CHAPTER

Multitemplate-based multiview learning for Alzheimer's disease diagnosis

9

M. Liu[1,2,3], R. Min[1], Y. Gao[4], D. Zhang[2], D. Shen[1]

University of North Carolina at Chapel Hill, Chapel Hill, NC, United States[1] Nanjing University of Aeronautics and Astronautics, Nanjing, China[2] Taishan University, Taian, China[3] School of Software, Tsinghua University, Beijing, China[4]

CHAPTER OUTLINE

Machine Learning and Medical Imaging. http://dx.doi.org/10.1016/B978-0-12-804076-8.00009-8

9.1 BACKGROUND

Alzheimer's disease (AD), characterized by progressive impairment of cognitive and memory function, is the sixth leading cause of death in the United States for Americans aged 65 years or older. According to a recent report from the Alzheimer's Association (2013), the total estimated prevalence of AD is expected to be 13.8 million in the United States by 2050. As there is no cure for AD to reverse its progression, early diagnosis and monitoring of AD at its early prodromal stage, that is, mild cognitive impairment (MCI), is of vital importance.

Over the past decade, advances in magnetic resonance imaging (MRI) have enabled significant progress in understanding neural changes that are related to AD (Chan et al., 2003; Davatzikos et al., 2001; Fan et al., 2008; Fox et al., 1996; Hinrichs et al., 2009; Magnin et al., 2009; Mueller et al., 2005). By directly accessing the structures provided by MRI, brain morphometry can identify the anatomical differences between populations of AD patients and normal controls (NCs) for assisting diagnosis and also evaluating the progression of MCI (Fox et al., 1996; Dickerson et al., 2001; Jack et al., 2008; Wang et al., 2014; Liu et al., 2015). In general, MRI-based classification methods can be roughly divided into two categories, that is, (1) methods using single-template-based morphometric representation of brain structures (Cuingnet et al., 2011; Liu et al., 2012; Argyriou et al., 2008; Zhang et al., 2011) and (2) methods using multitemplate-based representation of brain structures (Liu et al., 2015; Koikkalainen et al., 2011; Leporé et al., 2008; Min et al., 2014a,b).

In the first category of methods, researchers mainly utilize a single template as a benchmark space to provide a representative basis for comparing the common anatomical structures of different brain images. More specifically, they first obtain a morphometric representation of each brain image by spatially normalizing it onto a common space (eg, a predefined template) via nonlinear registration, and thus the corresponding regions in different brain images can be compared (Sotiras et al., 2013; Tang et al., 2009; Yap et al., 2009). Usually, such a predefined template is an image of a single subject, a general average template, or a specific template generated from a particular data set under study (Leporé et al., 2008; Chung et al., 2001; Teipel et al., 2007). In the literature, many single-template-based morphometric pattern analysis methods, such as voxel-based morphometry (VBM) (Davatzikos et al., 2001; Ashburner and Friston, 2000; Davatzikos et al., 2008; Thompson et al., 2001), deformation-based morphometry (DBM) (Chung et al., 2001; Ashburner et al., 1998; Gaser et al., 2001; Joseph et al., 2014), and tensor-based morphometry (TBM) (Koikkalainen et al., 2011; Leporé et al., 2008; Kipps et al., 2005; Whitford et al., 2006; Leow et al., 2006; Hua et al., 2008), have been proposed and have demonstrated promising results in AD diagnosis with different classification techniques (Bozzali et al., 2006; Frisoni et al., 2002; Hua et al., 2013). Specifically, in these methods, after nonrigidly transforming each individual brain image onto a common template space, VBM measures the local tissue density of the original brain image directly, while DBM and TBM measure the local deformation and the Jacobian of local deformation, respectively. For example, researchers in Fan et al. (2007) proposed a classification

of morphological patterns using adaptive regional elements (COMPARE) algorithm to extract volumetric features from self-organized and spatial-adaptive local regions based on a single template. However, due to the potential bias associated with the use of a particular template, the feature representation extracted from a single (particular) template may not be sufficient to reveal the underlying complicated differences between populations of disease-affected patients and NCs.

In the second category of methods, researchers attempt to use multiple templates to minimize the bias associated with the use of a single template. Although requiring higher computational cost, this kind of method can help reduce the negative impact of registration errors in morphometric analysis of brain images. Recently, several studies (Liu et al., 2015; Koikkalainen et al., 2011; Leporé et al., 2008; Min et al., 2014a,b) have shown that the multitemplate-based methods can often offer more accurate diagnosis results than the single-template-based methods. For example, researchers in Leporé et al. (2008) registered each brain image onto multiple templates (which had already been nonlinearly aligned to a new common template), and then averaged their respective Jacobian maps of the estimated deformation fields to improve the TBM-based monozygotic/dizygotic twin classification. In order to reduce errors caused by registration in the TBM-based classification, researchers in Koikkalainen et al. (2011) investigated the effects of utilizing mean deformation fields, mean volumetric features, and mean predicted responses of regression-based classifiers from multiple templates, and obtained improved results for AD analysis. However, one main disadvantage of the above-mentioned methods is that, after averaging the features from multiple templates, morphometric representations for a subject (although generated from different templates) could become less powerful in revealing the underlying complicated differences between AD patients and NCs, because they ignore the characteristics of each template.

It is worth noting that, due to the fact that anatomical structures among different templates can be very different from each other, a subject's corresponding representations generated from different templates (also named as views later) will also be distinctive, as shown in Fig. 9.1. Fig. 9.1 illustrates (1) how different morphometric patterns can be generated from different templates via nonlinear transformation, where we show an example of the tissue density map of white matter (WM) calculated from the registration by HAMMER (Shen and Davatzikos, 2002), and (2) also the amplified differences in comparison of two subjects when different templates are jointly considered. Actually, a similar philosophy is widely applied in other domains. For example, a side-view camera can capture the profile of an object, which is able to provide supplemental information for object recognition in addition to the frontal shot of the same object. In brain morphometry, multiple templates can be similarly regarded as different "cameras" in such measurements for the same "object" brain MRI.

On the other hand, in machine learning and pattern recognition domains, multiview-based learning methods have been well studied to make full use of features from multiple views to represent an object (Li et al., 2002; Liu et al., 2014; Thomas et al., 2006). For example, in multiview face recognition, a human

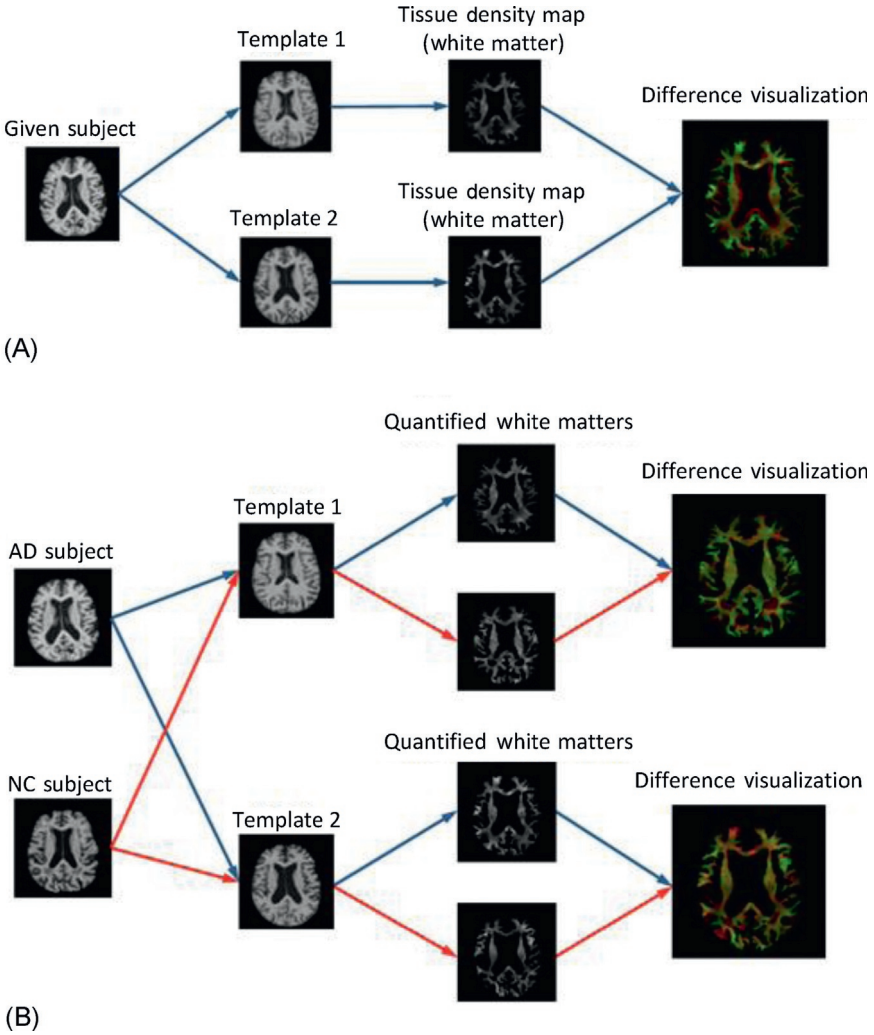

FIG. 9.1

Illustration of different morphometric patterns generated from different templates. (A) Registration of an image to different templates leads to different representations. It can be seen that the geometrical structures of white matter (WM) represented in different templates are different. In addition, tissue density distributions within each tissue are also different from the two different templates. (B) Registration of different images (eg, an AD subject and an NC subject) to different templates: the differences between their representations from individual templates are different (implying the amplified discriminative power when jointly considered in classification).

face can be represented by both the frontal- and side-view images. Since these images can provide different information for the same person, the use of multiple sets of features from different views can largely enrich the representation of the person and achieve significantly enhanced discriminative power, in comparison to features from only one view, as suggested in various studies (Li et al., 2002; Thomas et al., 2006; Basha et al., 2013; Gong et al., 2014; Xu et al., 2014). Similarly, in brain morphometry, multiple templates can also be regarded as multiple views for representing the same brain. Thus a representation generated from a specific template can be regarded as a profile for the brain, and can be used to provide supplementary or side information for other representations generated from the other templates (ie, views).

The rest of this chapter is organized as follows. In Section 9.2 we first introduce a multiview feature representation method for AD diagnosis, by using multiple templates selected from data. In Section 9.3 we present four multiview learning methods for the automatic diagnosis of AD and MCI. Section 9.4 introduces experiments and corresponding analysis. In Section 9.5 we conclude this chapter.

9.2 MULTIVIEW FEATURE REPRESENTATION WITH MR IMAGING

Min et al. (2014b) and Liu et al. (2015) propose a multiview feature representation method by using MRI data. Specifically, they propose to measure brain morphometry via multiple templates, in order to generate a rich representation of anatomical structures, which will be more discriminative to separate different groups of subjects. Unlike previous multitemplate-based works (Koikkalainen et al., 2011; Leporé et al., 2008), which register their templates to a common space via deformable registration, they retain the selected templates in their original (linearly aligned) spaces without nonlinearly registering them to the common space, in order to consider different information provided by different templates. In their method, affinity propagation (Frey and Dueck, 2007) is first applied to select the most distinctive and representative templates. Then, subjects from different groups are registered to different templates by using HAMMER (Shen and Davatzikos, 2002). By adopting a feature extraction method used in COMPARE (Fan et al., 2007), the most discriminative regional features can be extracted in different template spaces.

9.2.1 PREPROCESSING

A standard preprocessing procedure is applied to the T1-weighted MR brain images. First of all, nonparametric nonuniform bias correction (N3) (Sled et al., 1998) is applied to correct intensity inhomogeneity. Then, a skull-stripping method (Wang et al., 2011, 2014) is performed, followed by manual review or correction to ensure

clean skull and dura removal. Cerebellum removal is subsequently conducted by warping a labeled template to each skull-stripping image. Afterwards, each brain image is segmented into three tissues (gray matter (GM), WM, and cerebrospinal fluid (CSF)) by using FAST (Zhang et al., 2001), and finally all brain images are affine aligned by FLIRT (Jenkinson and Smith, 2001; Jenkinson et al., 2002).

9.2.2 TEMPLATE SELECTION

For obtaining a multitemplate-based human brain representation, the first question to address is how to select those multiple templates. In Koikkalainen et al. (2011), 30 templates are randomly selected from different categories (10 for AD, 10 for MCI, and 10 for NC). However, these randomly selected templates cannot guarantee to appropriately reflect the distribution of the whole population. Also, redundant information could be introduced with this random selection; moreover, the selection of unrepresentative images as templates could further cause large registration errors. To overcome these limitations, we propose a data-driven template selection scheme to obtain the most distinctive and representative templates.

