

# Multi-modal AD classification via self-paced latent correlation analysis

Qi Zhu<sup>a,b,\*</sup>, Ning Yuan<sup>a</sup>, Jiashuang Huang<sup>a</sup>, Xiaoke Hao<sup>a</sup>, Daoqiang Zhang<sup>a</sup>

<sup>a</sup> College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

<sup>b</sup> Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China



## ARTICLE INFO

### Article history:

Received 6 December 2017

Revised 7 September 2018

Accepted 28 April 2019

Available online 10 May 2019

Communicated by Shaoting Zhang

### Keywords:

Multi-modal fusion

Feature extraction

Low-rank

Self-paced learning

Computer-aided diagnosis

## ABSTRACT

As an irreparable brain disease, Alzheimer's disease (AD) seriously impairs human thinking and memory. The accurate diagnosis of AD plays an important role in the treatment of patients. Many machine learning methods have been widely used in classification of AD and its early stage. An increasing number of studies have found that multi-modal data provide complementary information for AD prediction problem. In this paper, we propose multi-modal rank minimization with self-paced learning for revealing the latent correlation across different modalities. In the proposed method, we impose low-rank constraint on the regression coefficient matrix, which is composed of regression coefficient vectors of all modalities. Meanwhile, we adaptively evaluate the contribution of each sample to the fusion model by self-paced learning (SPL). Finally, we utilize multiple-kernel learning (MKL) to classify the multi-modal data. Experiments on the Alzheimer's disease Neuroimaging Initiative (ADNI) databases show that the proposed method obtains better classification performance than the state-of-the-art methods.

© 2019 Elsevier B.V. All rights reserved.

## 1. Introduction

Alzheimer's disease (AD) [1–9] is a progressive nervous system degenerative disease. The typical clinical characteristics of AD include dysmnnesia, aphasia, visuospatial impairment, executive dysfunction, etc. As life expectancy increases, AD is becoming a serious health problem in the elder. There are many hypotheses about the pathogeny of AD, like familial inheritance, physical disease, head trauma, etc. It is still an open problem of exploring the biomarkers [4] and making accurate diagnosis of AD.

In recent years, many machine learning methods have been proposed for the diagnosis of AD and its prodromal stage, i.e., mild cognitive impairment (MCI). For example, by integrating the classifiers on different local patch subsets, Liu et al. [2] obtained more accurate classification performance on AD data. Sarraf and Tofghi adopted the convolutional neural network (CNN) to classify AD patient from normal control (NC) [10]. In the early research work, people tended to focus on classification methods based on single-modal data [2–5], such as structural Magnetic Resonance Imaging (MRI) [5] and functional imaging (e.g., Positron Emission Tomography, PET) [3], and ignored the complementary information from other modalities. To alleviate this deficiency, Zhang and Shen [6] developed multi-modal fusion method combining different modalities, including MRI, PET, and cerebrospinal fluid (CSF), for

AD diagnosis. Thung et al. proposed a deep learning model, which incorporates incomplete multi-modal AD data to improve the classification performance. Benefiting from utilizing the complementary information among different modalities, multi-modal based methods can achieve higher classification accuracy than single-modal based methods [11].

These classification methods used in AD classification are mainly based on the analysis of high dimensional feature, which may lead to the curse of dimensionality. Feature selection is an effective technique for dimensionality reduction by removing the irrelevant features. At present, a number of feature selection approaches have been applied into AD classification methods include multi-task feature selection (MTFS) [7], group lasso [12] and principal component analysis (PCA) [13]. In the multi-modal methods, these selected features of each modality are often directly combined to predict the class label [14]. However, most recent efforts made on feature selection methods ignore two important aspects: (1) the intrinsic correlation among different modalities, (2) the difference of sample significances.

For addressing these two problems, we proposed a novel multi-modal classification model, which is optimized in feature and sample levels simultaneously. For one AD patient or NC, it is reasonable to assume that the data of different modalities have some intrinsic correlation. The conventional multi-modal AD fusion method often integrates the different modalities linearly [6,7,11]. This approach may miss some latent important characteristics of different modalities. As we know, rank is the powerful global measure of matrix sparseness. Thus, in our proposed method,

\* Corresponding author at: College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China.

E-mail address: [zhuqi@nuaa.edu.cn](mailto:zhuqi@nuaa.edu.cn) (Q. Zhu).

we exploit low-rank technique [15–20] to effectively capture the latent correlation of multi-modal AD data. In the aspect of estimating the sample significance, we adopt the self-paced learning (SPL) [21–25] to our method. As a cognitive driven model, SPL can dynamically evaluate the learning difficulty of each sample, and it gradually increases the training set by introducing more hard-learning samples. In AD classification problem, some patients may have many neurological diseases, it makes the diagnosis of AD or MCI difficult. Therefore, we apply the SPL into processing high-confidence samples, noise samples and outliers respectively in AD data. In the multi-modal fusion process, sample significance analysis is also helpful to describe and capture the relevance across different modalities. These important samples will make a large contribution in building the classifier decision boundary than those insignificance and noise samples. In this way, the influence of noise samples and outliers to the classification model can be suppressed.

In the AD classification problem, we need to handle several tasks include AD vs. MCI, AD vs. NC etc. In our work, we use MTFs to remove those irrelevant features to these tasks. After MTFs, the usual practice is to directly connect those different feature spaces into one single matrix and then train a classifier. In our work, we use the proposed novel MKL method to classify the multi-modal data, which is effective to reveal the latent correlation among different modalities and can offer a general framework for data fusion.

Overall, the proposed method mainly has the following contributions:

- (1) The low-rank constraint is first employed to capture the intrinsic correlation across different modalities.
- (2) The self-paced learning is first adopted to estimate the sample significance of AD data. The SPL has been proven to be robust to noise samples and can speed up the objective function convergence.
- (3) The experimental results show that the proposed method achieves promising performance in AD classification.

The remaining parts are organized as follows. In Section 2, some related works are presented. The proposed method and its reasonability are given in Section 3. Section 4 demonstrates the experiment datasets and settings. The results are shown in Section 5. Finally, we draw the conclusion and future works in Section 6.

## 2. Related works

### 2.1. Low-rank representation

Low-rank technique can provide the intrinsic information of data, and it has been successfully used in face recognition, computer vision and other popular domains [15–20]. By using low-rank method, Wang et al. [16] solved the problem of robust face recognition, in which both training and test image data are corrupted. Haeffele et al. [15] developed a novel matrix factorization technique based on low-rank constraint, which is suitable for large datasets. Liu et al. [17] proposed low-rank representation (LRR) model which can capture the global structure of training samples. He et al. [26] proposed a unified framework that pursues the low-rank subspace for hyperspectral image restoration. For the problem of unsupervised domain transfer learning, in which no labels are available in the target domain, Xu et al. [27] proposed the model can preserve the structures information of source and target data. In this way, their method can avoid potentially negative transfer and is more robust to different types of noise. These research works have proven that low-rank method is effective for matrix completion and recovering the global structure of the data. Inspired by these methods, we apply low-rank method into the AD data to capture the correlation among different modalities.

### 2.2. Self-paced learning

Motivated by human learning, Bengio et al. proposed curriculum learning [28] whose main idea is to organize samples in an easy-to-hard learning order. Self-paced learning is the extension of curriculum learning, and it is more flexible and adaptive in estimating the learning difficulty of sample [21–25]. Jiang et al. [21] proposed a novel re-ranking approach for multi-modal data based on self-paced learning. Kumar et al. [23] introduced the self-paced learning to the latent variable models. Self-paced learning iteratively picks the easy-learning samples and updates the model parameters till all samples participate in training. The self-paced learning model can alleviate the problem of falling into a bad local optimum by presenting the data in an easy-to-hard learning order. The self-paced learning model can be simply represented as:  $\min_{\mathbf{P}} \mathbf{P} \text{ loss} + f(\mathbf{P}; k)$ , where  $\mathbf{P}$  denotes the self-paced weighting matrix,  $k$  is the learning pace parameter, and  $\text{loss}$  denotes the sample loss by the classifier. There are many types of self-paced functions, e.g., binary weighting scheme, linear weighting scheme, logarithmic weighting scheme, etc. No matter what kind of self-paced function  $f(\mathbf{P}; k)$ , it needs to satisfy the three conditions: (1)  $f(\mathbf{P}; k)$  is convex with respect to  $\mathbf{P} \in [0, 1]$ ; (2) the weight of each sample should be monotonically decreasing with respect to its relevant loss; and (3) the weight of each sample should be monotonically decreasing with respect to the self-paced parameter  $k$ . Inspired by self-paced learning, we dynamically estimate the contribution of each sample to the fusion model to avoid the influence of noise samples and outliers in early learning stage.

## 3. Proposed method

### 3.1. Objective function

In this paper, we propose self-paced sample weighting based multi-modal rank minimization (SPMRM) to construct a more discriminative and robust model for AD classification. In our work, we use multi-task feature selection (MTFS) [6,29,30] as data pre-processing method, which is used to coarsely select those relevant features of each modality in AD data. We denote the coarsely selected data  $m$  modalities as  $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(m)}\}$ , and each modality has  $n$  label samples  $\{\mathbf{x}_i^{(j)}, y_i\}_{i=1}^n$ , ( $j = 1, \dots, m$ ). After MTFs, the dimensionalities of these modalities may be different. For addressing this problem, we proposed a novel multi-kernel learning (MKL) framework to solve the heterogeneous data fusion problem. In addition, we can implicitly use feature space induced by kernel function to improve the classification performance. MKL framework can combine the complementary information among all modalities. More details can be found in Section 3.3. We use mapping function  $\phi$  induced by kernel function to project each modal data into high dimensional space. According to the practical application, many kernel functions can be chosen, such as linear kernel, polynomial kernel Gaussian kernel, etc.

