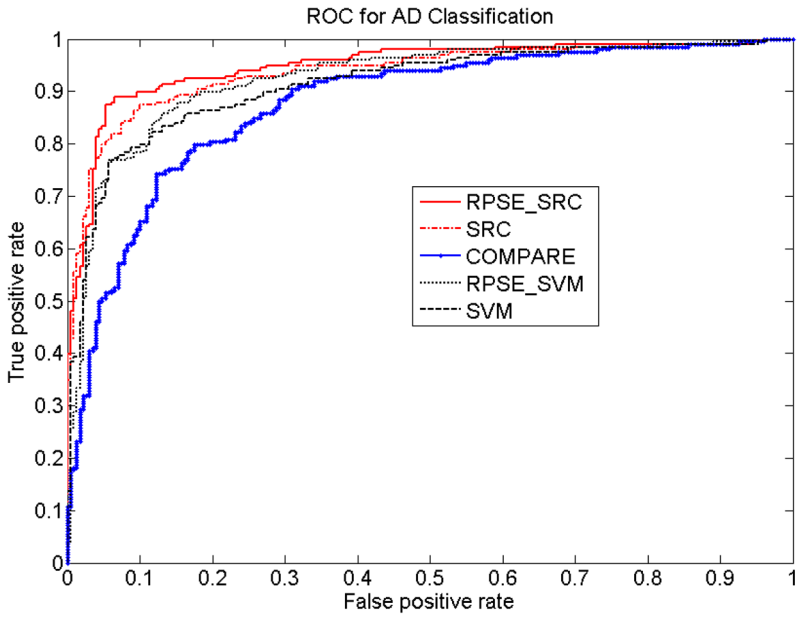


Figure 7.
ROC curves of five different methods for AD classification.

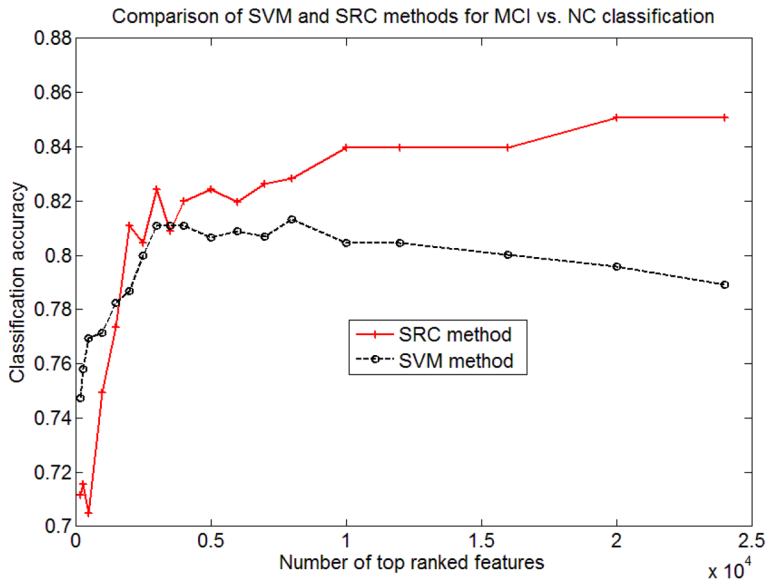


Figure 8.
Classification accuracy of SVM and SRC with respect to different numbers of top ranked features selected for MCI classification.

Table 1

Demographic characteristics of the studied population (from the ADNI database). The values are denoted as mean ± standard deviation.

Diagnosis	Number	Age	Gender (M/F)	MMSE
AD	198	75.7±7.7	103/93	23.3±2.0
MCI	225	75.2±7.4	154/71	26.7±1.8
NC	229	76.0±5.0	119/110	29.1±1.0

Table 2

Comparison of AD classification in five different classification methods.

Methods	ACC (%)	SEN (%)	SPE (%)	Area under ROC (%)
COMPARE	81.07	78.84	82.94	87.65
SVM	84.57	72.82	94.76	91.40
SRC	87.83	80.84	93.85	93.77
RPSE_SVM	85.53	75.47	94.24	92.39
RPSE_SRC	**90.80**	**86.32**	**94.76**	**94.86**

Table 3

Comparison of MCI classification on 4 different classification methods.

Methods	ACC (%)	SEN (%)	SPE (%)	Area under ROC (%)
SVM	81.33	73.00	89.53	87.58
SRC	85.08	82.77	87.35	91.66
RPSE_SVM	82.26	73.70	90.69	90.92
RPSE_SRC	87.85	85.26	90.40	92.90

Table 4

Comparison of AD and MCI classification results reported in the literature using MR imaging data of ADNI subjects.

Methods	Features	Classifier	Subjects	ACC (%)	SEN (%)	SPE (%)
(Hinrichs et al., 2009)	Voxel-wise GM	(LP) boosting	183 (NC+AD)	82.00	85.00	80.00
			–	–	–	–
(Cuingnet et al., 2011)	Voxel-wise GM	SVM	162NC+137AD	88.58	81.00	95.00
			162NC+76MCI	81.17	73.00	85.00
(Zhang et al., 2011)	93 ROI GM	SVM	52NC+51AD	86.20	86.0	86.3
			52NC+99MCI	72.00	78.5	59.6
RPSE_SRC	**Voxel-wise GM**	**SRC ensemble**	**229NC+198AD**	**90.80**	**86.32**	**94.76**
			229NC+225MCI	**87.85**	**85.26**	**90.40**