In order to select templates that can yield discriminative morphometric representations, differences among the selected templates should be maximized. On the other hand, to reduce registration errors, selected templates should be representative enough to cover the entire population. To this end, the affinity propagation (Frey and Dueck, 2007) algorithm is used to partition the entire population (of AD and NC images) into K (eg, $K = 10$ in this chapter) nonoverlapping clusters. Note that, by performing affinity propagation, an exemplar image will be automatically selected for each cluster, which can then be used as a representative image or template for this cluster. Finally, by combining all exemplar images from all different clusters, we can obtain a set of templates to form the template pool. In the clustering process, a bisection method (Frey and Dueck, 2007) is applied to find the appropriate preference value, and the image similarity is computed as normalized mutual information. The clustering results and the respective selected templates are shown in Fig. 9.2. It should be noted that, although it is possible to add more templates to the set of selected templates, those additional templates could introduce just the redundant information and thus affect the optimal representation of each subject. Here, only templates from the AD and NC subjects are selected, but not from the MCI subjects. This is because MCI can be considered as an intermediate stage between AD and NC and is associated with both AD and NC characteristics.

9.2.3 REGISTRATION AND QUANTIFICATION

The core steps in morphometric pattern analysis (eg, VBM, DBM, or TBM) include (1) a registration step for spatial normalization of different images into a common space and (2) a quantification step for morphometric measurement. Similar to Fan et al. (2007), a mass-preserving shape transformation framework

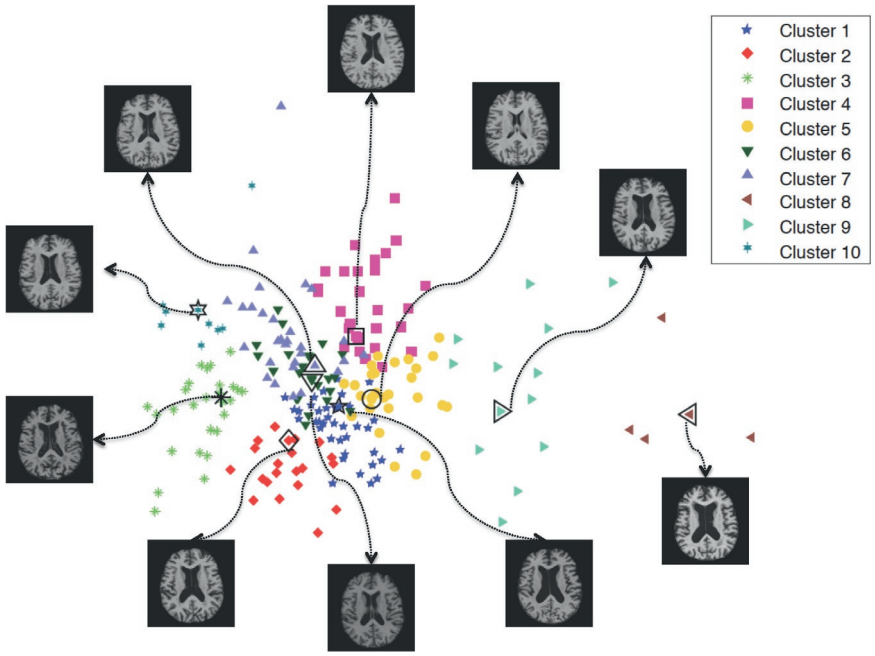

FIG. 9.2

The clustering result of AD/NC subjects, using affinity propagation with normalized mutual information. Each selected template corresponds to an exemplar image of the respective cluster. Points are visualized by multidimensional scaling (Kruskal, 1964).

(Shen and Davatzikos, 2003) is adopted to capture the morphometric patterns of any given subject on the spaces of different templates.

Fig. 9.3 illustrates the multitemplate registration and quantification steps. First, for a given subject with three segmented tissues (ie, GM, WM, and CSF), it will be registered onto multiple (ie, K) selected templates by using a high-dimensional elastic warping tool (ie, HAMMER (Shen and Davatzikos, 2002)). Then, based on those K estimated deformation fields, for each tissue one can quantify its voxel-wise tissue density map in each of the K different template spaces. All these quantified tissue density maps (Davatzikos et al., 2001; Goldszal et al., 1998; Davatzikos, 1998) can thus reflect the unique deformation behavior of the given subject with respect to each different template. In Fig. 9.3 it is clear that the K generated tissue density maps are different in terms of both their density values and tissue structures, which lead to different feature representations as introduced below. Since the GM is mostly affected by AD and thus widely investigated in the literature (Liu et al., 2012; Zhang et al., 2011; Zhang and Shen, 2012), only the GM density map is used for subsequent feature extraction and classification.

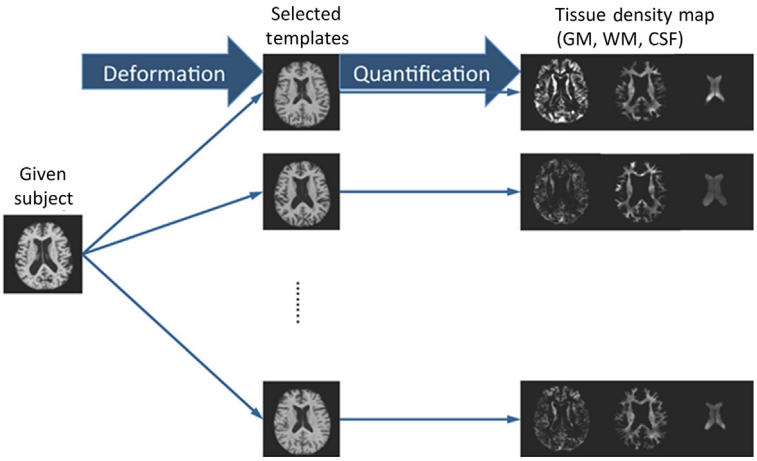

FIG. 9.3

Registration and quantification of a subject registered to multiple templates using HAMMER. Registration to different templates leads to different quantification results. In the figure, the generated tissue density maps (GM, WM, and CSF) are different from registration via different templates.

9.2.4 FEATURE EXTRACTION

Features are first extracted from each individual template space, and then integrated together for a more complete representation. In Section 9.2.4.1 a set of regions-of-interest (ROIs) in each template space is first adaptively determined by performing watershed segmentation (Vincent and Soille, 1991; Grau et al., 2004) on the correlation map obtained between the voxel-wise tissue density values and the class labels from all training subjects. Then, to improve both discrimination and robustness of the volumetric feature computed from each ROI, in Section 9.2.4.2 each ROI is further refined by picking only voxels with reasonable representation power. Finally, to show the consistency and difference of ROIs obtained in all templates, in Section 9.2.4.3 some analysis is provided to demonstrate the capability of the feature extraction method in extracting the complementary features from multiple templates for representing each subject brain.

9.2.4.1 Watershed segmentation

For robust feature extraction, it is important to group voxel-wise morphometric features into regional features. Voxel-wise morphometric features (such as the Jacobian determinants, voxel-wise displacement fields, and tissue density maps) usually have very high feature dimensionality, which includes a large amount of redundant/irrelevant information as well as noises that are due to registration errors.

On the other hand, using regional features can alleviate the above issues and thus provide more robust features in classification.

A traditional way to obtain regional features is to use prior knowledge, that is, predefined ROIs, which summarizes all voxel-wise features in each predefined ROI. However, this method is inappropriate in the case of using multiple templates for complementary representation of brain images, since in this way ROI features from multiple templates will be very similar (we use the volume-preserving measurement to calculate the template-specific morphometric pattern of tissue density change within the same ROI w.r.t. each different template). To capture different sets of distinctive brain features from different templates, a clustering method (Fan et al., 2007) is adopted for adaptive feature grouping. Since clustering will be performed on each template space separately, the complementary information from different templates can be preserved for the same subject image. As indicated in Fan et al. (2007), the clustering algorithm can improve the discriminative power of the obtained regional features, and reduce the negative impacts from registration errors.

Let $I_i^k(u)$ denote a voxel-wise tissue density value at voxel u in the kth template for the ith training subject, $i \in [1, N]$. The ROI partition for the kth template is based on the combined discrimination and robustness measure, $\mathrm{DRM}^k(u)$, computed from all N training subjects, which takes into account both feature relevance and spatial consistency as defined below:

$$\mathrm{DRM}^k(u) = P^k(u)C^k(u), \tag{9.1}$$

where $P^k(u)$ is the voxel-wise Pearson correlation (PC) between tissue density set $\{I_i^k(u), i \in [1, N]\}$ and label set $\{y_i \in [-1, 1], i \in [1, N]\}$ (1 for AD and -1 for NC) from all N training subjects, and $C^k(u)$ denotes the spatial consistency among all features in the spatial neighborhood (Fan et al., 2007).

Watershed segmentation is then performed on each calculated DRM^k map for obtaining the ROI partitions for the kth template. Note that, before applying watershed segmentation, we use a Gaussian kernel to smooth each map DRM^k, to avoid any possible oversegmentation, as also suggested in Fan et al. (2007). As a result, for example, we can partition the kth template into totally R^k nonoverlapping regions, $\{r_l^k, l \in [1, R^k]\}$, with each region $r_l{}^k$ owning U_l^k voxels. It is worth noting that each template will yield its own unique ROI partition, since different tissue density maps (of same subject) are generated in different template spaces.

Fig. 9.4 shows the partition results obtained from the same group of images registered to the two different templates. It is clear that the obtained ROIs are very different, in terms of both their structures and discriminative powers (as indicated by different colors). Those differences will naturally guide the subsequent steps of feature extraction and selection, and thus provide the complementary information to represent each subject and also improve its classification.

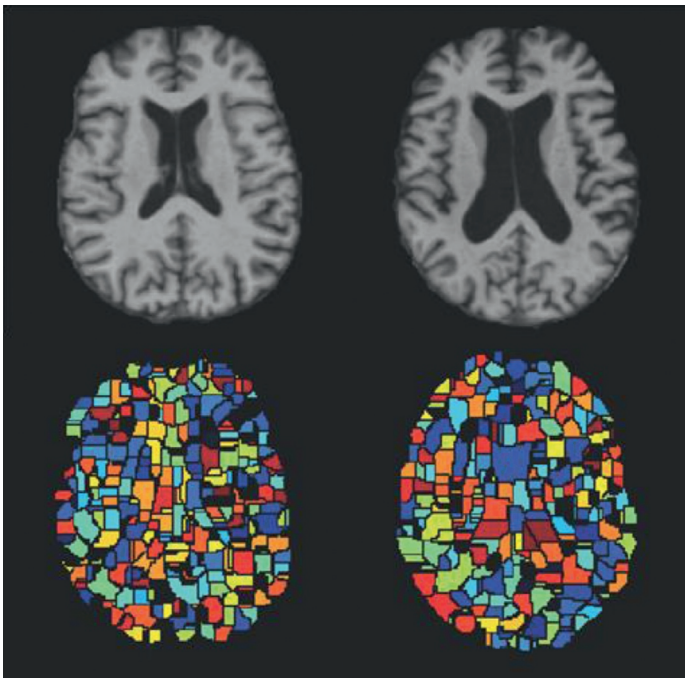

FIG. 9.4

Watershed segmentation of the same group of subjects on two different templates. Color indicates the discriminative power learned from the group of subjects (with the hotter color denoting more discriminative regions). Upper row: two different templates. Lower row: the corresponding partition results.