Then, we consider a linear regression model whose parameters are determined by minimizing a regularized sum-of-squares error function [31] given by:

$$\min_{\omega} \frac{1}{2} \sum_{i=1}^n (y_i - \omega^T \phi(\mathbf{x}_i))^2 + \frac{\lambda}{2} \omega^T \omega \quad (1)$$

where  $n$  is the number of samples,  $y_i$  is the label of  $i$ -th sample, and  $\omega$  is a vector of parameters. The solution of Eq. (1) is

$$\omega = \frac{1}{\lambda} \sum_{i=1}^n (y_i - \omega^T \phi(\mathbf{x}_i)) \phi(\mathbf{x}_i) = \Phi^T \alpha \quad (2)$$

where  $\Phi$  is the matrix whose  $i$ th row is given by  $\phi(\mathbf{x}_i)^T$ . And  $\alpha = (\alpha_1, \dots, \alpha_n)^T$ , where  $\alpha_i = \frac{1}{\lambda} (y_i - \omega^T \phi(\mathbf{x}_i))$ . Then we can

rewrite the linear regression model using the dual representation [31] and reformulate the Eq. (1) in terms of the parameter  $\alpha$  instead of working with the parameter vector  $\omega$ . If we substitute  $\omega = \Phi^T \alpha$  into Eq. (1), we obtain:

$$\min_{\alpha} \frac{1}{2} \alpha^T \Phi \Phi^T \Phi \Phi^T \alpha - \alpha^T \Phi \Phi^T \mathbf{y} + \frac{1}{2} \mathbf{y}^T \mathbf{y} + \frac{\lambda}{2} \alpha^T \Phi \Phi^T \alpha \quad (3)$$

where  $\mathbf{y} = (y_1, \dots, y_n)^T$ . Then we define the Gram matrix  $\mathbf{K} = \Phi \Phi^T$ , which is a  $n \times n$  symmetric matrix with elements  $\mathbf{K}_{i,j} = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)$  where  $i, j \in [1, n]$ . Then the Eq. (3) can be written as:

$$\min_{\alpha} \frac{1}{2} \alpha^T \mathbf{K} \mathbf{K} \alpha - \alpha^T \mathbf{K} \mathbf{y} + \frac{1}{2} \mathbf{y}^T \mathbf{y} + \frac{\lambda}{2} \alpha^T \mathbf{K} \alpha \quad (4)$$

Eq. (4) can be written as:

$$\min_{\alpha} \frac{1}{2} (\mathbf{y} - \mathbf{K} \alpha)^T (\mathbf{y} - \mathbf{K} \alpha) + \frac{\lambda}{2} \alpha^T \mathbf{K} \alpha \quad (5)$$

Considering all the modalities in AD data, we use a matrix  $\mathbf{A} = [\alpha^{(1)}, \dots, \alpha^{(m)}]$  to denote the parameter vectors of all modalities in AD data, where  $\alpha^{(j)}$  is the parameter vector of  $j$ th modality. It can be assumed that the matrix  $\mathbf{A}$  with low-rank constraint reserves the latent correlation information among different modalities in multi-modal data.

Therefore, we give the low-rank constraint based multi-modal fusion model:

$$\min_{\mathbf{A}} \sum_{j=1}^m (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)})^T (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)}) + \lambda_S \sum_{j=1}^m \alpha^{(j)T} \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)} + \lambda_R R(\mathbf{A}) \quad (6)$$

where  $\mathbf{K}(\mathbf{X}^{(j)})$  is the Gram matrix of  $j$ -th modality data. In our work, we adopt Gaussian kernel function, also known as radial basis function (RBF), which is defined as  $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$ .  $\lambda_S$  and  $\lambda_R$  are the regularization parameters.  $R(\mathbf{A})$  denotes the rank of parameter matrix  $\mathbf{A}$ .

In our work, we also measure the contribution of each sample to the model. Specifically, we will select the most discriminative samples, and employ them to construct a more robust multi-modal classification model. Therefore, we further improve the original model as follows:

$$\min_{\mathbf{A}, \mathbf{P}} \sum_{j=1}^m (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)})^T \mathbf{P} (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)}) + \lambda_S \sum_{j=1}^m \alpha^{(j)T} \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)} + \lambda_R R(\mathbf{A}) + mf(\mathbf{P}; k) \quad (7)$$

where  $\mathbf{P}$  is a diagonal matrix whose elements are the weight of samples. In our work, we employ self-paced learning to estimate the weight matrix. Specifically, we utilize a self-paced function  $f(\mathbf{P}; k)$  that allocates weight to samples to controls the number of samples considered in current learning stage. The parameter  $k$  determines the learning pace. Self-paced function can be defined in various forms, which is discussed in Section 3.2.

### 3.2. Optimization for SPMRM

Considering solving Eq. (7) is an NP-hard problem, we use the nuclear norm  $\|\mathbf{A}\|_*$ , which is the sum of the singular values of matrix  $\mathbf{A}$ , to replace  $R(\mathbf{A})$ . Then we adopt the alternating direction method of multipliers (ADMM) framework [32] to resolve the problem. After an auxiliary variable  $\mathbf{J}$  is introduced, the objective function can be reformulated as:

$$\min_{\mathbf{A}, \mathbf{P}} \sum_{j=1}^m (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)})^T \mathbf{P} (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)}) + \lambda_S \sum_{j=1}^m \alpha^{(j)T} \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)} + \lambda_R \|\mathbf{J}\|_* + mf(\mathbf{P}; k) \quad (8)$$

s.t.  $\mathbf{J} = \mathbf{A}$

The augmented Lagrangian function can be written as:

$$\begin{aligned} L(\mathbf{A}, \mathbf{J}, \lambda, \mathbf{P}) = & \min_{\mathbf{A}, \mathbf{J}, \lambda, \mathbf{P}} \sum_{j=1}^m (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)})^T \mathbf{P} (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)}) \\ & + \lambda_S \sum_{j=1}^m \alpha^{(j)T} \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)} + \lambda_R \|\mathbf{J}\|_* \\ & + \langle \lambda, \mathbf{A} - \mathbf{J} \rangle + \frac{\rho}{2} \|\mathbf{A} - \mathbf{J}\|_F^2 + mf(\mathbf{P}; k) \end{aligned} \quad (9)$$

where  $\lambda \in \mathbb{R}^{d \times m}$  is Lagrange multiplier, and  $\langle \mathbf{X}_1, \mathbf{X}_2 \rangle$  denotes the trace of  $\mathbf{X}_1^T \mathbf{X}_2$ .

The parameters of above problem are updated by:

$$\begin{cases} \mathbf{A} \leftarrow \underset{\mathbf{A}}{\operatorname{argmin}} L(\mathbf{A}, \mathbf{J}, \lambda, \mathbf{P}) \\ \mathbf{J} \leftarrow \underset{\mathbf{J}}{\operatorname{argmin}} L(\mathbf{A}, \mathbf{J}, \lambda, \mathbf{P}) \\ \lambda \leftarrow \lambda + \rho(\mathbf{A} - \mathbf{J}) \\ \mathbf{P} \leftarrow \underset{\mathbf{P}}{\operatorname{argmin}} L(\mathbf{A}, \mathbf{J}, \lambda, \mathbf{P}) \end{cases} \quad (10)$$

Each sub-problem of the original problem is convex, and we can obtain its optimal solution. The solutions of  $\mathbf{A}, \mathbf{J}$  and  $\mathbf{P}$  are derived as follows:

(1) The solution of  $\mathbf{A}$ :

$$\begin{aligned} \min_{\mathbf{A}} L(\mathbf{A}, \mathbf{J}, \lambda, \mathbf{P}) = & \min_{\mathbf{A}} \sum_{j=1}^m (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)})^T \mathbf{P} (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)}) \\ & + \lambda_S \sum_{j=1}^m \alpha^{(j)T} \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)} + \langle \lambda, \mathbf{A} - \mathbf{J} \rangle \\ & + \frac{\rho}{2} \|\mathbf{A} - \mathbf{J}\|_F^2 \end{aligned} \quad (11)$$

$$\begin{aligned} = & \min_{\mathbf{A}} \sum_{j=1}^m (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)})^T \mathbf{P} (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)}) \\ & + \lambda_S \sum_{j=1}^m \alpha^{(j)T} \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)} + \lambda_v^T (\alpha^{(j)} - \mathbf{J}^v) + \frac{\rho}{2} \|\alpha^{(j)} - \mathbf{J}^v\|_2^2 \end{aligned} \quad (12)$$

where  $\mathbf{J}^v$  is the column vector of  $\mathbf{J}$ . The above problem can be divided into  $m$  sub-problems regarding  $\alpha^{(j)}$ ,  $j = 1, 2, \dots, m$ .

$$\begin{aligned} \min_{\alpha^{(j)}} & (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)})^T \mathbf{P} (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)}) \\ & + \lambda_S \alpha^{(j)T} \mathbf{K}(\mathbf{X}^{(j)}) \alpha^{(j)} + \lambda_v^T (\alpha^{(j)} - \mathbf{J}^v) + \frac{\rho}{2} \|\alpha^{(j)} - \mathbf{J}^v\|_2^2 \end{aligned} \quad (13)$$

Eq. (13) is a convex problem with respect to  $\alpha^{(j)}$ , which can be easily solved by gradient descent method. Then, we obtain the matrix  $\mathbf{A}$  by normalizing each row.

(2) The solution of  $\mathbf{J}$

According to the augmented Lagrange multipliers (ALM) [32], we solve  $\mathbf{J}$  as:

$$\mathbf{J} = \underset{\mathbf{J}}{\operatorname{argmin}} \lambda_R \|\mathbf{J}\|_* + \langle \lambda, \mathbf{A} - \mathbf{J} \rangle + \frac{\rho}{2} \|\mathbf{A} - \mathbf{J}\|_F^2 = \mathbf{US}_\theta[\mathbf{S}]\mathbf{V}^T \quad (14)$$

where  $\theta = \lambda_R/\rho$ , and  $\mathbf{USV}^T$  is the results of singular value decomposition (SVD) of  $\mathbf{A} - \lambda/\rho$ .

(3) The solution of  $\mathbf{P}$ :

$$\min_{\mathbf{P}} \sum_{j=1}^m (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)})\boldsymbol{\alpha}^{(j)})^T \mathbf{P} (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)})\boldsymbol{\alpha}^{(j)}) + mf(\mathbf{P}; k) \quad (15)$$

Before introducing the self-paced function, we first discuss a most common weighting scheme, the binary weighting (also called hard weighting). In binary weighting,  $f(\mathbf{P}; k)$  is defined based on the  $L_1$ - norm of  $\mathbf{P} \in [0, 1]^n$ :

$$f(\mathbf{P}; k) = -\frac{1}{k} \|\mathbf{P}\|_1 = -\frac{1}{k} \sum_{i=1}^n p_i \quad (16)$$

where  $\mathbf{P}$  is a diagonal matrix, whose diagonal elements are  $\{p_1, \dots, p_n\}$ , and the other elements of  $\mathbf{P}$  are all zeros. Substituting Eq. (16) into Eq. (15), the solution of  $p_i$  can be calculated as:

$$p_i = \begin{cases} 1 & \frac{1}{m} \sum_{j=1}^m \ell_{ij} < \frac{1}{k} \\ 0 & \frac{1}{m} \sum_{j=1}^m \ell_{ij} \geq \frac{1}{k} \end{cases} \quad (17)$$

where  $\ell_{ij}$  is the squared loss of  $i$ th sample in  $j$ th modality.