9.2.4.2 Regional feature aggregation

Instead of using all U_l^k voxels in each region r_l^k for total regional volumetric measurement, only a subregion $\tilde{r}_l^k$ in each region r_l^k is aggregated to further optimize the discriminative power of the obtained regional feature, by employing an iterative voxel selection algorithm. Specifically, one first selects a most relevant voxel, according to the PC calculated between this voxel's tissue density values and class labels from all N training subjects. Then the neighboring voxels are iteratively included to increase the discriminative power of all selected voxels, until no increase is found when adding new voxels. Note that this iterative voxel selection process will finally lead to a voxel set (called the optimal subregion) $\tilde{r}_l^k$ with $\tilde{U}_l^k$ voxels, which are selected from the region r_l^k. In this way, for a given subject i, its lth regional feature $V_{i,l}^k$ in the region $\tilde{r}_l^k$ of the kth template can be computed as

$$V_{i,l}^k = \sum_{\forall u \in \tilde{r}_l^k} \frac{I_i^k(u)}{\tilde{U}_l^k}. \tag{9.2}$$

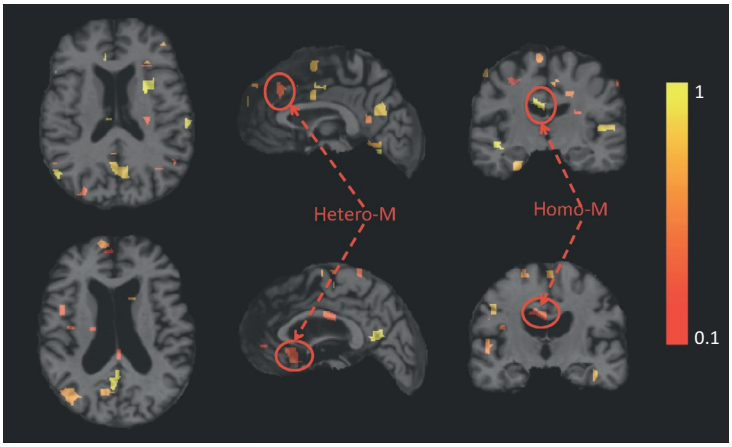

FIG. 9.5

Illustration of the top 100 regions identified using the regional feature aggregation scheme, where the same subject is registered to two different templates. The axial, sagittal, and coronal views of the original MR image of the subject after warping to each of the two different templates are displayed. Color indicates the discriminative power of the identified region (with the hotter color denoting more discriminative region). Upper row: image registered to template 1. Lower row: image registered to template 2. (For the definitions of both hetero-M and homo-M, please refer to Section 9.2.4.3.)

Each regional feature is then normalized to have zero mean and unit variance, across all N training subjects. Finally, from each template, M (out of R^k) most discriminative features are selected using their PC. Thus for each subject, its feature representation from all K templates consists of $M \times K$ features, which will be further selected for classification. Fig. 9.5 shows the top 100 regions selected using the regional feature aggregation scheme, for the same image registered to two templates (as shown in Fig. 9.4). It clearly shows the structural and discriminative differences of regional features from different templates.

9.2.4.3 Anatomical analysis

It is important to understand how the identified regions (ROIs) from different templates are correlated with the target brain abnormality (ie, AD), in order to better reveal the advantages of using multiple templates for morphometric pattern analysis in comparison to using only a single template. Accordingly, we categorize the identified regions (ROIs) into two classes: (1) the class with homogeneous measurements (*homo-M*) and (2) the class with heterogeneous measurements (*hetero-M*) (see Fig. 9.5). The homo-M refers to the regions that are simultaneously identified from different templates, whereas the hetero-M refers to the regions identified in a certain

template but not in other templates. In Fig. 9.5, it can be observed that a region within the left corpus callosum is identified in both templates 1 and 2 (see the coronal view). On the other hand, a region within the frontal lobe is only identified in template 1, and a region within the temporal lobe is only identified in template 2 (see the sagittal view). When jointly considering all identified regions from different templates in the classification, the integration of homo-M features is helpful to improve both robustness and generalization of feature extraction for the unseen subjects, while the combination of hetero-M features can provide complementary information for distinguishing subjects during the classification.

9.3 MULTIVIEW LEARNING METHODS FOR AD DIAGNOSIS

9.3.1 FEATURE FILTERING-BASED MULTIVIEW LEARNING

Although the most representative regional features are selected from each template, many regional features, after combination with other features from other templates, could be redundant or even deteriorate the classification of unseen subjects. Therefore selecting a subset of robust regional features (from all templates) is an essential step to achieve good classification performance. Min et al. (2014a) develop a feature filtering-based method to make use of those multiview feature representations.

It has been demonstrated via Fig. 9.5 that the regional features identified from different templates could be heterogeneous. Therefore selecting features jointly from multiple templates can potentially aggregate complementary information that is helpful for the classification. Specifically, for the N training images that have been registered to K templates, all features extracted from K templates can be denoted as $V = \{v_{n,m}^k, m \in [1, M], k \in [1, K], n \in [1, N]\}$, where M top selected features are extracted *independently* from each template by using the method described in Section 9.2.4.2. For each subject, that is, the nth subject, its feature vector $V_n = \{v_{n,m}^k, m \in [1, M], k \in [1, K]\}$ has in total $\xi = M \times K$ features. The goal is to select the top T features out of ξ features to gather the most discriminative and robust information jointly from all templates. The detail of selecting the top M features is provided in the following paragraph.

Because the regional features extracted from different templates are finally used for the same classification task, a "good" feature should be agreed not only by one template, but also by the other templates. In other words, a "good" feature selected from one template should strongly correlate to the "good" features selected from the other templates. Meanwhile, features that are helpful for classification should also strongly correlate with the training labels. To this end, for feature selection (FS), we propose to maximize both the feature relevance w.r.t. labels (ie, according to the PC), and the correlation with features from other templates. This can be done by introducing the "intertemplate" correlation ψ, and combining it with the PC ω by imposing a balancing factor λ as follows:

$$\Delta_m^k = \omega_m^k + \lambda \psi_m^k, \tag{9.3}$$

where Δ_m^k indicates the importance of the mth feature computed from the kth template. The FS can then be achieved by ranking this feature importance for all $\xi = M \times K$ features, $\{\Delta_m^k, m \in [1, M], k \in [1, K]\}$. In Eq. (9.3), ω_m^k denotes the PC between the mth feature from the tth template and the class label from all training subjects. Similarly, the "intertemplate" correlation ψ_m^k can be obtained by first computing the correlation between this mth feature in the kth template and each feature in *other* templates, and then integrating all these correlation coefficients (via summation and normalization) as the final measure. By using the above scheme, one can select a total of M top features with the highest feature importance values.

9.3.2 MAXIMUM-MARGIN-BASED REPRESENTATION LEARNING

The high-dimensional representations generated from multiple templates in their original spaces can form a low-dimensional manifold, in which the optimal representation for classification might be neither a representation generated from one of the existing templates nor the average representation located at the manifold centroid. Instead, the optimal representation could lie somewhere within the manifold of representations from multiple templates, which is most discriminative for the classification. Accordingly, Min et al. (2014b) propose a maximum-margin-based representation learning (MMRL) method to learn the optimal representation from multiple templates for AD classification, which can not only reduce the negative impact due to registration errors but also aggregate the complementary information captured from different template spaces. First, multiple templates are selected to serve as unique common spaces based on affinity propagation (Frey and Dueck, 2007). Then each studied subject is nonlinearly registered to the selected templates, and multiple representations from different template spaces are further generated by an autonomous feature extraction algorithm (Fan et al., 2007). Afterwards, the optimal representation from multiple representations (of multiple templates) in conjunction with the learning of a support vector machine (SVM) (Cortes and Vapnik, 1995) is learned based on the maximum-margin criteria. Finally, the learned representation and SVM are used for classification. Unlike traditional methods enforcing a prior in the representation learning (eg, variance maximization in PCA-based dimensionality reduction (DR) (Jolliffe, 2002), or the locality-preserving property in Laplacian score (LS)-based FS (He et al., 2005), which is independent from the classification stage), the MMRL method learns both the optimal representation and the classifier jointly, in order to make the two different tasks consistently conform to the same classification objective.

Fig. 9.6 illustrates the main idea, where a subject is first nonlinearly registered to multiple templates. Volumetric features are then extracted within each template space, so that multiple representations are generated from different templates. Based on the representations obtained, an optimal representation is finally learned to maximize the classification accuracy. To this end, an MMRL method is introduced to jointly learn both the optimal representation and the classifier for AD classification.

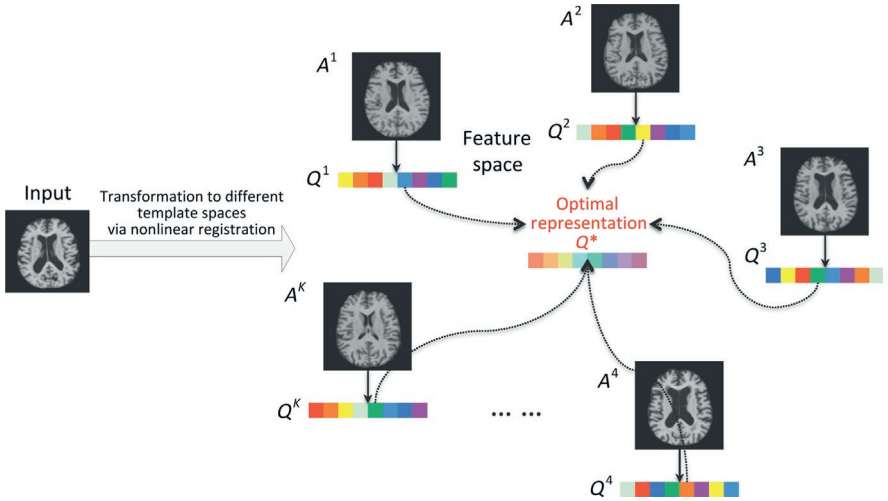

FIG. 9.6

Framework of the MMRL method: learning an optimal representation (Q^*) from the representations ($Q^1 - Q^K$) generated in multiple template spaces ($A^1 - A^K$).

In the multiview feature extraction method introduced in Section 9.2, it is assumed that different templates can then be used to capture complementary information for the same subject, by performing feature extraction in each individual template space. Given the set of representations of a subject generated from K different templates $\boldsymbol{X} = \{\boldsymbol{x}^k \in \mathbb{R}^M, k \in [1, K]\}$, we want to find a new representation $\boldsymbol{x}^* \in \mathbb{R}^L$, which can yield the best classification result. Suppose that the new representation can be generated by applying a mapping to the set of original representations as

$$\boldsymbol{x}^* = f(\boldsymbol{X}). \tag{9.4}$$

The goal is to learn the optimal mapping function $f(\cdot)$ which can yield the best representation $\boldsymbol{x}^*$ for classification. To achieve this goal, we propose an MMRL method to learn $f(\cdot)$ in conjunction with the learning of an SVM classifier, where the jointly learned mapping and classifier are both optimal for the targeted classification task.

Given a training set $\{(\boldsymbol{x}_i, y_i),\ i \in [1, N]\}$, where $\boldsymbol{x}_i \in \mathbb{R}^M$ and $y_i \in \{-1, 1\}$ denote the feature vector and label of the ith subject, respectively, a soft-margin SVM tries to find a hyperplane that maximizes the margin between two classes of samples and also minimizes the cost of misclassification:

$$\arg\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||_2^2 + C\sum_{i=1}\left[1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)\right]_+, \tag{9.5}$$

where $\{\boldsymbol{w}, b\}$ defines the SVM hyperplane, $[\cdot]_+$ denotes the hinge loss function, and C is the balancing factor between the hinge loss and the margin regularization.

When the feature vector $\tilde{x}$ of an unknown subject is input, its associated label $\tilde{y}$ can be predicted using the learned hyperplane $\{\boldsymbol{w}, b\}$ as

$$\tilde{y} = \text{sign}\left(\boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}} + b\right). \tag{9.6}$$

Given a training set $\{(\boldsymbol{X}_i, y_i), i \in [1, N]\}$, where $\boldsymbol{X}_i = \left\{x_i^k, k \in [1, K]\right\}$ is the set of representations generated from all K templates, and $\boldsymbol{x}_i^k \in \mathbb{R}^M$ denotes one representation extracted from the kth template for the ith subject. In order to learn the optimal representation $\boldsymbol{x}^*$ (ie, learning the optimal mapping function $f(\cdot)$) jointly with the classification model as defined in Eq. (9.5), the mapping to the new representation is first defined as a linear combination of the K existing representations generated from different templates:

$$f\left(\boldsymbol{X}_i | P^k, \forall k\right) = \sum_{k=1}^{K} P^k \boldsymbol{x}_i^k, \tag{9.7}$$

where $P^k \in \mathbb{R}^{L \times L}$ is a diagonal coefficient matrix to assign different weights to different features of the kth representation (with all nondiagonal elements equal to zero). Then the goal is to find the optimal mapping $f\left(\cdot | P^k, \forall k\right)$ and hyperplane $\{\boldsymbol{w}, b\}$ that maximize the margin between different classes and also reduce the misclassification rate on the training set:

$$\begin{aligned} &\arg\min_{\boldsymbol{w}, b, \{P^k, \forall k\}} \frac{1}{2}||\boldsymbol{w}||_2^2 + C\sum_{i=1}\left[1 - y_i\left(\boldsymbol{w}^{\mathrm{T}} f\left(X_i | P^k, \forall k\right) + b\right)\right]_+ \\ &\text{s.t.} \quad \forall j, P_j^k > 0 \quad \text{and} \quad \sum_{k=1}^{K} P_j^k = 1, \end{aligned} \tag{9.8}$$

where P_j^k denotes the jth diagonal element of the coefficient matrix P^k (ie, the weight for the jth feature from the kth template). The constraints in Eq. (9.8) confine the estimated weights into the first quadrant of a unit square, so that the generated representation lies within the polygon in the manifold of original representations.