The sample weight will be set 0 if the squared loss is higher than the threshold value  $1/k$ ; otherwise it will be set 1. It should be noted that this weighting scheme is sensitive to sample loss. For example, when the threshold is 0.35 and the squared loss of sample  $m_1$  and  $m_2$  are 0.34 and 0.36, respectively, the weight of  $m_1$  is 1 whereas the weight of  $m_2$  is 0. Obviously, the binary weighting scheme is unreasonable in above case. To alleviate this deficiency, in our work, we adopt mixture weighting scheme to assign weights of samples. Mixture weighting scheme is the combination of soft weighting scheme and binary weighting scheme. The sample weight assigned by soft weighting can reflect the importance of sample more faithfully, e.g. linear weighting scheme, logarithmic weighting scheme etc. Specifically, in mixed weighting scheme, if the loss is either too small or too large, the binary weighting is applied. Otherwise, the soft weighting is applied in the middle area. The mixed weighted scheme is defined as:

$$f(p_i; k; k') = -\xi \sum_{i=1}^n \log(p_i + \xi k) \quad (18)$$

where  $\xi = 1/(k' - k)$ ,  $k'$  is an auxiliary parameter ( $k' > k > 0$ ). Then the derivation of  $p_i$  in Eq. (15) is:

$$\frac{\partial L}{\partial p_i} = \sum_{j=1}^m \ell_{ij} - \frac{m\xi}{p_i + k\xi} \quad (19)$$

Let Eq. (20) be 0, the closed-form optimal solution of  $p_i$  can be calculated by:

$$p_i = \begin{cases} 1 & \frac{1}{m} \sum_{j=1}^m \ell_{ij} \leq \frac{1}{k'} \\ 0 & \frac{1}{m} \sum_{j=1}^m \ell_{ij} \geq \frac{1}{k} \\ \frac{m\xi}{\sum_{j=1}^m \ell_{ij}} - k\xi & \text{otherwise} \end{cases} \quad (20)$$

Fig. 1 shows the comparison of binary weighting and mixture weighting. Comparing with soft weighting scheme, the mixture weighting can tolerate some errors in loss function. In our work,

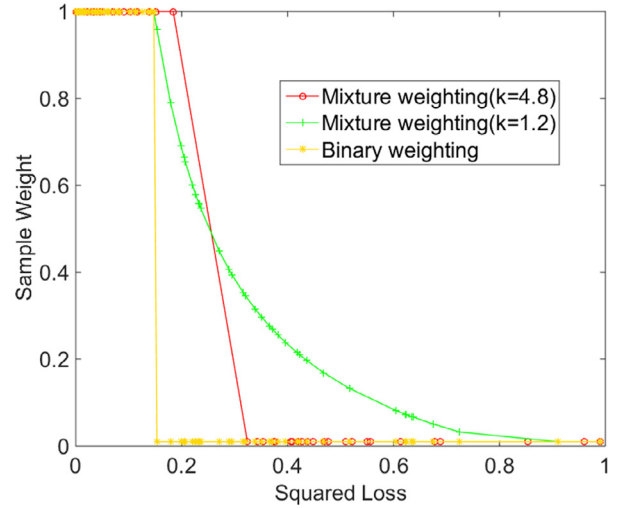


Fig. 1. Comparison of different weighting schemes ( $k = 4.8$ ,  $k = 1.2$ ).

#### Algorithm 1 SPMRM.

**Input:** Training data  $\{\mathbf{x}_1^j, \mathbf{x}_2^j, \dots, \mathbf{x}_m^j\}_{j=1}^m$ , self-paced parameters  $k, k'$ , parameters  $\lambda_S, \lambda_R$ .  
**Output:** parameter  $\mathbf{J}$ , self-paced weighted matrix  $\mathbf{P}$ .  
1: Initialize  $\lambda_0$ , self-paced  $k, k'$  and  $t = 0$ .  
2: **while** ( $t \neq$  the threshold) **do**  
3: **for**  $j \leftarrow \{1, \dots, m\}$  **do**  
4:  $(\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(m)})_{t+1} \leftarrow$  solution by Eq. (13).  
5:  $(\mathbf{P})_{t+1} \leftarrow$  solution by Eq. (21).  
6: **end for**  
7:  $\mathbf{J}_{t+1} \leftarrow$  solution by Eq. (14).  
8: Update  $\lambda_t$  to  $\lambda_{t+1}$  in Eq. (10).  
9: **end while**  
10:  $\mathbf{J} = \mathbf{J}_{t+1}$ ,  $\mathbf{P} = \mathbf{P}_{t+1}$

we first set  $k = 4.8$ , and gradually decrease the value of  $k$  to 1.2. As we can see from Fig. 1, when  $k = 4.8$ , only a small part of samples participate in model. When  $k = 1.2$ , almost all the samples are considered. Then the Eq. (15) can be denoted as:

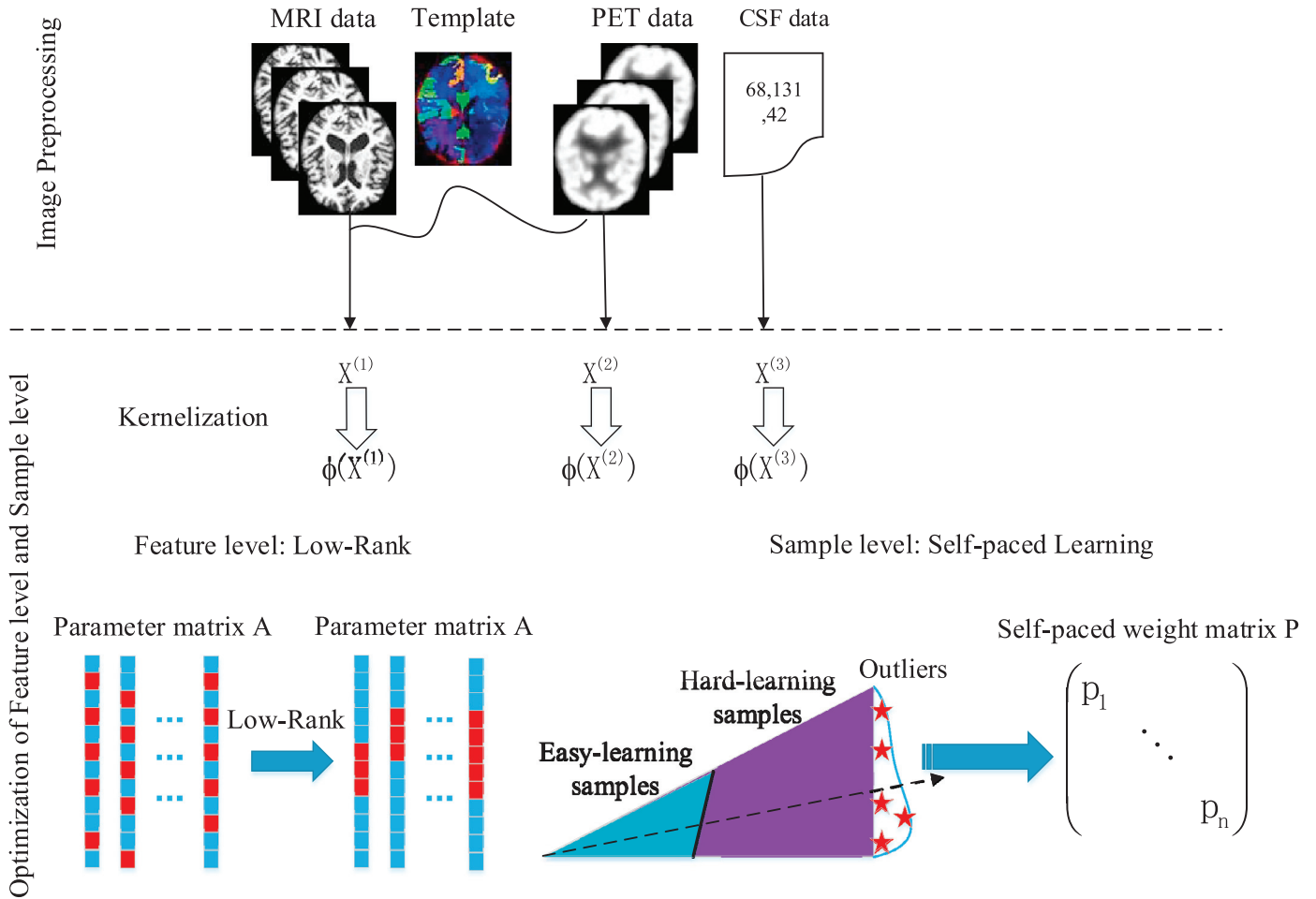
$$\min_{\mathbf{P}} \sum_{j=1}^m (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)})\boldsymbol{\alpha}^{(j)})^T \mathbf{P} (\mathbf{y} - \mathbf{K}(\mathbf{X}^{(j)})\boldsymbol{\alpha}^{(j)}) - m\xi \log(\mathbf{P} + \xi k) \quad (21)$$

### 3.3. Multi-modal data fusion for classification

According to Algorithm 1, we can obtain the parameter  $\mathbf{J}$  and the weight matrix  $\mathbf{P}$ . We also need to calculate the regression coefficient of the fusion data. It is easy to extend the single kernel linear regression to multiple-kernel linear regression. We denote  $(\mathbf{x}_a, \mathbf{x}_b) = \sum_j \beta_j k^{(j)}(\mathbf{x}_a^{(j)}, \mathbf{x}_b^{(j)})$  as a mixed kernel between the multi-modal training samples  $\mathbf{x}_a$  and  $\mathbf{x}_b$ . Meanwhile, let  $k(\mathbf{x}_a, \mathbf{x}_c) = \sum_j \beta_j k^{(j)}(\mathbf{x}_a^{(j)}, \mathbf{x}_c^{(j)})$  as a mixed kernel between the multimodal training sample  $\mathbf{x}_a$  and the test sample  $\mathbf{x}_c$ . Our method integrates multiple kernels into one kernel, which can be regarded as an approach to kernel combination. In our work, we adopt grid search with the constraint  $\sum_j \beta_j = 1$  to find the optimal combination parameters. Substituting the mixed kernel to Eq. (13), we can get the regression coefficient of the fusion data.

### 3.4. Advantages of our proposed model

We utilize the matrix  $\mathbf{A}$  to store the regression coefficients of all modalities. By imposing the low-rank constraint on the matrix  $\mathbf{A}$ , our model can better capture the latent correlation among different modalities. Based on the intrinsic structure of different modalities, we can construct a more discriminative classifier for



**Fig. 2.** Illustration of the proposed SPMRM model. After image pre-processing, we can obtain data matrices from the MRI, PET images and CSF biomarkers. By solving problem of Eq. (13), we can get the parameter vector  $\alpha$  of each modal data. In the feature level, the potential correlation among the parameter vectors of different modalities in AD data can be extracted. The process is shown in the lower left corner of Fig. 2. Meanwhile, in the sample level, as the value of  $k$  decreases, we gradually introduce those hard-learning samples into the model. Finally, all samples are considered in building the model. Therefore, the influence of noise samples and outliers can be suppressed in early learning stage. The self-paced weight matrix  $\mathbf{P}$  contains the weights of the samples, whose the elements  $p_i$  ( $i = 1, 2, \dots, n$ ) are inversely proportional to sample residual. The process is shown in the lower right corner of Fig. 2.

AD classification problem. Meanwhile, we introduce the self-paced learning to our model, which benefits the robustness of the optimization algorithm. We define a diagonal weight matrix  $\mathbf{P}$  whose diagonal elements indicate the importance of samples to the fusion model. Motivated by the human learning process, self-paced learning will adaptively learn the model from “easy” samples to more complicated samples. The main idea of self-paced learning is to determine the “easy” samples. In our work, AD classification problem is the supervised model. We employ the difference between predicted values and label to determine the “easy” samples. In the training process, we first select those small residual samples to construct the model. Gradually, we add the samples with larger residuals to train the model till all samples are considered. Our proposed model can be views as a framework of two parameter models. (1) When the self-paced learning is not considered, the model can be regarded as the multi-modal low-rank minimization based method (MRM-based model); (2) when the low-rank is not considered, the model can be regarded as the self-paced learning based method for multi-modal data (SP-based model). It is worth pointing out that our method is not the combination of two independent parameter models. In our method, these two models promote each other. The exploring of latent correlation among different modalities helps determining the real weights in SPL. Meanwhile, SPL considering the contribution of each sample

also benefits the optimization of low-rank problem. The Fig. 2 demonstrates the framework of the proposed SPMRM method.

#### 4. Experiments

##### 4.1. Datasets

The data used in the experiments were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset ([www.loni.ucla.edu](http://www.loni.ucla.edu)). A total of 202 subjects in the ADNI (202-ADNI) with MRI, PET and CSF modalities are used in the study, which includes 50 AD subjects, 53NCs, and 99 MCI subjects. The 99 MCI patients can be further divided into two types, MCI converts and MCI non-converts. Specially, MCI converts (MCI-C) will develop to AD patients within 18 months whereas the MCI non-converts (MCI-NC) will maintain the original status. Meanwhile, another AD database (913-ADNI) is also involved in the experiment. This dataset contains five modalities, ID (serial number), single nucleotide polymorphism (SNPdata), voxel based morphometry (VBM), fluorodeoxyglucose positron emission tomography (FDG) and F-18 florbetapir PET scans amyloid imaging (AV45) with AD, MCI and NC, which includes 160 ADs, 542 MCIs and 211 NCs. The 542 MCI patients have three phases, like significant memory concern (SMC), early mild cognitive impairment (EMCI) and late mild cognitive



**Table 1**  
Demographic Characteristics of the Studied Sample in the 913-ADNI Database. (The values are denoted as mean  $\pm$  standard deviation. NC= Normal Control, SMC=Significant Memory Concern, EMCI=Early Mild Cognitive Impairment, LMCI=Late Mild Cognitive Impairment, AD=Alzheimer's disease.).

Subjects	HC	SMC	EMCI	LMCI	AD
Number	210	82	272	187	160
Gender(M/F)	109/101	33/49	153/119	108/79	95/65
Age	76.13 $\pm$ 6.54	72.45 $\pm$ 5.67	71.51 $\pm$ 7.11	73.86 $\pm$ 8.44	75.18 $\pm$ 7.88
Education	16.44 $\pm$ 2.62	16.78 $\pm$ 2.67	16.07 $\pm$ 2.62	16.38 $\pm$ 2.81	15.86 $\pm$ 2.75

impairment (LMCI). The ID is the unique attribute of the patient and the SNPdata is gene dataset. In our work, we just select the VBM, FDG and AV modalities to construct our model. Table 1 lists the demographic characteristics of subjects in 913-ADNI dataset.

#### 4.2. Image preprocessing

In our experiment, we perform image pre-processing to all MR and PET image on 202-ADNI dataset. Firstly, the anterior commissure (AC) – posterior commissure (PC) correlation on all images is implemented on all images, and then the N3 algorithm [33] is employed to correct the intensity inhomogeneity. Next, we combine the brain surface extractor (BSE) [34] and brain extraction tool (BET) [35] to perform skull-stripping on structural MR images. The skull-stripping results were further manually to ensure clean skull. After removal of cerebellum, FAST in the FSL package [36] is used to divide structural MR images into three different tissues: grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF). Afterward, we utilize the 4D HAMMER [37] which is a fully automatic 4-dimensional atlas warping method to obtain the subject-labeled image based on a template with 93 manually labeled ROIs [38]. Then, all images based on the 93 labeled ROIs in the template can be labeled. For each region of the 93 ROIs in the labeled MR image, we compute the volume of GM as a feature. For PET image, we first align it to its respective MR image of the same subject using a rigid transformation, and then compute the average intensity of each ROI region in the PET image as a feature. Finally, for each subject, we can achieve 93 features in MRI image, 93 features in PET image, and 3 features in CSF image.

For 913-ADNI database, we aligned the preprocessed multi-modality image data (VBM, FDG, AV45) to same visit scan. Then, in the standard Montreal Neurological Institute (MNI) space as  $2 \times 2 \times 2 \text{ mm}^3$  voxels, we created normalized gray matter density maps from MRI data, and registered the FDG-PET and AV45-PET scans into same space by SPM software package [39]. Based on the MarsBaR AAL atlas [40], 116 ROI level measurements of mean gray matter densities. The FDG-PET glucose utilization, and AV45 amyloid values were further extracted. After removing of cerebellum, the imaging measures on each modality (VBM, FDG, and AV45) with 90 ROIs were used as quantitative traits (QTs) in our experiments.

#### 4.3. Experimental settings

To evaluate the effectiveness of our proposed model, we perform two sets of experiments. In the first set of experiment, we aim to validate if the effectiveness of our proposed model is influenced by the combination of the two independent parameter models. We calculate four statistical measures, including the accuracy (ACC), sensitivity (SEN), specificity (SPE) and area under the receiver operating characteristic curve (AUC). Then, we separately use the two independent parameter models to build the classification models. To evaluate the performance of our proposed method, we adopt a 10-fold cross validation strategy to calculate these measures. Specifically, we first divide the dataset  $D$  into ten same size mutually exclusive subsets, like  $D = D_1 \cup D_2 \cup \dots \cup D_{10}$ .

**Table 2**  
MCI-C vs. MCI-NC classification results of two independent parameter models and SPMRM on 202-ADNI dataset.

Method	Modality	MCI-C vs. MCI-NC			
		ACC (%)	SEN (%)	SPE (%)	AUC (%)
SP-based	MRI	69.89	44.76	<b>88.15</b>	58.93
	PET	62.89	30.00	70.00	61.25
	CSF	62.56	50.89	63.39	59.88
	CONCAT	<b>72.78</b>	<b>66.37</b>	79.82	<b>72.92</b>
MRM-based	MRI	70.89	40.00	60.00	66.90
	PET	70.89	40.00	60.00	64.58
	CSF	68.89	60.29	73.57	67.78
	CONCAT	<b>77.89</b>	<b>62.74</b>	<b>77.32</b>	<b>77.87</b>
SPMRM	MRI	70.89	40.00	60.00	69.98
	PET	71.89	40.00	62.50	69.85
	CSF	72.89	68.63	78.57	74.13
	CONCAT	<b>78.89</b>	<b>74.40</b>	<b>83.99</b>	<b>76.91</b>

In each experiment, we select the union of 9 subsets as training set and the remaining subset is viewed as the test set. Then, we can conduct 10 times training and test based on the ten groups of training and test sets. The mean of 10 times test results is treated as the final test result.

In the second set of experiment, we compare our model with other existing baselines in AD classification. Similar to the above experiment, we adopt 10-fold cross-validation in calculating the four statistical measures, ACC, SEN, SPE and AUC. We conduct the comparative experiment on the single-modal and the multi-modal data. In each classification task, we also perform the two independent parameter models (MRM-based model and SP-based model) and our proposed model SPMRM on the dataset.

## 5. Results and discussion

### 5.1. Comparisons with independent parameter models and SPMRM

Tables 2 and 3 list the comparison results of single-modal and multi-modal of the two independent parameter models and SPMRM on 202-ADNI dataset. Specifically, Table 2 shows the MCI-C vs. MCI-NC task, and Table 3 shows the AD vs. NC and MCI vs. NC tasks. Fig. 4(a) demonstrates the ROC curves of single-modal and the multi-modal of SPMRM model in the MCI-C vs. MCI-NC task. Obviously, the performance of multi-modal model is superior to the single-modal model. Meanwhile, compared to the two independent parameter models, the SPMRM model obtains the better experiment results. In 913-ADNI dataset, we perform two classification tasks, AD vs. NC and MCI vs. NC. Table 4 lists the experiment results. As we can observe from the Table 4, we get the similar experiment results as on 202-ADNI dataset.