To avoid overfitting, features are further partitioned into different groups, where features within the same group will be assigned to the same mapping weights (ie, $P_{j_1}^k = P_{j_2}^k$ if the j_1th and j_2th features are in the same group). Introducing this additional constraint can efficiently reduce the degree of freedom of the proposed model, thus achieving improved generalization with limited training samples. The feature grouping strategy used in this chapter is implemented by performing affinity propagation on the feature covariance matrix calculated from the training set.

To optimize Eq. (9.8), one can adopt the coordinate descent method to estimate the parameters. The mapping weights $\left\{P^k, \forall k\right\}$ and the hyperplane $\{\boldsymbol{w}, b\}$ are optimized in an iterative manner. In each iteration, one term is optimized while the other is fixed, and thus each optimization step is convex. With the learned mapping $f\left(\cdot | P^k, \forall k\right)$ and the learned decision boundary $\{\boldsymbol{w}, b\}$, given the feature vectors $\tilde{\boldsymbol{X}}$ of

an unknown sample extracted from multiple templates, the associated label $\tilde{y}$ can be predicted as

$$\tilde{y} = \text{sign}\left(\boldsymbol{w}^{\mathrm{T}} f\left(\boldsymbol{X}_i | P^k, \forall k\right) + b\right). \tag{9.9}$$

9.3.3 VIEW-CENTRALIZED MULTIVIEW LEARNING

Given multiview feature representation, one can observe that a representation generated from a specific template can be regarded as a profile for the brain, and can be used to provide supplementary or side information for other representations generated from the other templates (ie, views). Accordingly, Liu et al. (2015) develop a view-centralized multitemplate (VCM) classification method, with flowchart illustrated in Fig. 9.7.

As can be seen from Fig. 9.7, brain images are first nonlinearly registered to multiple templates individually, and then their volumetric features are extracted within each template space. In this way, multiple feature representations can be generated

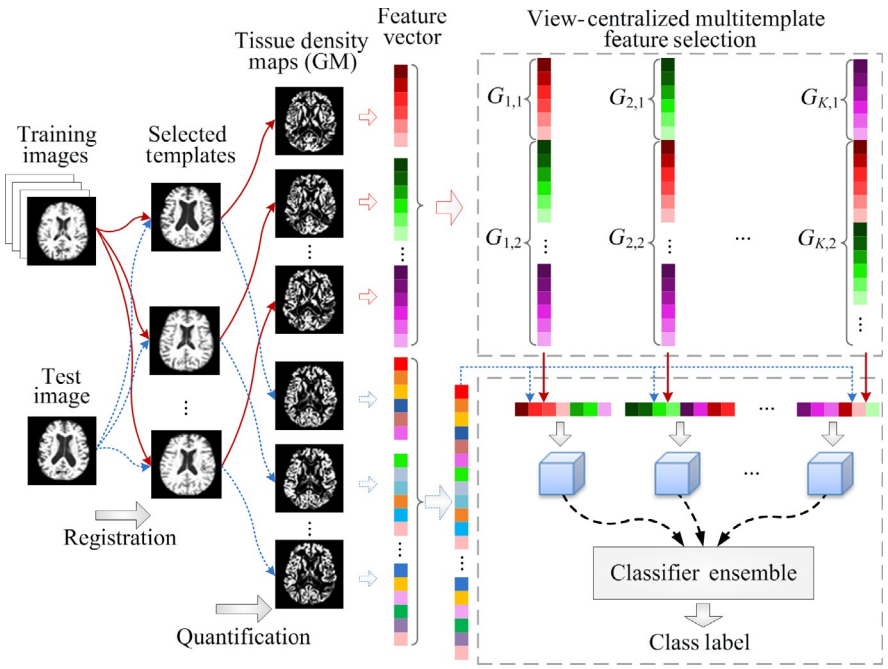

FIG. 9.7

The framework of the view-centralized multitemplate classification method, which includes four main steps: (1) preprocessing and template selection, (2) feature extraction, (3) feature selection, and (4) ensemble classification.

from different templates for each specific subject. Based on such representations, the proposed VCM FS method can be applied to select the most discriminative features, by focusing on the main-view template along with the extra guidance from side-view templates. Finally, multiple SVM classifiers are constructed based on multiple sets of selected features, followed by a classifier ensemble strategy to combine multiple outputs from all SVM classifiers for making the final decision.

Given N training images that have been registered to K templates, we denote $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{N} \in \mathbb{R}^{D\times N}$ ($D = M \times K$ in this chapter) as the training data, where $\boldsymbol{x}_i \in \mathbb{R}^{D}$ is the feature representation generated from K templates for the ith training image. Let $\boldsymbol{Y} = \{y_i\}_{i=1}^{N} \in \mathbb{R}^{N}$ be the class labels of N training data, and $\boldsymbol{w} \in \mathbb{R}^{D}$ be the weight vector for the FS task. For clarity, we divide the feature representations from multiple templates into a main-view group and a side-view group, as illustrated in Fig. 9.8. As can be seen from Fig. 9.8, the main-view group (corresponding to the main template) contains features from a certain template, while the side-view group (corresponding to other supplementary templates) contains features from all other (supplementary) templates.

Denote $a^{(1)}$ as the weighting value for the main-view (ie, main template) group and $a^{(2)}$ as the weighting value for the side-view (ie, supplementary templates) group. By setting different weighting values for features from the main view and the side views, we can incorporate the prior information into the following learning model:

$$\begin{aligned} &\min_{\boldsymbol{w}} \frac{1}{2N}\sum_{i=1}^{N}||y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i||_2^2 + \lambda_1||\boldsymbol{w}||_1 + \lambda_2\sum_{g=1}^{2}a^{(g)}||\boldsymbol{w}^{(g)}||_2 \\ &\text{s.t.} \sum_{g=1}^{2}a^{(g)} = 1;\ a^{(g)} > 0,\ g = 1,2, \end{aligned} \tag{9.10}$$

where $\boldsymbol{w}^{(g)}$ represents the weight vector for the gth group. The first term in Eq. (9.10) is the empirical loss on the training data, and the second one is the l_1-norm regularization term that enforces some elements of $\boldsymbol{w}$ to be zero. It is worth noting

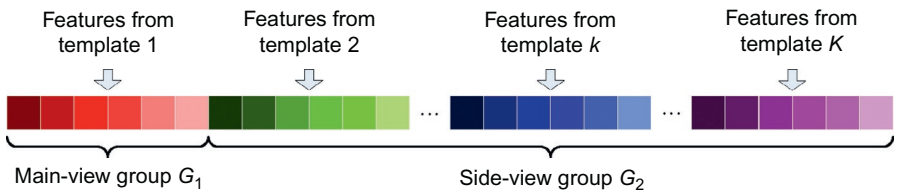

FIG. 9.8

Illustration of group information for feature representations generated from multiple templates. The first group G_1 (ie, the main-view group) consists of features from a certain template, while the second group G_2 (ie, the side-view group) contains features from all other (supplementary) templates.

that the last term in Eq. (9.10) is a view-centralized regularization term, which treats features in the main-view group and the side-view group differently by using different weighting values (ie, $a^{(1)}$ and $a^{(2)}$). For example, a small $a^{(1)}$ (as well as a large $a^{(2)}$) implies that the coefficients for features in the main-view group will be penalized lightly, while features in the side-view group will be penalized severely, because the goal of the model defined in Eq. (9.10) is to minimize the objective function. Accordingly, most elements in the weight vector corresponding to the side-view group will be zero, while those corresponding to the main-view group will not. In this way, the prior knowledge that one focuses on the representation from the main template (ie, main view) with extra guidance from other templates can be incorporated into the learning model naturally. In addition, two constraints in Eq. (9.10) are used to ensure that the weighting values for different groups are greater than 0 and not greater than 1. By introducing such constraints, one can efficiently reduce the degrees of freedom of the proposed model, and avoid overfitting with limited training samples.

Based on the VCM FS model defined in Eq. (9.10), one can obtain a feature subset by selecting features with nonzero coefficients in $\boldsymbol{w}$. Each time, one performs the above-mentioned FS procedure by focusing on one of multiple templates, with other templates used as extra guidance. Accordingly, given K templates, one can get K selected feature subsets, with each of them reflecting the information learned from a certain main template and corresponding supplementary templates.

Ensemble classification

After obtaining K feature subsets by using the view-centralized FS algorithm, one can then learn K base classifiers individually. In this study, a linear SVM classifier is used to identify AD patients from NCs, and progressive MCI patients from stable MCI patients, since the linear SVM model has good generalization capability across different training data, as shown in extensive studies (Zhang and Shen, 2012; Burges, 1998; Pereira et al., 2009). Finally, a classifier ensemble strategy is used to combine these K base classifiers to construct a more accurate and robust learning model, where the majority voting strategy is employed for the fusion of multiple classifiers. Thus the class label of an unseen test sample can be determined by majority voting for the outputs of base classifiers.

9.3.4 RELATIONSHIP-INDUCED MULTIVIEW LEARNING

The main limitation of existing multiview learning models is that only the relationship between samples and their corresponding class labels is considered. Actually, there exist some other important structure information in multiview feature representation using multitemplate MR imaging data, for example, (1) the relationship among multiple templates and (2) the relationship among different subjects. Accordingly, (Liu et al., 2016) develop a relationship-induced multitemplate learning (RIML) method to explicitly model the relationships among templates and among subjects. The flowchart of the RIML method is provided in Fig. 9.9. As can be seen, there are

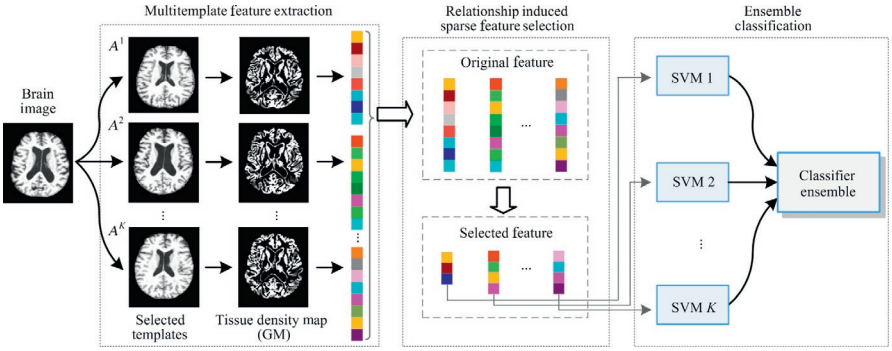

FIG. 9.9

The framework of the relationship-induced multitemplate learning (RIML) method, which consists of three main steps: (1) multitemplate feature extraction, (2) feature selection, and (3) ensemble classification.

three main steps in the RIML method: (1) feature extraction, (2) feature selection, and (3) ensemble classification.

To model the relationships among templates and among subjects, a *relationship-induced sparse* (RIS) FS method is proposed under the multitask learning framework (Argyriou et al., 2008; Zhang and Shen, 2012), by treating the classification in each template space as a specific task. In this chapter, we have K learning tasks corresponding to K templates. Denote $\boldsymbol{X}^k = [\boldsymbol{x}_1^k, \ldots, \boldsymbol{x}_n^k, \ldots, \boldsymbol{x}_N^k]^{\mathrm{T}} \in \mathbb{R}^{N \times M}$ as training data for the kth learning task (corresponding to the kth template) containing a total of N subjects, where $\boldsymbol{x}_n^k \in \mathbb{R}^M$ represents a feature vector of the nth subject in the kth template space ($n = [1, N]$). Similarly, denote $\boldsymbol{Y} = [y_1, \ldots, y_n, \ldots, y_N]^{\mathrm{T}} \in \mathbb{R}^N$ as the response vector for training data $\boldsymbol{X}^k$, where $y_n \in \{-1, 1\}$ is the class label (ie, NC or patient) for the nth subject. Let $\boldsymbol{W} = \left[\boldsymbol{w}^1, \ldots, \boldsymbol{w}^k, \ldots, \boldsymbol{w}^K\right] \in \mathbb{R}^{M \times K}$ represent the weight vector matrix, where $\boldsymbol{w}^k \in \mathbb{R}^M$ parameterizes a linear discriminant function for the kth task ($k = [1, K]$). Then, the multitask feature learning model can be formulated by solving the following objective function (Zhang and Shen, 2012; Caruana, 1997; Baxter, 1997):

$$\min_{\boldsymbol{W}} \sum_{k=1}^{K} ||\boldsymbol{Y} - \boldsymbol{X}^k \boldsymbol{w}^k||_2^2 + \lambda ||\boldsymbol{W}||_{2,1}. \tag{9.11}$$

The first term in Eq. (9.11) is the empirical loss on the training data. The second one is a group-sparsity regularizer to encourage the weight matrix $\boldsymbol{W}$ to have many zero rows, where $||\boldsymbol{W}||_{2,1}$ is the sum of the l_2-norm of the rows in matrix $\boldsymbol{W}$. For FS purposes, only those features corresponding to those rows with nonzero coefficients in $\boldsymbol{W}$ are selected, after solving Eq. (9.11). That is, the $l_{2,1}$-norm regularization term ensures that only a small number of common features are jointly selected across different tasks. The parameter λ is a regularization parameter that is used to balance

the relative contributions of those two terms in Eq. (9.11). Specifically, a large λ means that a lower number of features will be selected, while a small λ denotes more features will be selected.

It is worth noting that, due to the anatomical differences across templates, different sets of features obtained for each brain image generally come from different ROIs. Thus the $l_{2,1}$-norm regularization in Eq. (9.11) is not appropriate for the case of using multiple templates, since it jointly selects features across different tasks (ie, templates). To encourage the sparsity of the weight matrix $\boldsymbol{W}$, as well as to select the most informative features corresponding to each template space, we propose the following multitask sparse feature learning model:

$$\min_{\boldsymbol{W}} \sum_{k=1}^{K} ||\boldsymbol{Y} - \boldsymbol{X}^k \boldsymbol{w}^k||_2^2 + \lambda ||\boldsymbol{W}||_{1,1}, \tag{9.12}$$

where $||\boldsymbol{W}||_{1,1}$ is the sum of l_1-norm of the rows in matrix $\boldsymbol{W}$. It is worth noting that the $l_{1,1}$-norm does not necessarily ensure many rows in $\boldsymbol{W}$ are zero, but can help select features that are discriminative for specific tasks.

In Eqs. (9.11) and (9.12), a linear mapping function (ie, $f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}$) is learned to transform data in the original high-dimensional feature space to a one-dimensional label space. The main limitation of these models is that only the relationship between samples and their corresponding class labels is considered. Actually, there exists some important structure information in the multitemplate data, for example, (1) the relationship among multiple templates (*template relationship*) and (2) the relationship among different subjects (*subject relationship*).

(1) As illustrated in Fig. 9.10A, a subject $\boldsymbol{x}_n$ is represented as $\boldsymbol{x}_n^{k_1}$ and $\boldsymbol{x}_n^{k_2}$ in the k_1th template space and in the k_2th template spaces, respectively. After being mapped to the label space, they should be close to each other (ie, $f(\boldsymbol{x}_n^{k_1})$ should be similar to $f(\boldsymbol{x}_n^{k_2})$), since they represent the same subject.

(2) Similarly, as shown in Fig. 9.10B, if two subjects $\boldsymbol{x}_{n_1}^k$ and $\boldsymbol{x}_{n_2}^k$ in the same kth template space are very similar, the distance between $f(\boldsymbol{x}_{n_1}^k)$ and $f(\boldsymbol{x}_{n_2}^k)$ should be small, implying that estimated labels of these two subjects are similar.

Accordingly, a novel *template relationship*-induced regularization term is defined as follows:

$$\begin{aligned} &\sum_{n=1}^{N} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \left(f(\boldsymbol{x}_n^{k_1}) - f(\boldsymbol{x}_n^{k_2}) \right)^2 \\ &= \sum_{n=1}^{N} tr((\boldsymbol{B}_n \boldsymbol{W})^{\mathrm{T}} \boldsymbol{L}_n (\boldsymbol{B}_n \boldsymbol{W})), \end{aligned} \tag{9.13}$$

where $tr(\cdot)$ denotes the trace of a square matrix, $\boldsymbol{B}_n = [\boldsymbol{x}_n^1, \ldots, \boldsymbol{x}_n^k, \ldots, \boldsymbol{x}_n^K]^{\mathrm{T}} \in \mathbb{R}^{K \times M}$ represents multiple sets of features derived from K templates for the nth subject,

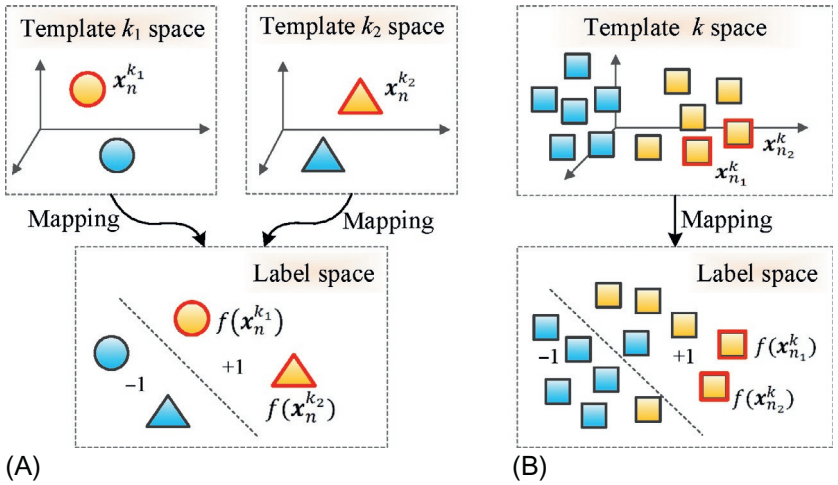

FIG. 9.10

Illustration of structure information, conveyed by: (A) relationship between features of two templates (ie, features of the nth subject in the k_1th template space and the k_2th template space, respectively), and (B) relationship between features of two subjects in the same template (ie, features of the n_1th subject and the n_2th subject in the kth template space). Here, yellow denotes positive training subjects, while blue denotes negative training subjects. Different shapes (circle, triangle, and square) denote samples in three different template spaces (ie, the k_1th template, k_2th template, and kth template).

and $\boldsymbol{L}_n \in \mathbb{R}^{K\times K}$ is a matrix with diagonal elements being $K-1$ and all other elements being -1. By using Eq. (9.13), we can model the relationship among multiple templates explicitly.

Similarly, the *subject relationship*-induced regularization term is defined as follows:

$$\sum_{k=1}^{K}\sum_{n_1=1}^{N}\sum_{n_2=1}^{N} S_{n_1,n_2}^k \left(f\left(\boldsymbol{x}_{n_1}^k\right)-f\left(\boldsymbol{x}_{n_2}^k\right)\right)^2 = \sum_{k=1}^{K}\left(\boldsymbol{X}^k\boldsymbol{w}^k\right)^{\mathrm{T}}\boldsymbol{L}^k\left(\boldsymbol{X}^k\boldsymbol{w}^k\right), \tag{9.14}$$

where $\boldsymbol{X}^k$ is the data matrix in the kth learning task (ie, kth template) as mentioned above, and $\boldsymbol{S}^k = \left\{S_{n_1,n_2}^k\right\}_{n_1,n_2=1}^N \in \mathbb{R}^{N\times N}$ denotes a similarity matrix with elements defining the similarity among N training subjects in the kth template space. Here, $\boldsymbol{L}^k = \boldsymbol{D}^k - \boldsymbol{S}^k$ represents the Laplacian matrix for task k, where $\boldsymbol{D}^k$ is a diagonal matrix with diagonal element $D_{n_1,n_1}^k = \sum_{n_2=1}^N S_{n_1,n_2}^k$, and S_{n_1,n_2}^k is defined as

$$S^k_{n_1,n_2} = \begin{cases} e^{-\frac{||x^k_{n_1} - x^k_{n_2}||^2}{\sigma}}, & \text{if } \boldsymbol{x}^k_{n_1} \text{ and } \boldsymbol{x}^k_{n_2} \text{ are } q \text{ neighbors,} \\ 0, & \text{otherwise,} \end{cases} \tag{9.15}$$

where σ is a constant to be set, and q is set as 3 empirically. It is easy to see that Eq. (9.14) aims to preserve the local neighboring structure of the original data during the mapping, through which one can capture the relationship among different subjects explicitly.

By incorporating two relationship-induced regularization terms defined in Eqs. (9.13) and (9.14) into Eq. (9.12), the objective function of the RIS FS model can be obtained as follows:

$$\begin{aligned} \min_{\boldsymbol{W}} \sum_{k=1}^{K} ||\boldsymbol{Y} - \boldsymbol{X}^k \boldsymbol{w}^k||_2^2 + \lambda_1 ||\boldsymbol{W}||_{1,1} + \lambda_2 \sum_{n=1}^{N} tr\left((\boldsymbol{B}_n \boldsymbol{W})^{\mathrm{T}} \boldsymbol{L}_n (\boldsymbol{B}_n \boldsymbol{W}) \right) \\ + \lambda_3 \sum_{k=1}^{K} \left(\boldsymbol{X}^k \boldsymbol{w}^k \right)^{\mathrm{T}} \boldsymbol{L}^k \left(\boldsymbol{X}^k \boldsymbol{w}^k \right), \end{aligned} \tag{9.16}$$

where λ_1, λ_2, and λ_3 are positive constants used to balance the relative contributions of the four terms in the proposed RIS model, and their values can be determined via inner cross-validation on training data. In Eq. (9.16), the $l_{1,1}$-norm regularization term (the second term) ensures only a small number of features to be selected for each task. The *template relationship*-induced regularization term (the third term) is used to capture the relationship among different templates, while the *subject relationship* regularization term (the fourth term) is employed to preserve the local neighboring structure of data in each template space. After FS using the proposed RIS FS algorithm, an ensemble classification process (similar to Section 9.3.3) is adopted for making a final decision for a test subject.

9.4 EXPERIMENTS

9.4.1 SUBJECTS

The Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.ucla.edu) (Jack et al., 2008) is employed to evaluate the performance of the proposed classification algorithm. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early 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. Since we focus on the morphometric study of AD, T1-weighted MRI data from ADNI is used in the experiments. In total, 459 subjects, scanned with a 1.5T scanner,

Table 9.1 Demographic Information of the Studied Subjects From ADNI Database

Diagnosis	Number	Age	Gender (M/F)	MMSE
AD	97	75.90 ± 6.84	48/49	23.37 ± 1.84
NC	128	76.11 ± 5.10	63/65	29.13 ± 0.96
pMCI	117	75.18 ± 6.97	67/50	26.45 ± 1.66
sMCI	117	75.09 ± 7.65	79/38	27.42 ± 1.78

Note: Values are denoted as mean ± deviation; MMSE means mini-mental state examination; M and F represent male and female, respectively.

were randomly selected, comprising of 97 AD, 128 NC, and 234 MCI (117 pMCI and 117 sMCI) subjects. The demographic information of the used dataset is shown in Table 9.1.

As mentioned in Section 9.2, K ($K = 10$ in the experiments) representative templates are first selected from AD and NC subjects. In the multitemplate feature extraction stage, there are a total of M ($M = 1500$ in the experiments) features extracted in each template space. Thus for each subject, its feature representation from K templates consists of 1500×10 features, which will be further selected for classification.

9.4.2 EXPERIMENTAL SETTINGS

The evaluation of different methods is conducted on two different problems: (1) AD diagnosis such as AD versus NC classification and (2) progressive MCI diagnosis such as pMCI versus sMCI classification. Note that the second problem is considered more difficult than the first problem, but received relatively less attention in previous works. However, it is important to identify progressive MCI patients from stable MCI patients, in order to possibly prevent the progression of MCI to AD via timely therapeutic interventions.

In general, a 10-fold cross-validation strategy is adopted to evaluate the performances of different methods, which has been widely used in recent studies (Zhang and Shen, 2012; Burges, 1998; Pereira et al., 2009). Specifically, all samples are partitioned into 10 subsets (with each subset having a roughly equal size), and *each time* samples in one subset are successively selected as the test data, while those in all other nine subsets are used as the training data to perform FS and classifier construction. This process is repeated 10 times independently to avoid any bias introduced by the random partitioning of the original data in the cross-validation process. Finally, the mean values of corresponding classification results are recorded for comparison.

The performance of different methods is evaluated via four evaluation criteria, that is, classification accuracy (ACC), classification sensitivity (SEN), classification specificity (SPE), and the area under the receiver operating characteristic (ROC)

curve (AUC). More specifically, the accuracy measures the proportion of subjects that are correctly predicted among all studied subjects, the sensitivity denotes the proportion of patients that are correctly predicted, and the specificity represents the proportion of NCs that are correctly predicted.