### 5.2. Performance with different combinations schemes

Fig. 3 shows the performance of our proposed method on different combination of weights, i.e.,  $\beta_{MRI}$ ,  $\beta_{PET}$  and  $\beta_{CSF}$ , in MCI-C vs. MCI-NC task. In the condition of  $\beta_{MRI} + \beta_{PET} + \beta_{CSF} = 1$ , we allot all the possible values, changing from 0 to 1 at a step of 0.1. According to above constraint condition, only the upper triangular

**Table 3**

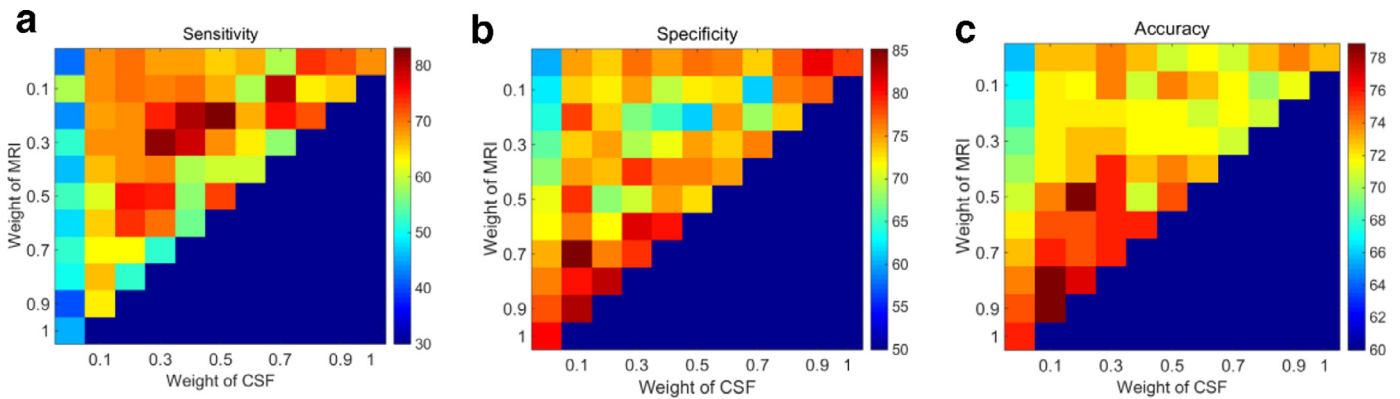
AD vs. NC and MCI vs. NC classification results of two independent parameter models and SPMRM on 202-ADNI dataset.

Method	Modality	AD vs. NC				MCI vs. NC			
		ACC (%)	SEN (%)	SPE (%)	AUC (%)	ACC (%)	SEN (%)	SPE (%)	AUC (%)
SP-based	MRI	89.23	86.00	92.75	91.14	78.25	81.56	71.14	85.02
	PET	90.23	90.00	90.57	93.44	74.25	85.67	51.71	75.74
	CSF	87.46	88.33	86.57	87.33	74.21	77.56	66.57	77.05
	CONCAT	<b>91.23</b>	<b>92.00</b>	<b>90.57</b>	<b>92.04</b>	<b>78.92</b>	<b>83.89</b>	<b>70.86</b>	<b>77.43</b>
MRM-based	MRI	87.46	82.00	93.14	86.92	74.33	84.44	53.71	74.64
	PET	90.69	90.33	91.14	90.91	76.92	91.89	48.86	70.61
	CSF	88.46	88.33	88.57	91.52	70.87	74.56	50.86	73.71
	CONCAT	<b>92.49</b>	<b>92.33</b>	<b>92.57</b>	<b>92.46</b>	<b>80.21</b>	<b>96.00</b>	<b>62.57</b>	<b>75.72</b>
SPMRM	MRI	90.23	86.00	94.57	88.99	78.33	87.33	59.14	80.27
	PET	93.69	94.33	93.14	91.25	81.54	91.00	64.86	72.88
	CSF	89.23	88.33	90.00	92.42	72.21	77.56	60.57	75.50
	CONCAT	<b>95.46</b>	<b>96.33</b>	<b>94.57</b>	<b>96.61</b>	<b>83.54</b>	<b>95.00</b>	<b>62.86</b>	<b>78.15</b>

**Table 4**

Classification results of different parameter models and SPMRM on 913-ADNI dataset.

Methods	Modality	AD vs. NC				MCI vs. NC			
		ACC (%)	SEN (%)	SPE (%)	AUC (%)	ACC (%)	SEN (%)	SPE (%)	AUC (%)
SP-based	VBM	80.55	88.17	71.79	95.73	82.14	93.84	48.78	81.79
	FDG	77.49	84.66	67.67	94.04	83.40	93.48	52.20	88.23
	AV	81.09	88.23	73.37	94.15	81.61	92.48	45.00	84.61
	CONCAT	<b>84.09</b>	<b>90.25</b>	<b>76.48</b>	<b>96.11</b>	<b>83.47</b>	<b>93.98</b>	<b>53.29</b>	<b>88.64</b>
MRM-based	VBM	85.06	92.28	75.10	96.20	82.61	90.81	52.24	84.01
	FDG	79.08	85.97	70.16	95.23	82.47	91.78	50.34	84.79
	AV	84.88	93.06	75.79	95.73	80.01	90.15	44.51	84.02
	CONCAT	<b>85.65</b>	<b>93.58</b>	<b>76.46</b>	<b>96.31</b>	<b>82.93</b>	<b>92.28</b>	<b>53.74</b>	<b>85.71</b>
SPMRM	VBM	87.76	90.78	78.23	95.25	82.14	88.75	40.45	86.23
	FDG	79.78	86.87	69.67	93.88	83.47	94.30	52.50	80.65
	AV	86.79	92.68	75.77	96.43	82.41	89.94	49.95	88.89
	CONCAT	<b>88.02</b>	<b>94.14</b>	<b>80.00</b>	<b>97.21</b>	<b>84.14</b>	<b>94.31</b>	<b>55.26</b>	<b>89.10</b>

**Fig. 3.** MCI classification results with respect to different combing weights of MRI, PET and CSF.

parts have valid values. In Fig. 3, the better experiment results gather around the inner squares of the upper triangular parts. It demonstrates the multi-modal data provide complementary information for AD classification. In other words, each modality in our model is essential to achieve good classification.

### 5.3. Comparison of other AD classification methods

To demonstrate the superiority of our algorithm, we compare the SPMRM with nine baseline algorithms in AD classification, i.e., Canonical Correlation Analysis (CCA) [41], Naïve Bayesian (NB) [42], Random Forest [43], Multi-kernel learning (MKL) [44], Sparse representation classification (SRC) [45], Locality Preserving Projections (LPP) [46], Sparsity Preserving Projections (SPP) [47], Locally Linear Embedding (LLE) [48], Stacked auto-encoders (SAE) [49], Stacked denoising sparse auto-encoder (DSAE) [50], deep

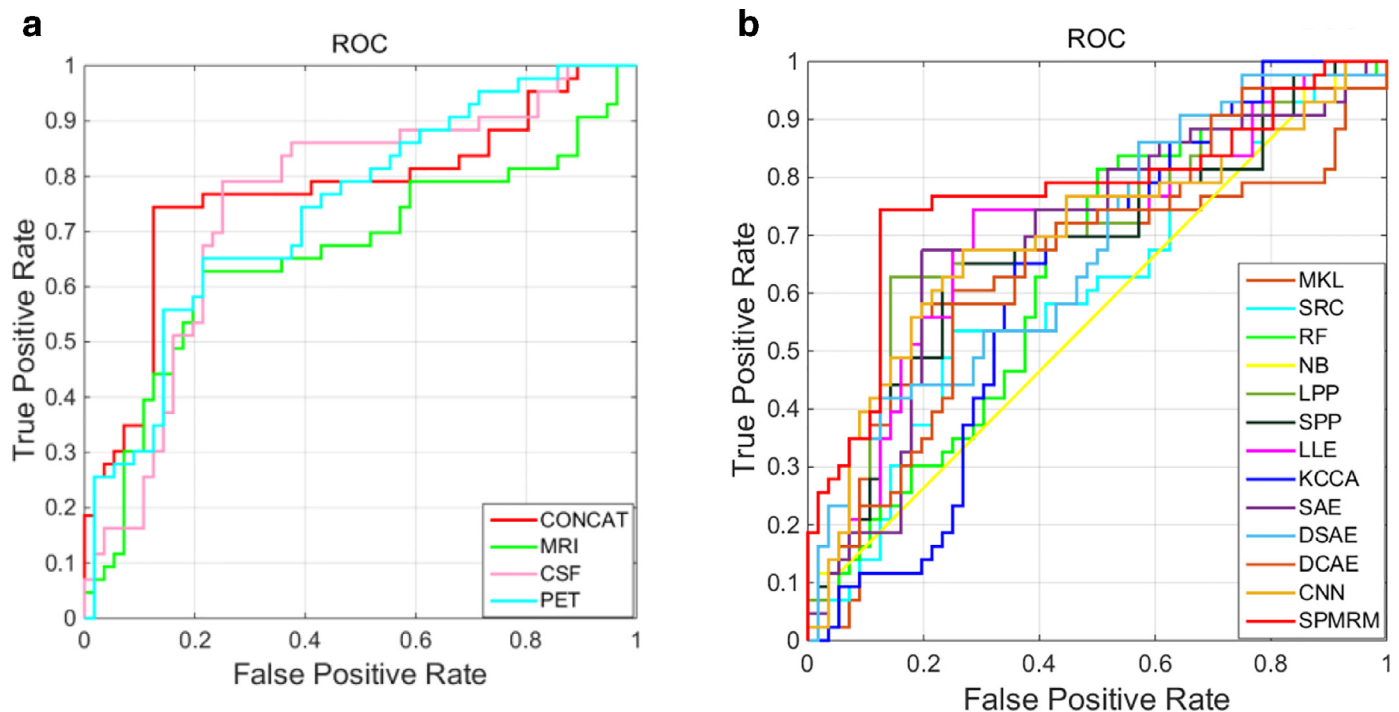
convolutional auto-encoder (DCAE) [51] and convolutional neural network (CNN) [10,52]. Through imposing linear transformation for two variables, CCA method can maximize the correlation in the subspace. LPP, SPP and LLE are the typical manifold feature extraction methods. MKL can construct a fusion classification model by integrating the features in different modalities. Random Forest is a sort of Ensemble Learning (EL), integrating a number of decision trees. A sample will have  $n$  classification results based on the  $n$  decision trees. Random Forest integrates all the results and specifies the major labels as the output. SRC method is the classifier based on sparse representation. Naïve Bayesian classifier can calculate the probability of each category it belongs to and picks the maximum probability category as the final label. Stacked auto-encoders (SAE) is a deep learning algorithm which analyzes multiple classes in one setting with a few labeled samples. DSAE and DCAE are two variations of auto-encoder. CNN is one of the