9.4.3 RESULTS OF FEATURE FILTERING-BASED METHOD FOR AD/MCI DIAGNOSIS

In this group of experiments, the balancing factor λ in Eq. (9.3) is set to 0.38. The SVM classifier used here is implemented by the LIBSVM library (Chang and Lin, 2011), using a linear kernel and $C = 1$ (the default cost). Finally, $M = 1{:}1500$ features are tested, and the best results are reported for quantitative comparison.

Table 9.2 first shows the results using a single template for AD/NC classification, to demonstrate the variability of classification results when using different templates even for the same classification task, where the best results are marked in boldface. Because the proposed FS method integrates not only the PC but also the "intertemplate" correlation from the multiple templates, two conventional FS methods are examined based on single templates. The first FS method is simply based on the ranking of PC, and the second method combines PC with SVM-RFE-based FS (Guyon et al., 2002) (as proposed in Fan et al. (2007)) for jointly considering multiple features in the selection. It should be noted that, in the single template case, the feature extraction performed in the proposed method is the same as COMPARE (Fan et al., 2007). Therefore in this chapter, the PC+SVM-RFE-based method using a single template is denoted as COMPARE.

Table 9.2 reports the best classification accuracies (ACC) for each of the 10 templates using PC and COMPARE, along with their respective sensitivities (SEN) and specificities (SPEC). Note that the sensitivity and the specificity refer to the portions of correctly identified AD patients and correctly classified NC subjects, respectively. From Table 9.2, it is clear that COMPARE outperforms PC when using their own best templates (ie, A^5 for PC and A^7 for COMPARE). However, for some templates (ie, A^1, A^2, A^5, A^9, and A^{10}), the use of additional SVM-RFE-based FS (in COMPARE) cannot further improve the simple PC-based classification (in terms of the best classification accuracy). That is, the result improvement brought by SVM-RFE is limited, but at a cost of increased computational burden.

Furthermore, the results of AD versus NC and pMCI versus sMCI classification using multiple templates are given in Table 9.3. The proposed (multitemplate-based) FS method (namely MA_Proposed) that considers both PC and "intertemplate" correlation is compared with both PC- and COMPARE-based FS methods using either a single template (namely SA_PC and SA_COMPARE) or multiple templates (namely MA_PC and MA_COMPARE). For fair comparison, the averaged results of single-template-based methods (SA_PC and SA_COMPARE) across all 10 templates are reported. In MA_PC, all regional features extracted from 10 different templates are used, thus resulting in a feature representation with $M \times K = 15{,}000$ dimensions for each subject; afterwards, the top 1500 features are selected out

Table 9.2 Results of AD Versus NC and pMCI Versus sMCI Classification Using Single Templates (A^1–A^{10})

	AD vs. NC Classification						pMCI vs. sMCI Classification					
	PC			COMPARE			PC			COMPARE		
Template	**ACC**	**SEN**	**SPE**	**ACC**	**SEN**	**SPE**	**ACC**	**SEN**	**SPE**	**ACC**	**SEN**	**SPE**
A^1	84.09	78.33	88.40	83.16	75.33	89.17	68.93	64.62	73.18	71.03	68.79	73.18
A^2	84.94	80.56	88.30	81.95	73.67	88.40	68.87	68.56	69.09	71.46	71.97	70.76
A^3	83.12	77.33	87.56	84.50	78.44	89.17	69.34	65.15	73.41	69.81	69.47	70.08
A^4	84.87	80.44	88.33	85.72	82.22	88.40	**72.71**	73.56	71.82	71.82	**72.58**	71.06
A^5	**85.85**	82.56	88.46	84.05	76.22	90.00	70.66	69.39	71.82	71.93	71.21	72.80
A^6	84.38	78.33	**89.04**	85.35	**83.56**	86.73	71.04	65.98	**75.98**	72.86	69.62	76.14
A^7	82.23	77.22	86.09	**87.07**	81.33	**91.54**	71.08	**73.94**	68.18	**74.56**	70.8	**78.64**
A^8	83.59	79.44	86.86	84.48	79.44	88.46	70.27	68.71	71.67	71.88	68.56	75.00
A^9	83.65	77.33	88.40	82.27	78.44	85.38	68.55	66.36	70.68	71.10	66.97	75.15
A^{10}	83.28	**83.78**	83.01	83.20	76.56	88.46	69.00	72.05	65.83	71.74	70.15	73.41

Table 9.3 Results of AD Versus NC and pMCI Versus sMCI Classification Using Single Templates (SA_PC, SA_COMPARE, SA_Proposed) and Multiple Templates (MA_PC, MA_COMPARE, MA_Proposed)

	AD vs. NC			pMCI vs. sMCI		
Method	**ACC (%)**	**SEN (%)**	**SPE (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**
SA_PC	82.01	75.88	86.76	68.49	67.80	69.10
SA_COMPARE	81.52	77.11	84.92	70.06	68.08	72.02
MA_PC	85.91	81.56	89.23	72.78	74.62	70.91
MA_COMPARE	87.19	80.56	92.31	73.35	**75.76**	70.83
MA_Proposed	**91.64**	**88.56**	**93.85**	**72.41**	72.12	**72.58**

of 15,000 features based on the PC, and $M = 1{:}1500$ features are subsequently selected and used for classification. In MA_COMPARE, the top 1500 features are first selected in the same way as MA_PC, but additionally using SVM-RFE to further refine the selected features, before inputting them to the SVM for classification.

For both AD versus NC and pMCI versus sMCI classification, the best classification accuracies (ACC) as well as the corresponding sensitivities (SEN) and specificities (SPEC) of all methods are illustrated in Table 9.3. The results clearly show that MA_Proposed is better than any other methods in terms of all metrics. It should be noted that the sensitivities of SA_PC, SA_COMPARE, MA_PC, and MA_COMPARE are much lower in comparison to their corresponding specificities. A low sensitivity value indicates low confidence on AD diagnosis, which will greatly limit their practical usage. On the other hand, MA_Proposed gives a significantly improved sensitivity value. Together with its high specificity (93.85% for AD vs. NC classification), the MA_Proposed method produces more confident AD diagnosis results.

In addition, Fig. 9.11 illustrates the results of SA_PC, SA_COMPARE, MA_PC, MA_COMPARE, and MA_Proposed in AD versus NC and pMCI versus sMCI classification with respect to different numbers of top selected features. From Fig. 9.11, it is clear that the results of multitemplate-based methods (MA_PC, MA_COMPARE, and MA_Proposed) outperform the results of single-template-based methods (SA_PC and SA_COMPARE) by a significant margin. Specifically, in Fig. 9.11 (left), SA_PC and SA_COMPARE reach their best classification accuracy with a small portion of top selected features, and their performances decline rapidly when more features are included in AD versus NC classification. This indicates that many of their selected features are noisy and redundant, if using only a single template. In contrast, multitemplate-based methods consistently increase or maintain their performance with the increase of the number of features used, which demonstrates that the complementary information from different templates is aggregated together to improve the classification. In addition, with the assistance of SVM-RFE, the COMPARE-based methods (SA_COMPARE and MA_COMPARE) achieve

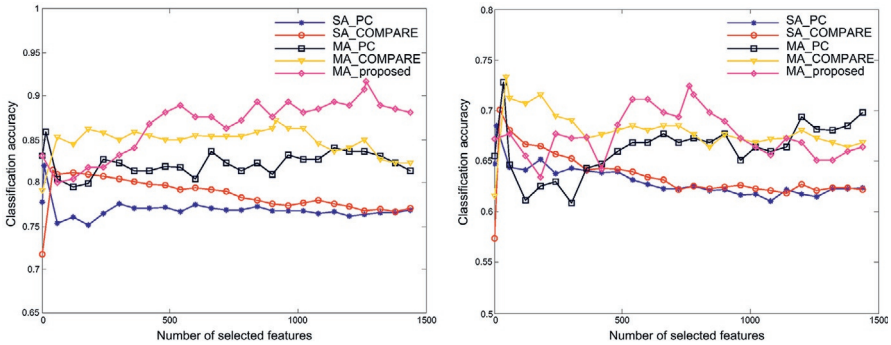

FIG. 9.11

Results of SA_PC, SA_COMPARE, MA_PC, MA_COMPARE, and MA_Proposed in (left) AD versus NC classification and (right) pMCI versus sMCI classification.

better performance than the PC-based methods (SA_PC and MA_PC) in both cases of using single template and multiple templates. Fig. 9.11 (left) also demonstrates that MA_Proposed significantly outperforms all other comparison methods. Although only a small portion of features can give good classification accuracy for the single-template-based methods, the performance of the MA_Proposed method is consistently improved with use of more features (ie, 91.64% when using 1268 features for AD versus NC classification). This phenomenon shows that the redundant features from a single template can be integrated with the features from other templates (in an effective way) to yield more robust and discriminative representations. From Fig. 9.11 (right), we can observe again that all three multitemplate-based methods (MA_PC, MA_COMPARE, and MA_Proposed) perform significantly better than the two single-template-based methods (SA_PC, SA_COMPARE) in pMCI versus sMCI classification, indicating the power of using multiple templates in aggregating more useful information for classification. Among all three multitemplate-based methods, MA_Proposed demonstrates comparable performance to both MA_PC and MA_COMPARE. When using the $M = 500{:}1000$ top selected features, the proposed method (MA_Proposed) gives the best overall classification results. On the other hand, MA_COMPARE gets its best results when using $M = 1{:}500$ features, and MA_PC achieves its best results when using $M = 1000{:}1500$ features.

9.4.4 RESULTS OF MAXIMUM-MARGIN-BASED LEARNING FOR AD/MCI DIAGNOSIS

In this group of experiments, the number of selected templates is $K = 10$ from affinity propagation, and the number of biomarkers identified on each template is $L = 20$. Table 9.4 compares the classification performance of the learned representation

Table 9.4 Comparison of MMRL to Representation Generated From Single Template (SA) and the Average Representation From Multiple Templates for AD/NC Classification and pMCI/sMCI Classification

	AD vs. NC			pMCI vs. sMCI		
Method	**ACC (%)**	**SEN (%)**	**SPE (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**
Mean_SA	83.23	82.28	84.06	67.56	70.30	67.71
Best_SA	85.35	82.33	87.69	71.08	72.88	69.02
Average	86.68	85.67	87.63	70.70	73.11	68.18
MMRL	**90.69**	**87.56**	**93.01**	**73.69**	**76.44**	**70.76**

(using MMRL from multiple templates) with single template (SA) representations and the average representation of multiple templates. The classification rate of the best template (Best SA) and the average result across the 10 templates (Mean SA) are also reported in Table 9.4. Additionally, the classification performance obtained by the average representation from multiple representations generated from all 10 templates is given in Table 9.4. From Table 9.4 it is clear that the representation learned by MMRL significantly outperforms both Best_SA and Average according to all evaluation metrics (accuracy, sensitivity, and specificity) for both AD/NC classification and pMCI/sMCI classification.

Table 9.5 shows the results comparing MMRL with four popular DR and FS methods when multiple templates are used. In Table 9.5, PCA (Jolliffe, 2002) and AutoEncoder (Bengio, 2009) are DR methods, whereas LS (He et al., 2005) and mRMR (Peng et al., 2005) are widely used FS techniques. For fair comparison, all techniques reduce the feature dimension to 20 (the same as the MMRL learned representation). For AutoEncoder, a widely used configuration with a three-layer architecture (Bengio, 2009) was adopted. These results demonstrate that the proposed joint learning method yields the best classification results (90.69% for AD vs. NC and

Table 9.5 Comparison of MMRL to Different Dimensionality Reduction and Feature Selection Methods for AD/NC Classification and pMCI/sMCI Classification

	AD vs. NC			pMCI vs. sMCI		
Method	**ACC (%)**	**SEN (%)**	**SPE (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**
PCA	83.08	81.22	84.55	68.96	72.12	68.68
AutoEncoder	87.60	86.56	88.40	66.42	70.45	62.20
LS	87.19	84.56	89.23	67.73	71.36	63.94
mRMR	85.31	83.33	86.79	69.42	69.62	69.17
MMRL	**90.69**	**87.56**	**93.01**	**73.69**	**76.44**	**70.76**

73.69% for pMCI vs. sMCI) in comparison to the others, whose representations are learned prior to the final classification.