**Table 5**  
Classification results (ACC and AUC) of baselines and SPMRM on 202-ADNI dataset.

Methods	Tasks					
	AD vs. NC		MCI vs. NC		MCI-C vs. MCI-NC	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
MKL	87.23	88.08	72.25	83.66	70.00	63.91
SRC	87.46	92.19	74.17	80.74	63.67	61.00
Random Forest	86.46	95.71	70.29	81.52	64.89	63.37
Naïve Bayes	85.46	91.06	71.54	72.90	67.78	59.55
LPP	85.23	79.26	72.96	71.62	75.89	70.64
SPP	76.00	69.95	72.88	72.60	70.89	67.65
LLE	79.69	76.40	70.25	73.89	72.67	69.64
KCCA	74.92	81.67	71.46	78.15	58.89	62.33
SAE	91.40	92.89	82.10	83.88	75.00	68.59
DSAE	88.73	82.45	80.91	75.98	75.86	68.31
DCAE	87.47	90.22	78.47	70.37	75.19	65.14
CNN	94.50	88.52	76.60	78.32	74.20	69.37
SPMRM	<b>95.46</b>	<b>96.61</b>	<b>83.54</b>	<b>88.43</b>	<b>78.89</b>	<b>76.91</b>

**Table 6**  
Classification results between baselines and our proposed model SPMRM on 913-ADNI dataset.

Methods	Tasks							
	AD vs. NC				MCI vs. NC			
	ACC (%)	SEN (%)	SPE (%)	AUC (%)	ACC (%)	SEN (%)	SPE (%)	AUC (%)
MKL	79.89	93.93	42.23	92.08	81.34	92.23	40.80	86.44
SRC	74.36	81.95	66.58	88.77	68.55	72.04	54.56	88.15
Random Forest	79.83	86.57	74.12	96.55	82.54	94.26	45.21	89.08
Naïve Bayes	80.27	86.20	72.43	95.09	82.68	91.80	54.91	88.24
LPP	86.27	93.70	78.28	95.99	83.21	89.68	45.84	85.14
SPP	88.02	94.14	80.00	97.21	75.20	83.89	54.96	82.14
LLE	76.33	81.53	71.29	91.12	81.47	93.59	46.30	76.16
KCCA	74.03	90.21	46.76	86.89	79.47	88.43	42.09	82.44
SAE	87.79	87.32	78.47	88.74	77.90	92.92	49.46	73.85
DSAE	81.44	75.21	76.31	80.19	75.19	79.07	50.59	65.32
DCAE	83.05	75.97	78.43	82.33	75.43	64.46	51.69	63.44
CNN	83.80	83.63	74.88	81.97	69.51	74.12	51.41	70.28
SPMRM	<b>89.20</b>	<b>95.56</b>	<b>81.27</b>	<b>99.48</b>	<b>84.14</b>	<b>94.31</b>	<b>55.26</b>	<b>89.10</b>



**Fig. 4.** (a) The ROC curves of three single-modal and the multi-modal methods. (b) The ROC curves of our algorithm and other baselines on 202-ADNI.



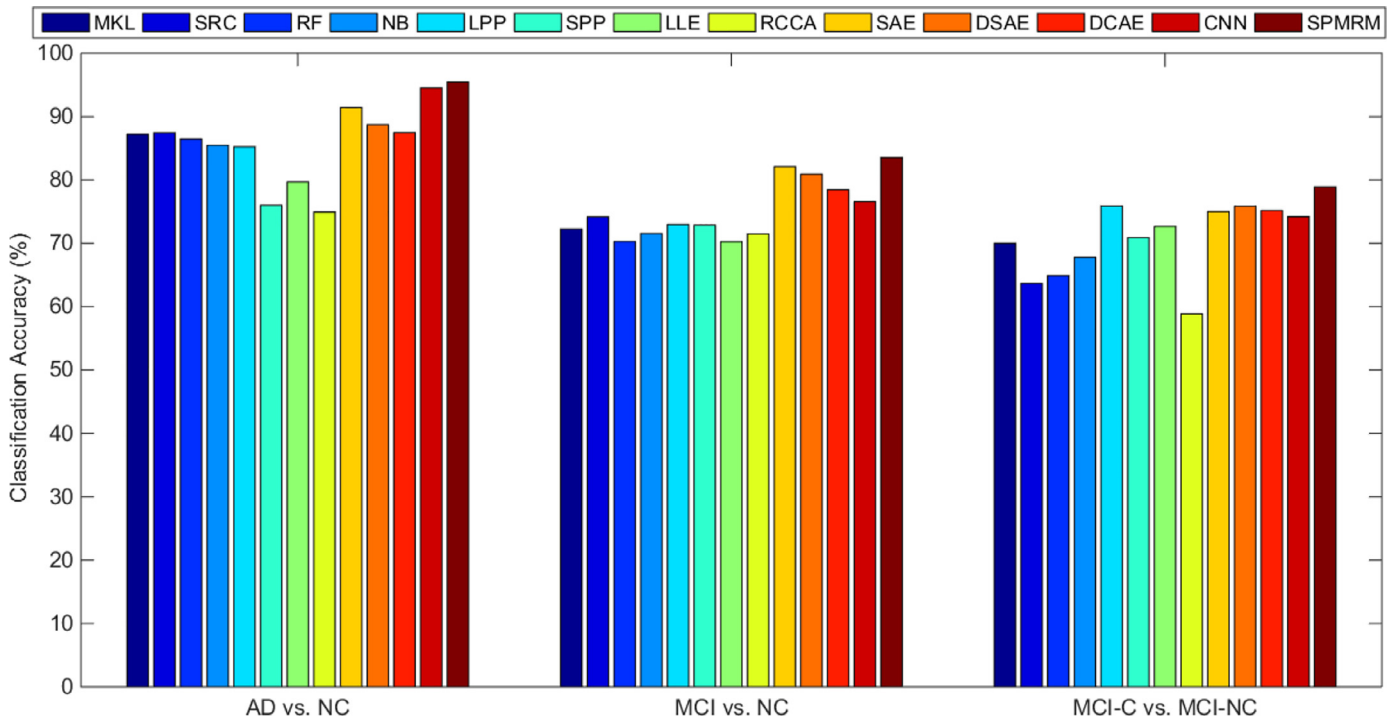


Fig. 5. Comparisons on baselines and our model on ADNI dataset.

most popular models in deep learning and has been applied in AD classification problem in recent years.

The experiment results (the mean of accuracy and AUC) of thirteen algorithms on 202-ADNI dataset are listed in Table 5 and Fig. 5. Table 6 shows the results on 913-ADNI dataset. For different algorithms, we list the highest mean of accuracy (ACC) and area under curve (AUC) in bold. Fig. 4(b) demonstrates the ROC curves of our algorithm and baselines in the task MCI-C vs. MCI-NC on 202-ADNI dataset. In three different tasks, AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC, our algorithm obtains all the best performance on accuracy and AUC. Specifically, our proposed

model achieves the best classification accuracy of 95.46%, 83.54% and 78.89% for the three different tasks in multi-modal data. The highest AUC obtained by our model are 96.61%, 88.43% and 76.91%, respectively.

Compared to baseline algorithms on AD classification problem, our proposed method obtains the higher results in the three tasks. For deep learning models, they usually show superiority over large datasets, such as the recognition of high-definition images. However, the AD classification problem is based on small sample data and it makes the large number of parameters in the deep learning model cannot be optimized, which constrains the effectiveness of

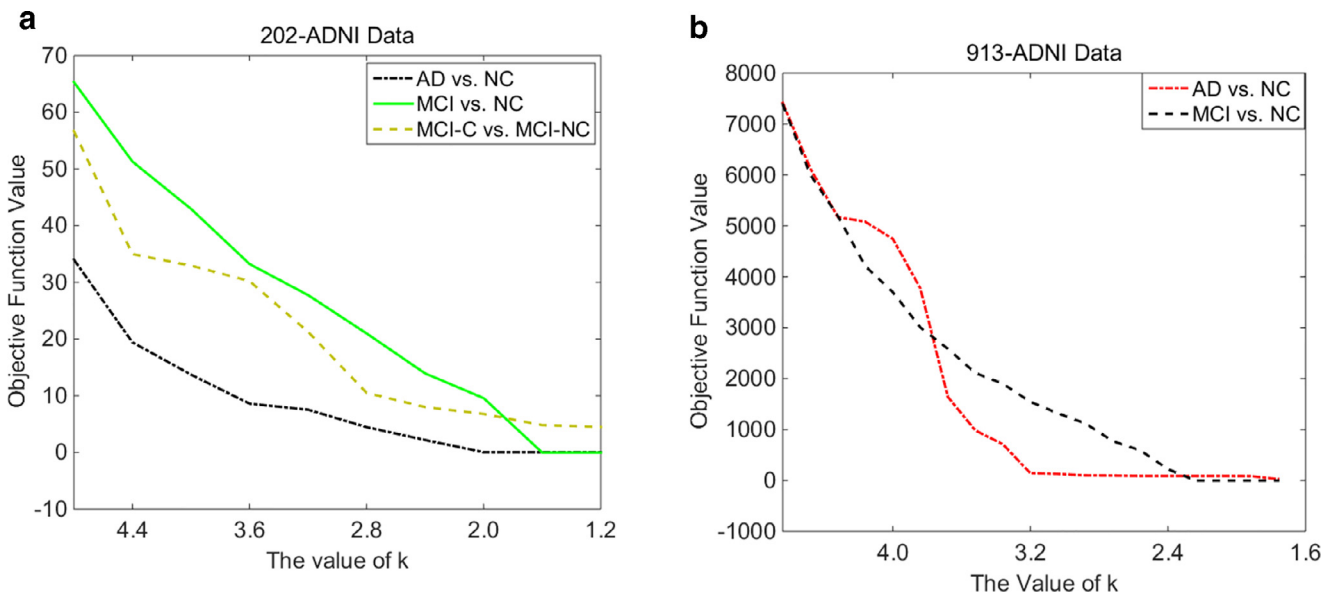


Fig. 6. The objective function values with the self-paced parameter k.