9.4.5 RESULTS OF VIEW-CENTRALIZED LEARNING FOR AD/MCI DIAGNOSIS

To demonstrate the variability of classification results using different templates, we first report the results of AD versus NC and pMCI versus sMCI classification based on a single template. Note that, in the single template case, the VCM FS method (Liu et al., 2015) only uses features from the selected template space, while features from other template spaces are completely ignored (ie, $a^{(1)} = 1$ and $a^{(2)} = 0$); thus Eq. (9.1) is similar to the formulation of an elastic net (Zou and Hastie, 2005). Two conventional FS methods are employed for comparison. The first one is based on the ranking of PC coefficients, and the second one is the COMPARE method proposed in Fan et al. (2007) that combines PC and SVM-RFE (Guyon et al., 2002). For fair comparison, the linear SVM with default parameter ($C = 1$) is adopted as a classifier after FS using PC, COMPARE, and the VCM method (Liu et al., 2015), respectively. Fig. 9.12 reports the distribution of classification results achieved by PC, COMPARE, and VCM using 10 single templates in AD versus NC classification.

As can be seen from Fig. 9.12, classification results based on different single templates are very different, regardless of the use of any FS methods. The underlying reason may be that the anatomical structure of a certain template may be more representative of the entire population, compared with other templates. In this case, the overall registration errors to this template are smaller and thus the feature representation generated from this template includes less noise. Another possible reason could be that the AD-related patterns generated from a certain template may

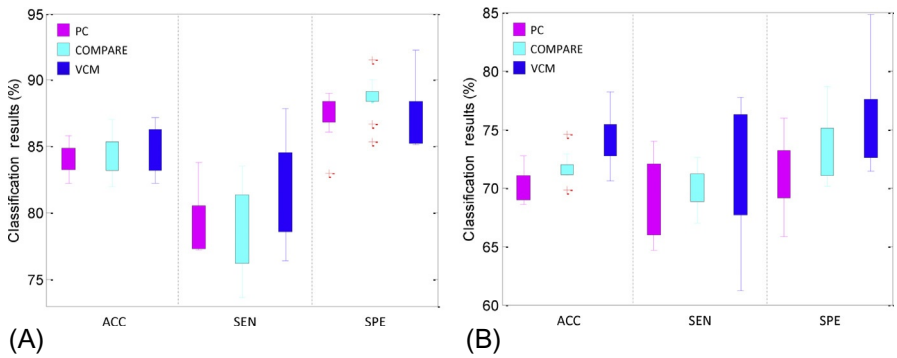

FIG. 9.12

Distribution of accuracy (ACC), sensitivity (SEN), and specificity (SPE) achieved by different single-template-based methods in (A) AD versus NC classification and (B) pMCI versus sMCI classification.

be more discriminative than those generated from other templates, thus having better generalization capability in identifying unseen test subjects.

The results for AD versus NC and pMCI versus sMCI classification by using multiple templates are shown in Table 9.6. The VCM method is compared with six FS methods, including (1) single-template-based PC (PC_SA); (2) single-template-based COMPARE (COMPARE_SA); (3) multiple-template-based PC (PC_MA); (4) multiple-template-based COMPARE (COMPARE_MA); (5) random subspace (RS) (Ho, 1998) that randomly selects features from the original feature space; and (6) Lasso (Tibshirani, 1996) that is a widely used FS method in neuroimaging analysis. Specifically, the averaged classification results of single-template-based methods (ie, PC and COMPARE) among all 10 templates are reported for PC_SA and COMPARE_SA. For both PC_MA and COMPARE_MA methods, we first concatenate all regional features (ie, 15,000-dimensional) extracted from multiple templates. Then, the top M ($M = \{1, 2, \ldots, 1500\}$) features are sequentially selected according to the PC (with respect to class labels) for PC_MA and PC+SVM-RFE for COMPAE_MA, and the best classification results are reported. It is worth noting that, in the proposed method, we learn 10 SVM classifiers based on different template-centralized feature subsets determined by the VCM FS method, and then construct a classifier ensemble with these learned base classifiers. For fair comparison, for the RS method, we randomly select M ($M = \{1, 2, \ldots, 1500\}$) features from each template for classification, and then record the best result. For the Lasso method, one first learns a Lasso model in a specific template space, and then selects features with nonzero coefficient in the learned weight vector. Given 10 templates, 10 classifiers can be constructed based on the features selected by RS and Lasso. Finally, for the RS and Lasso methods, these classifiers are combined using the same ensemble strategy as in the proposed method. The experimental results are summarized in Table 9.6.

From Table 9.6, it is clear to see that multitemplate-based methods (ie, PC_MA, COMPRE_MA, RS, Lasso, and VCM) generally achieve much better performance than single-template-based methods (ie, PC_SA and COMPARE_SA). Specifically, the best accuracies in AD versus NC classification achieved by PC_SA and COMPARE_SA are only 84.00% and 84.18%, respectively, which are much lower than those of COMPARE_MA, Lasso, and VCM. On the other hand, Table 9.6 shows that the VCM method consistently outperforms other methods in terms of classification accuracy, sensitivity, and AUC value. Obviously, by focusing on the representation from a certain template with other templates as extra guidance, the VCM method achieves better performance than the compared methods. In addition, from Table 9.6, one can observe that the sensitivities of PC_SA, COMPARE_SA, PC_MA, CPMPARE_MA, RS, and Lasso are much lower than their corresponding specificities. Here, low sensitivity values indicate low confidence in AD diagnosis, which will greatly limit practical usage in real-world applications. In contrast, VCM achieves a significantly improved sensitivity value in AD versus NC classification (ie, nearly 8% higher than the second best sensitivity achieved by Lasso).

Table 9.6 Results of AD Versus NC Classification Using Single Template and Multiple Templates

	AD vs. NC				pMCI vs. sMCI			
Method	**ACC (%)**	**SEN (%)**	**SPE (%)**	**AUC**	**ACC (%)**	**SEN (%)**	**SPE (%)**	**AUC**
PC_SA	84.00	79.53	87.45	0.7692	68.49	67.80	69.10	0.6285
COMPARE_SA	84.18	75.33	89.17	0.7870	70.06	68.08	72.02	0.6356
PC_MA	85.91	81.56	89.23	0.8191	72.78	74.62	70.91	0.7245
COMPARE_MA	87.19	80.56	92.31	0.8495	73.35	75.76	70.83	0.7405
RS	85.44	69.00	92.75	0.7688	69.05	68.10	72.94	0.6912
Lasso	87.27	84.78	89.23	0.9004	75.32	81.36	69.17	0.7602
VCM	**92.51**	**92.89**	88.33	**0.9583**	**78.88**	**85.45**	**76.06**	**0.8069**

9.4.6 RESULTS OF RELATIONSHIP-INDUCED LEARNING FOR AD/MCI DIAGNOSIS

To better make use of multiple sets of features generated from multiple templates, the following two strategies are used in this group of experiments, including (1) the feature concatenation method and (2) ensemble-based method. Specifically, in the feature concatenation method, features from multiple templates are *simply* concatenated into a long vector, and the corresponding classifier is constructed by using this feature vector. In the ensemble-based method, each feature set generated from a specific template space is treated individually, and multiple SVM classifiers based on these feature sets are constructed separately, followed by an ensemble strategy to combine the outputs of all SVMs for making a final classification decision.

In addition, the RIML method using the RIS FS algorithm (Liu et al., 2016) is compared with four methods, that is, (1) PC, (2) COMPARE method proposed in Fan et al. (2007) that combines PC and SVM-RFE, (3) statistical *t*-test method (Guyon et al., 2002), and (4) Lasso (Tibshirani, 1996) that is widely used for sparse FS in neuroimaging analysis. Here, we use PC<con>, COMPARE<con>, *t*-test<con>, and Lasso<con> to denote the four methods using four different FS algorithms (ie, PC, COMPARE, *t*-test, and Lasso) and the feature concatenation strategy (ie, <con>). Similarly, we use PC<ens>, COMPARE<ens>, *t*-test<ens>, and Lasso<ens> as another four methods using four different FS algorithms in each of the multiple template spaces during FS and then the proposed ensemble method (ie, <ens>) in the final classification step.

For comparison, the averaged classification results of single-template-based methods (including PC, COMPARE, *t*-test, and Lasso) are reported, with results given in Tables 9.7 and 9.8. Furthermore, the ROC curves achieved by five ensemble-based methods (RIML and four comparison methods) are plotted in Fig. 9.13.

Table 9.7 Performance of AD Versus NC Classification With Multiple Templates

	Method	ACC (%)	SEN (%)	SPE (%)	AUC
Single-template-based methods	PC	84.00	79.53	87.45	0.7692
	COMPARE	84.18	75.33	89.17	0.7870
	t-test	76.27	68.50	83.01	0.7496
	Lasso	84.32	81.66	86.36	0.8402
Multitemplate-based methods	PC<con>	84.01	81.56	89.23	0.8191
	COMPARE<con>	84.93	80.11	87.03	0.7907
	t-test<con>	81.87	70.77	**90.71**	0.8178
	Lasso<con>	86.62	84.78	89.80	0.8729
	PC<ens>	85.59	82.44	89.93	0.9151
	COMPARE<ens>	86.61	85.44	89.23	0.9085
	t-test<ens>	84.31	74.56	89.70	0.8878
	Lasso<ens>	87.27	84.78	89.23	0.9279
	RIML	**93.06**	**94.85**	90.49	**0.9579**

Table 9.8 Performance of pMCI Versus sMCI Classification With Multiple Templates

	Method	ACC (%)	SEN (%)	SPE (%)	AUC
Single-template-based methods	PC	68.49	67.80	69.10	0.6285
	COMPARE	70.06	68.08	72.02	0.6356
	t-test	61.99	64.93	73.11	0.6516
	Lasso	72.06	72.04	72.02	0.7203
Multitemplate-based methods	PC<con>	72.78	74.62	70.91	0.7245
	COMPARE<con>	73.35	75.76	70.83	0.7405
	t-test<con>	61.60	64.32	75.01	0.7163
	Lasso<con>	71.49	76.06	66.67	0.7136
	PC<ens>	73.92	73.38	72.32	0.7629
	COMPARE<ens>	75.56	75.75	73.48	0.7658
	t-test<ens>	63.36	60.60	71.74	0.6333
	Lasso<ens>	75.32	81.36	69.17	0.7602
	RIML	**79.25**	**87.92**	**75.54**	**0.8344**

From Table 9.7 and Fig. 9.13 (left), we can observe three main points. First, multitemplate-based methods generally achieve much better performance than single-template-based methods (ie, PC, COMPARE, *t*-test, and Lasso) in AD versus NC classification. For example, the best accuracy achieved by single-template-based methods is only 84.32% (achieved by Lasso), which is usually lower than those of multitemplate-based methods. This demonstrates that, compared with single-template-based methods, multitemplate-based methods can help promote the classification performance by taking advantage of richer representations for each subject. Second, using multiple templates, methods that adopt the proposed ensemble strategy (ie, PC<ens>, COMPARE<ens>, *t*-test<ens>, and Lasso<ens>) usually perform better than their counterparts that simply employ the feature concatenation strategy (ie, PC<con>, COMPARE<con>, *t*-test<con>, and Lasso<con>), in terms of four evaluation criteria. This implies that the feature concatenation strategy may not be a good choice to make use of multiple sets of features generated from multiple templates. Finally, RIML using the RIS FS algorithm achieves consistently better results than other methods in terms of classification accuracy, sensitivity, and AUC. To be specific, RIML achieves a classification accuracy of 93.06%, a sensitivity of 94.85%, and an AUC of 0.9579, while among all other methods the best accuracy is 87.27%, the best sensitivity is 85.44%, and the best AUC is 0.9279.

From Table 9.8 and Fig. 9.13 (right), one can observe again that multitemplate-based methods usually outperform single-template-based methods in pMCI versus sMCI classification. For example, the best accuracy of multitemplate-based methods (achieved by RIML) is 79.25%, which is much higher than the best accuracy of single-template-based methods, that is, 72.06% achieved by Lasso. In addition, among all nine multitemplate-based methods, RIML consistently achieves a better

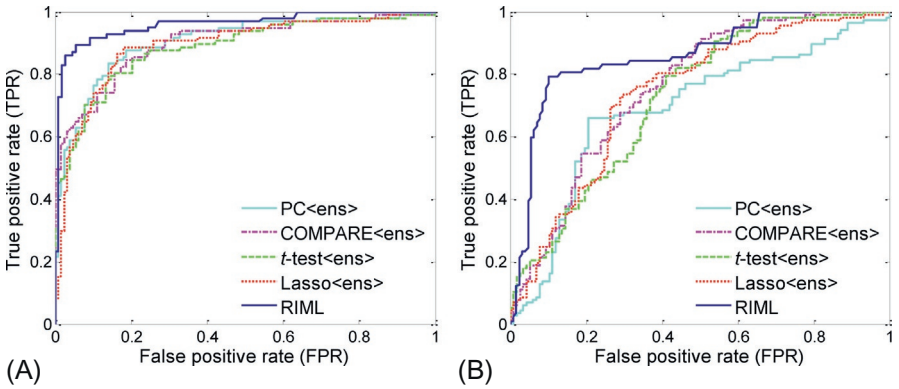

FIG. 9.13

ROC curves achieved by five ensemble-based methods using multiple templates in (A) AD versus NC classification and (B) pMCI versus sMCI classification.

performance than any other method, in terms of four evaluation criteria. In particular, RIML achieves an AUC of 0.8344, while the best AUC achieved by the comparison methods is only 0.7658 (achieved by COMPARE<ens>).