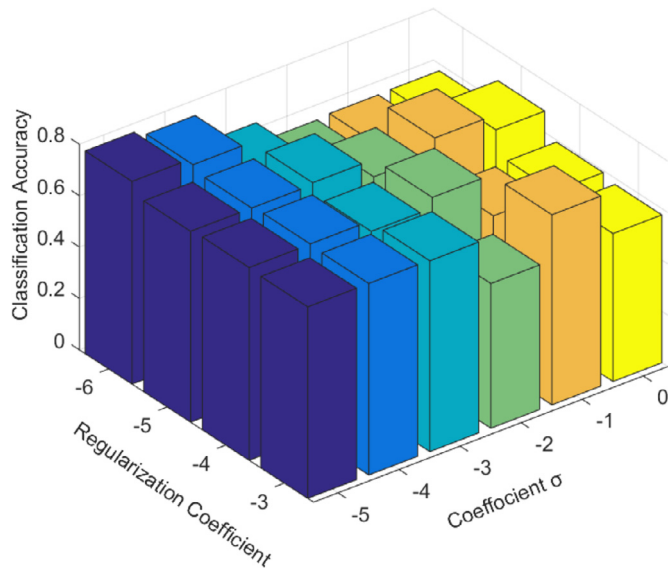


Fig. 7. The classification results of SPMRM with the combination of  $\lambda_S$  and  $\sigma$ .

deep learning models. In our follow-up work, we will adjust some deep learning models to fit our domain.

5.4. The influence of parameters

As the value of  $k$  decreases, the model becomes more mature. To visualize this change, we show the changes of the objective function value as  $k$  decreases in Fig. 6. We set the initial value of  $k$  to 4.8 and the decrease step is 0.4. Obviously, as the value of  $k$

decreases, the value of the objective function also decreases. This also shows that with more and more complicated samples are considered into the model, the objective function tends to convergence.

In SPMRM,  $\lambda_S$  and  $\lambda_R$  are the regularization parameters. In this paper, we employ radial basis function (RBF) to map the original data to a feature space, and the parameter  $\sigma$  is the width parameter of RBF. To evaluate the parameters effect on our algorithm, we measure different combinations of parameters  $\lambda_S$ ,  $\lambda_R$  and  $\sigma$ . Specifically,  $\lambda_S$  varies from  $10^{-i}$  ( $i = -6, -5, -4, -3$ ),  $\lambda_R$  varies from  $0, 5, \dots, 5 \times n$  ( $n = 0, \dots, 10$ ) and  $\sigma$  changes from  $2^{-i}$  ( $i = -6, -5, \dots, 0$ ). From the analysis of experiment results, our method is not sensitive to parameter  $\lambda_R$ , and the change of parameter  $\lambda_R$  almost does not produce any effect on the results. In our work, we fix the parameter  $\lambda_R$  to 10. Fig. 7 shows the experiment results under the influence of parameters  $\lambda_S$  and  $\sigma$ .

5.5. The performance of our model with respect to the number of selected ROI features

The above results have shown the effectiveness of our proposed model based on the whole brain ROI features in the AD classification problem. In this section, we will visually demonstrate the most important feature subsets. Specifically, we employ the value of regression coefficient vector of each modality to measure the importance of each brain region. After performing 10 times 10-fold cross-validation, we can obtain the mean value of the ten regression coefficient vectors, denote as  $\omega^*$ . By normalizing the value of vector  $\omega^*$  to  $[0, 1]$  and sorting it in descending order, we select the top 10 significant brain regions. Fig. 8 shows the top 10 significant brain regions (ROIs) detected from the FDG, VBM and AV modality of AD classification problem in the template space. From Fig. 8, these selected brain regions, i.e., hippocampal, Wernicke, amygdale

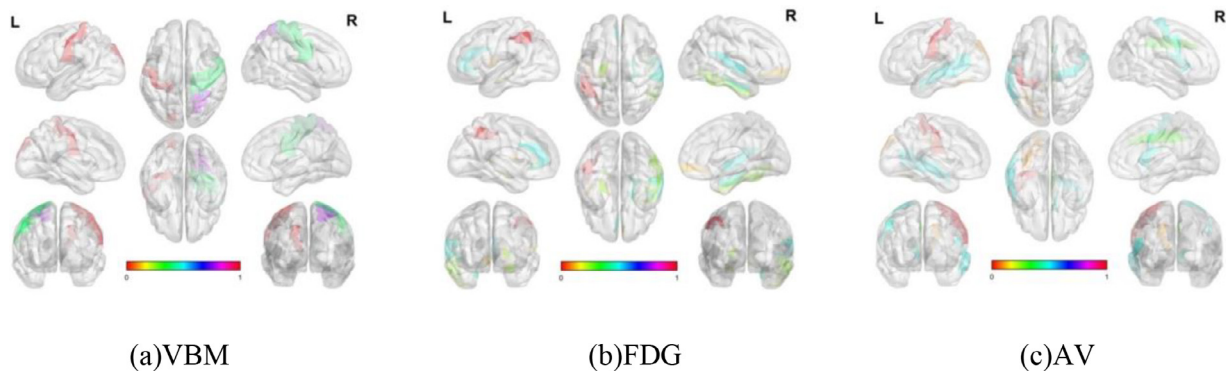


Fig. 8. Top 10 brain regions selected for AD classification detected from VBM, FDG and AV.

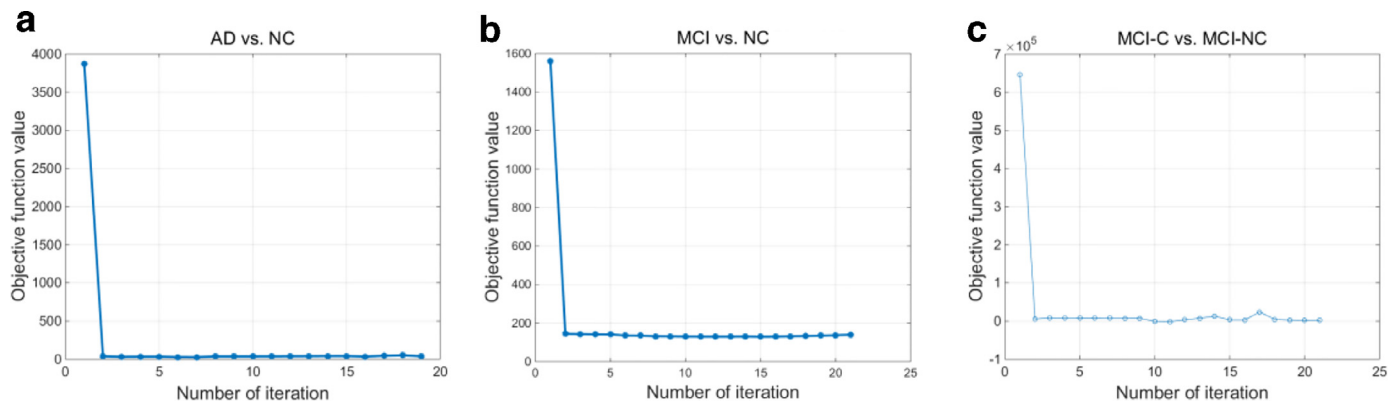


Fig. 9. The convergence property of the proposed algorithm on three classification tasks.

and para-hippocampal regions, have proved to be related to the AD. For example, the changes of the hippocampal will damage the memory of human, which is always affected in AD patients.

### 5.6. Convergence study

In the theoretical analysis of previous section, the objective function of SPMRM will converge to the local optimum in certain steps. In this section, we will show the convergence iterations steps of the proposed method. Fig. 9 demonstrates the variation of objective function value on 202-ADNI dataset in the three classification tasks, AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC. It can be found that SPMRM can converge to a determined value within fairly few iterations.

## 6. Conclusion

This study proposed SPMRM method for multi-modal AD classification problem. Compared to other methods, our model introduces two novel aspects: (1) imposing low-rank representation on AD data to capture the intrinsic latent structure across different modalities and (2) self-paced learning is adopted to adaptively measure the importance of sample to the fusion model. Experiment results demonstrate that our model achieves better performance in accuracy and AUC compared with other state-of-the-art methods on 202-ADNI dataset and 913-ADNI dataset. Benefiting from above two aspects, our proposed algorithm can construct a more discriminative and robust model. In future, we will investigate how to extend the proposed method to solve the incomplete multi-modal AD classification problem.

### Conflict of interest

The authors declare that they have no conflict of interest.

### Acknowledgments

This article is partly supported by National Natural Science Foundation of China (nos. 61501230, 61732006 and 61473149), Natural Science Foundation of Jiangsu Province (no. BK20150751), China Post-doctoral Science Foundation funded project (no. 2015M570446), Jiangsu Planned Projects for Postdoctoral Research Funds (no. 1402047B), Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (no. MJUKF201726) and Fundamental Research Funds for the Central Universities and the Foundation of Graduate Innovation Center in NUAA (KFJJ20171609).