9.5 SUMMARY

In recent years, brain morphometric pattern analysis using MRI has been widely investigated for automatic diagnosis of AD and MCI. Existing MRI-based studies can be categorized into single-template-based and multitemplate-based approaches. It is widely accepted that different templates can convey complementary information, which is useful for AD and MCI diagnosis. In particular, by regarding each template as a specific view, some recent developments in multiview learning using multitemplate MRI data have been discussed in this chapter. This chapter first introduces a multiview feature representation method by using multiple templates (selected automatically from data). Then, four recent multiview learning approaches for AD/MCI diagnosis are presented. Specifically, a feature filter-based method provides a direct way to make use of multiview feature representations, with experimental results demonstrating significant improvements in AD and MCI diagnosis. For learning an optimal feature representation based on features generated from multitemplates, an MMRL method is proposed for improving the discriminative power of the original multiview features. To take advantage of the guidance information provided by different templates, a view-centralized learning method is developed to treat each template as a main view, while the other templates are used as side information source. Then, a relationship-induced multiview learning method is presented to model both the relationships among templates and those among subjects, to further

boost the performance of the AD/MCI classification model. The increased accuracy, sensitivity, and specificity achieved by these approaches indicate that multiview learning methods are a viable alternative to clinical diagnosis of brain alterations associated with cognitive impairment.

REFERENCES

Argyriou, A., Micchelli, C.A., Pontil, M., Ying, Y., 2008. A spectral regularization framework for multi-task structure learning. In: Advances in Neural Information Processing Systems 20. MIT Press, MA, USA, pp. 25–32.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. NeuroImage 11, 805–821.

Ashburner, J., Hutton, C., Frackowiak, R., Johnsrude, I., Price, C., Friston, K., 1998. Identifying global anatomical differences: deformation-based morphometry. Hum. Brain Map. 6, 348–357.

Alzheimer's Association, 2013. Alzheimer's disease facts and figures. Alzheimer's Dement. 9, 208–245.

Basha, T., Moses, Y., Kiryati, N., 2013. Multi-view scene flow estimation: a view centered variational approach. Int. J. Comput. Vision 101, 6–21.

Baxter, J., 1997. A Bayesian/information theoretic model of learning to learn via multiple task sampling. Mach. Learn. 28, 7–39.

Bengio, Y., 2009. Learning deep architectures for AI. In: Foundations and Trends® in Machine Learning, vol. 2. Now Publishers, Boston, USA, pp. 1–127.

Bozzali, M., Filippi, M., Magnani, G., Cercignani, M., Franceschi, M., Schiatti, E., et al., 2006. The contribution of voxel-based morphometry in staging patients with mild cognitive impairment. Neurology 67, 453–460.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. 2, 121–167.

Caruana, R., 1997. Multitask learning. Mach. Learn. 28, 41–75.

Chan, D., Janssen, J.C., Whitwell, J.L., Watt, H.C., Jenkins, R., Frost, C., et al., 2003. Change in rates of cerebral atrophy over time in early-onset Alzheimer's disease: longitudinal MRI study. Lancet 362, 1121–1122.

Chang, C.C., Lin, C.J., 2011. Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 27.

Chung, M., Worsley, K., Paus, T., Cherif, C., Collins, D., Giedd, J., et al., 2001. A unified statistical approach to deformation-based morphometry. NeuroImage 14, 595–606.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., et al., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage 56, 766–781.

Davatzikos, C., 1998. Mapping image data to stereotaxic spaces: applications to brain mapping. Hum. Brain Map. 6, 334–338.

Davatzikos, C., Genc, A., Xu, D., Resnick, S.M., 2001. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. NeuroImage 14, 1361–1369.

Davatzikos, C., Fan, Y., Wu, X, Shen, D., Resnick, S.M., 2008. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiol. Aging 29, 514–523.

Dickerson, B.C., Goncharova, I., Sullivan, M., Forchetti, C., Wilson, R., Bennett, D., et al., 2001. MRI-derived entorhinal and hippocampal atrophy in incipient and very mild Alzheimer's disease. Neurobiol. Aging 22, 747–754.

Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C., 2007. COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imaging 26, 93–105.

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

Fox, N., Warrington, E., Freeborough, P., Hartikainen, P., Kennedy, A., Stevens, J., et al., 1996. Presymptomatic hippocampal atrophy in Alzheimer's disease: a longitudinal MRI study. Brain 119, 2001–2007.

Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. Science 315, 972–976.

Frisoni, G., Testa, C., Zorzan, A., Sabattoli, F., Beltramello, A., Soininen, H., et al., 2002. Detection of grey matter loss in mild Alzheimer's disease with voxel based morphometry. J. Neurol. Neurosurg. Psychiat. 73, 657–664.

Gaser, C., Nenadic, I., Buchsbaum, B.R., Hazlett, E.A., Buchsbaum, M.S., 2001. Deformation-based morphometry and its relation to conventional volumetry of brain lateral ventricles in MRI. NeuroImage 13, 1140–1145.

Goldszal, A.F., Davatzikos, C., Pham, D.L., Yan, M.X., Bryan, R.N., Resnick, S.M., 1998. An image-processing system for qualitative and quantitative volumetric analysis of brain images. J. Comput. Assist. Tomograp. 22, 827–837.

Gong, Y., Ke, Q., Isard, M., Lazebnik, S., 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. Int. J. Comput. Vision 106, 210–233.

Grau, V., Mewes, A., Alcaniz, M., Kikinis, R., Warfield, S.K., 2004. Improved watershed transform for medical image segmentation using prior information. IEEE Trans. Med. Imaging 23, 447–458.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422.

He, X., Cai, D., Niyogi, P., 2005. Laplacian score for feature selection. In: Advances in Neural Information Processing Systems. MIT Press, MA, USA, pp. 507–514.

Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M.K., Johnson, S.C., 2009. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. NeuroImage 48, 138–149.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844.

Hua, X., Leow, A.D., Lee, S, Klunder, A.D., Toga, A.W., Lepore, N., et al., 2008. 3D characterization of brain atrophy in Alzheimer's disease and mild cognitive impairment using tensor-based morphometry. NeuroImage 41, 19–34.

Hua, X., Hibar, D.P., Ching, C.R., Boyle, C.P., Rajagopalan, P., Gutman, B.A., et al., 2013. Unbiased tensor-based morphometry: improved robustness and sample size estimates for Alzheimer's disease clinical trials. NeuroImage 66, 648–661.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Mag. Reson. Imaging 27, 685–691.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.

Jolliffe, I., 2002. Principal Component Analysis. Wiley Online Library.

Joseph, J., Warton, C., Jacobson, S.W., Jacobson, J.L., Molteno, C.D., Eicher, A., et al., 2014. Three-dimensional surface deformation-based shape analysis of hippocampus and caudate nucleus in children with fetal alcohol spectrum disorders. Hum. Brain Map. 35, 659–672.

Kipps, C., Duggins, A., Mahant, N., Gomes, L., Ashburner, J., McCusker, E., 2005. Progression of structural neuropathology in preclinical Huntington's disease: a tensor based morphometry study. J. Neurol. Neurosurg. Psychiat. 76, 650–655.

Koikkalainen, J., Lötjönen, J., Thurfjell, L., Rueckert, D., Waldemar, G., Soininen, H., 2011. Multi-template tensor-based morphometry: application to analysis of Alzheimer's disease. NeuroImage 56, 1134–1144.

Kruskal, J.B., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29, 1–27.

Leow, A.D., Klunder, A.D., Jack, C.R., Toga, A.W., Dale, A.M., Bernstein, M.A., et al., 2006. Longitudinal stability of MRI for mapping brain change using tensor-based morphometry. NeuroImage 31, 627–640.

Leporé, N., Brun, C., Chou, Y.Y., Lee, A., Barysheva, M., Zubicaray, G.I.D., et al., 2008. Multi-atlas tensor-based morphometry and its application to a genetic study of 92 twins. In: Medical Image Computing and Computer-Assisted Intervention Workshop on Mathematical Foundations of Computational Anatomy, New York, USA, pp. 48–55.

Li, S.Z., Zhu, L., Zhang, Z, Blake, A., Zhang, H., Shum, H., 2002. Statistical learning of multi-view face detection. In: European Conference on Computer Vision, pp. 67–81.

Liu, M., Zhang, D., Shen, D., 2012. Ensemble sparse classification of Alzheimer's disease. NeuroImage 60, 1106–1116.

Liu, M., Zhang, D., Shen, D., 2014. Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. Hum. Brain Map. 35, 1305–1319.

Liu, M., Zhang, D., Shen, D., 2015. View-centralized multi-atlas classification for Alzheimer's disease diagnosis. Hum. Brain Map. 36, 1847–1865.

Liu, M., Zhang, D., Shen, D., 2016. Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment. IEEE Trans. Med. Imaging, http://dx.doi.org/10.1109/TMI.2016.2515021.

Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., et al., 2009. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 51, 73–83.

Min, R., Wu, G., Cheng, J., Wang, Q., Shen, D., 2014a. Multi-atlas based representations for Alzheimer's disease diagnosis. Hum. Brain Map. 35, 5052–5070.

Min, R., Wu, G., Shen, D., 2014b, Maximum-margin based representation learning from multiple atlases for Alzheimer's disease classication. In: Presented at the Medical Image Computing and Computer-Assisted Intervention, New York, USA, pp. 635–642.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., et al., 2005. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). Alzheimer's Dement. 1, 55–66.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1226–1238.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45, S199–S209.

Shen, D., Davatzikos, C., 2002. HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imaging 21, 1421–1439.

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

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17, 87–97.

Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: a survey. IEEE Trans. Med. Imaging 32, 1153–1190.

Tang, S., Fan, Y., Wu, G., Kim, M., Shen, D., 2009. RABBIT: rapid alignment of brains by building intermediate templates. NeuroImage 47, 1277–1287.

Teipel, S.J., Born, C., Ewers, M., Bokde, A.L., Reiser, M.F., Möller, H.J., et al., 2007. Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. NeuroImage 38, 13–24.

Thomas, A., Ferrar, V., Leibe, B., Tuytelaars, T., Schiel, B., Van Gool, L., 2006. Towards multi-view object class detection. In: IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, pp. 1589–1596.

Thompson, P.M., Mega, M.S., Woods, R.P., Zoumalan, C.I., Lindshield, C.J., Blanton, R.E., et al., 2001. Cortical change in Alzheimer's disease detected with a disease-specific population-based brain atlas. Cerebral Cortex 11, 1–16.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B 267-288.

Vincent, L., Soille, P., 1991. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. 13, 583–598.

Wang, Y., Nie, J., Yap, P.T., Shi, F., Guo, L., Shen, D., 2011. Robust deformable-surface-based skull-stripping for large-scale studies. In: Medical Image Computing and Computer-Assisted Intervention, Toronto, Canada, pp. 635–642.

Wang, Y., Nie, J., Yap, P.T., Li, G., Shi, F., Geng, X., et al., 2014. Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. PLoS ONE 9 (1), e77810.

Whitford, T.J., Grieve, S.M., Farrow, T.F., Gomes, L., Brennan, J., Harris, A.W., et al., 2006. Progressive grey matter atrophy over the first 2-3 years of illness in first-episode schizophrenia: a tensor-based morphometry study. NeuroImage 32, 511–519.

Xu, C., Tao, D., Xu, C., 2014. Large-margin multi-view information bottleneck. IEEE Trans. Pattern Anal. Mach. Intell. 36, 1159–1572.

Yap, P.T., Wu, G., Zhu, H., Lin, W., Shen, D., 2009. Timer: tensor image morphing for elastic registration. NeuroImage 47, 549–563.

Zhang, D., Shen, D., 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage 59, 895–907.

Zhang, Y., Brady, M., Smith, S, 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20, 45–57.

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

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. B 67, 301–320.