### References

- [1] Y.L. Hsu, P.C. Chung, W.H. Wang, M.C. Pai, C.Y. Wang, C.W. Lin, et al., Gait and balance analysis for patients with Alzheimer's disease using an inertial-sensor-based wearable instrument, *IEEE J. Biomed. Health Inf.* 18 (6) (2014) 1822–1830.
- [2] M. Liu, D. Zhang, D. Shen, Ensemble sparse classification of Alzheimer's disease, *Neuroimage* 60 (2) (2012) 1106–1116.
- [3] A. Veeramuthu, S. Meenakshi, P. Manjusha, A new approach for Alzheimer's disease diagnosis by using association rule over PET images, *Int. J. Comput. Appl.* 91 (9) (2014).
- [4] R. Chaves, D.S. Gonzales, F. Segovia, P. Padilla, Effective diagnosis of Alzheimer's disease by means of association rules, in: *Proceedings of the International Conference on Hybrid Artificial Intelligence Systems*, Springer, 2010, pp. 452–459.
- [5] R. Chaves, J. Ramirez, J. Gorriz, Integrating discretization and association rule-based classification for Alzheimer's disease diagnosis, *Expert Syst. Appl.* 40 (5) (2013) 1571–1578.
- [6] D. Zhang, D. Shen, Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease, *Neuroimage* 59 (2) (2012) 895–907.
- [7] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, Multimodal classification of Alzheimer's disease and mild cognitive impairment, *Neuroimage* 55 (3) (2011) 856–867.
- [8] O.B. Ahmed, J. Benois-Pineau, M. Allard, G. Catheline, C.B. Amar, Recognition of Alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and Multiple Kernel Learning, *Neurocomputing* 220 (2017) 98–110.
- [9] D. Zhang, Q. Zhu, D. Zhang, Multi-modal dimensionality reduction using effective distance, *Neurocomputing* 259 (2017) 130–139.
- [10] S. Sarraf and G. Tofghi, Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks, 2016, arXiv:1603.08631.
- [11] M. Liu, J. Zhang, P.-T. Yap, D. Shen, Diagnosis of Alzheimer's disease using view-aligned hypergraph learning with incomplete multi-modality data, in: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2016, pp. 308–316.
- [12] Z. Sun, Y. Fan, B.P. Lelieveldt, M. van de Giessen, Detection of Alzheimer's disease using group lasso SVM-based region selection, in: *Medical Imaging 2015: Computer-Aided Diagnosis*, 9414, International Society for Optics and Photonics, 2015 941414.
- [13] M. Pagani, D. Salmasso, G. Rodriguez, D. Nardo, F. Nobili, Principal component analysis in mild and moderate Alzheimer's disease—A novel approach to clinical diagnosis, *Psychiatry Res. Neuroimaging* 173 (1) (2009) 8–14.
- [14] X. Chen, Z.J. Wang, M.J. McKeown, A three-step multimodal analysis framework for modeling corticomuscular activity with application to Parkinson's disease, *IEEE J. Biomed. Health Inf.* 18 (4) (2014) 1232–1241.
- [15] B. Haeffele, E. Young, R. Vidal, Structured low-rank matrix factorization: optimality, algorithm, and applications to image processing, in: *Proceedings of the International Conference on Machine Learning*, 2014, pp. 2007–2015.
- [16] Y.-C.F. Wang, C.-P. Wei, C.-F. Chen, Low-rank matrix recovery with structural incoherence for robust face recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2618–2625.
- [17] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: *Proceedings of the Twenty-seventh International Conference on Machine Learning (ICML-10)*, 2010, pp. 663–670.
- [18] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, C. Yuan, Low-rank preserving projections, *IEEE Trans. Cybern.* 46 (8) (2016) 1900–1913.
- [19] W. Yang, Y. Gao, Y. Shi, L. Cao, MRM-lasso: a sparse multiview feature selection method via low-rank analysis, *IEEE Trans. Neural Netw. Learn. Syst.* 26 (11) (2015) 2801–2815.
- [20] X. Cao, Q. Zhao, D. Meng, Y. Chen, Z. Xu, Robust low-rank matrix factorization under general mixture noise distributions, *IEEE Trans. Image Process.* 25 (10) (2016) 4677–4690.
- [21] L. Jiang, D. Meng, T. Mitamura, A.G. Hauptmann, Easy samples first: self-paced reranking for zero-example multimedia search, in: *Proceedings of the Twenty-second ACM International Conference on Multimedia*, ACM, 2014, pp. 547–556.
- [22] L. Lin, K. Wang, D. Meng, W. Zuo, L. Zhang, Active self-paced learning for cost-effective and progressive face identification, *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (1) (2018) 7–19.
- [23] M.P. Kumar, B. Packer, D. Koller, Self-paced learning for latent variable models, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2010, pp. 1189–1197.
- [24] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, A.G. Hauptmann, Self-paced learning for matrix factorization, in: *Proceedings of the AAAI*, 2015, pp. 3196–3202.
- [25] D. Meng, Q. Zhao, L. Jiang, A theoretical understanding of self-paced learning, *Inf. Sci.* 414 (2017) 319–328.
- [26] W. He, H. Zhang, L. Zhang, H. Shen, Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration, *IEEE Trans. Geosci. Remote Sens.* 54 (1) (2016) 178–188.
- [27] Y. Xu, X. Fang, J. Wu, X. Li, D. Zhang, Discriminative transfer subspace learning via low-rank and sparse representation, *IEEE Trans. Image Process.* 25 (2) (2016) 850–863.
- [28] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: *Proceedings of the Twenty-sixth Annual International Conference on Machine Learning*, ACM, 2009, pp. 41–48.
- [29] G. Obozinski, B. Taskar, M. Jordan, Multi-task feature selection, Technical Report, Statistics Department, UC Berkeley, 2006, pp. 1–15.
- [30] Y. Zhang, D.-Y. Yeung, Q. Xu, Probabilistic multi-task feature selection, in: *Proceedings of the Advances in neural information processing systems*, 2010, pp. 2559–2567.
- [31] C.M. Bishop, *Pattern Recognition and Machine Learning*, 103, Publications of the American Statistical Association, 2006, pp. 886–887.
- [32] Z. Lin, M. Chen, and Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, 2010, arXiv:1009.5055.
- [33] J.G. Sled, A.P. Zijdenbos, A.C. Evans, A nonparametric method for automatic correction of intensity nonuniformity in MRI data, *IEEE Trans. Med. Imaging* 17 (1) (1998) 87–97.
- [34] D.W. Shattuck, S.R. Sandor-Leahy, K.A. Schaper, D.A. Rottenberg, R.M. Leahy, Magnetic resonance image tissue classification using a partial volume model, *Neuroimage* 13 (5) (2001) 856–876.
- [35] S.M. Smith, Fast robust automated brain extraction, *Hum. Brain Mapp.* 17 (3) (2002) 143–155.
- [36] Y. Zhang, M. Brady, S. Smith, Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm, *IEEE Trans. Med. Imaging* 20 (1) (2001) 45–57.
- [37] D. Shen, S.M. Resnick, C. Davatzikos, 4D HAMMER image registration method for longitudinal study of brain changes, in: *Proceedings of the Human Brain Mapping*, 2003, pp. 1–8.

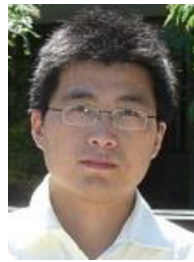
- [38] N.J. Kabani, 3D anatomical atlas of the human brain, *Neuroimage* 7 (1998) P-0717.
- [39] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. ETARD, Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, *Neuroimage* 15 (1) (2002) 273–289.
- [40] J. Ashburner, K.J. Friston, Voxel-based morphometry—the methods, *Neuroimage* 11 (6) (2000) 805–821.
- [41] B. Lei, S. Chen, D. Ni, T. Wang, Discriminative learning for Alzheimer's disease diagnosis via canonical correlation analysis and multimodal fusion, *Front. Aging Neurosci.* 8 (2016) 77.
- [42] S.B. Shree, H. Sheshadri, Diagnosis of Alzheimer's disease using Naive Bayesian Classifier, *Neural Comput. Appl.* 29 (1) (2018) 123–132.
- [43] L. Huang, Y. Jin, Y. Gao, K.-H. Thung, D. Shen, Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest, *Neurobiol. Aging* 46 (2016) 180–191.
- [44] L. Xu, X. Wu, K. Chen, L. Yao, Multi-modality sparse representation-based classification for Alzheimer's disease and mild cognitive impairment, *Comput. Methods Programs Biomed.* 122 (2) (2015) 182–190.
- [45] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, *IEEE Trans. Pattern Anal. Mach. Intell.* 29 (1) (2007) 40–51.
- [46] F. Nie, D. Xu, I.W.-H. Tsang, C. Zhang, Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction, *IEEE Trans. Image Process.* 19 (7) (2010) 1921–1932.
- [47] X. Liu, D. Tosun, M.W. Weiner, N. Schuff, Locally linear embedding (LLE) for MRI based Alzheimer's disease classification, *Neuroimage* 83 (2013) 148–157.
- [48] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, D. Feng, Early diagnosis of Alzheimer's disease with deep learning, in: *Proceedings of the IEEE Eleventh International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2014, pp. 1015–1018.
- [49] S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease, *IEEE Trans. Biomed. Eng.* 62 (4) (2015) 1132–1140.
- [50] B. Shi, Y. Chen, P. Zhang, C.D. Smith, J. Liu, Nonlinear feature transformation and deep fusion for Alzheimer's disease staging analysis, *Pattern Recognit.* 63 (2017) 487–498.
- [51] H. Huang, X. Hu, Y. Zhao, M. Makkie, Q. Dong, S. Zhao, L. Guo, T. Liu, Modeling task fMRI data via deep convolutional autoencoder, *IEEE Trans. Med. Imaging* 37 (7) (2018) 1551–1561.
- [52] A. Khvostikov, K. Aderghal, J. Benois-Pineau, A. Krylov, and G. Catheline, 3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies, 2018. arXiv:1801.05968.



**Jiashuang Huang** is current working toward the Ph.D. degree in computer science from Nanjing University of Aeronautics and Astronautics. His current research interests include machine learning and pattern recognition.



**Xiaoke Hao** received the Ph.D. in Computer Science from Nanjing University of Aeronautics and Astronautics, China, in 2017. His current research interests include machine learning and pattern recognition.



**Daoqiang Zhang**, born in 1978 and received his B.S. degree and Ph.D. degree in Computer Science from Nanjing University of Aeronautics and Astronautics in 1999 and 2004, respectively. And he is a Professor and a PhD supervisor in Nanjing University of Aeronautics and Astronautics at present. His research interests include machine learning, pattern recognition, data mining and medical imaging analysis etc. In these areas, he has published over 100 scientific articles in refereed international journals such as *Neuroimage*, *Pattern Recognition*, *Artificial Intelligence in Medicine*, *IEEE Trans. Neural Networks*; and conference proceedings such as *IJCAI*, *AAAI*, *SDM*, *ICDM*. He is a member of the Machine Learning Society of the Chinese

Association of Artificial Intelligence (CAAI), and the Artificial Intelligence & Pattern Recognition Society of the China Computer Federation (CCF).



**Qi Zhu** received his B.S. degree, M.S. degree, and Ph.D. degree from Harbin Institute of Technology in 2007, 2010 and 2014, respectively. Recently, he is an associate professor at College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. His interests include pattern recognition, feature extraction and medical image analysis.



**Ning Yuan** received the B.S. degree from Jiangsu University of Science and Technology, Jiangsu, China, in 2016. He is current working toward the M.S. degree in computer science from Nanjing University of Aeronautics and Astronautics. His current research interests include machine learning, pattern recognition, medical image processing.