Computerized Medical Imaging and Graphics 80 (2020) 101663

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

# Hypergraph based multi-task feature selection for multimodal classification of Alzheimer's disease

Wei Shao, Yao Peng, Chen Zu, Mingliang Wang,
Daoqiang Zhang*, the Alzheimer's Disease Neuroimaging Initiative

*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China*

ARTICLE INFO

*Article history:*
Received 11 March 2018
Received in revised form 14 June 2019
Accepted 1 October 2019

*Keywords:*
Hypergraph learning
Multi-task learning
Feature selection
Multimodal classification
Alzheimer's disease

ABSTRACT

Multi-modality based classification methods are superior to the single modality based approaches for the automatic diagnosis of the Alzheimer's disease (AD) and mild cognitive impairment (MCI). However, most of the multi-modality based methods usually ignore the structure information of data and simply squeeze them to pairwise relationships. In real-world applications, the relationships among subjects are much more complex than pairwise, and the high-order structure containing more discriminative information will be intuitively beneficial to our learning tasks. In light of this, a hypergraph based multi-task feature selection method for AD/MCI classification is proposed in this paper. Specifically, we first perform feature selection on each modality as a single task and incorporate group-sparsity regularizer to jointly select common features across multiple modalities. Then, we introduce a hypergraph based regularization term for the standard multi-task feature selection to model the high-order structure relationship among subjects. Finally, a multi-kernel support vector machine is adopted to fuse the features selected from different modalities for the final classification. The experimental results on the Alzheimer's Disease Neuroimaging Initiative (ADNI) demonstrate that our proposed method achieves better classification performance than the start-of-art multi-modality based methods.



## 1. Introduction

Alzheimer's disease (AD), which is usually associated with elderly people, is the sixth leading cause of death in the United States (Association, 2012). The progression of AD gradually leads to a widespread loss of mental function such as memory loss, language impairment, disorientation and personality change, ultimately leading to death. In Association (2013), it was reported that the total estimated prevalence of AD is expected to be 13.8 million by 2050. However, no treatment has so far been reported to be able to reverse or stop the progress of AD. Therefore, many studies focus on the early diagnosis of AD and MCI based on neuroimaging data (Sui et al., 2012; Ye et al., 2011) which plays a crucial role in the later treatments.

Neuroimaging which offers great potential to discover features corresponding to the early course of dementing illness is a powerful tool for disease diagnosis in neurodegenerative diseases such as AD. Recently, magnetic resonance imaging (MRI) and positron emission topography (PET) are indicated to be useful to investigate neurophysiological characteristics of AD and MCI (Davatzikos et al., 2011; Fan et al., 2008; Chetelat et al., 2003; Foster et al., 2007). Furthermore, multiple biomarkers have been shown to be sensitive to the diagnosis of AD and MCI by neuroimaging, such as the structural atrophy, pathological amyloid depositions, and metabolic alterations in the brain.

In recent decades, machine learning and pattern recognition methods have been widely used in neuroimaging analysis of AD and MCI, including group-based comparison approaches and individual classification (Ye et al., 2011; Orrú et al., 2012). However, most of the existing studies mainly focus on extracting features from single modality. For example, researchers extracted the features from the structural MRI, such as voxel-wise tissue (Desikan et al., 2009; Fan et al., 2007) and hippocampal volumes (Gerardin et al., 2009) for diagnosis of AD. In addition to the structural MRI, PET images are also utilized for the classification of AD and MCI (Chetelat et al., 2003; Foster et al., 2007; Hinrichs et al., 2009).

However, since the structure and function of brain are very complex, it is challenging to accurately detect all the disease-related

* Corresponding author.
*E-mail addresses:* shaowei20022005@nuaa.edu.cn (W. Shao), pengyao320@nuaa.edu.cn (Y. Peng), chen0zu@gmail.com (C. Zu), wml489@nuaa.edu.cn (M. Wang), dqzhang@nuaa.edu.cn (D. Zhang).

https://doi.org/10.1016/j.compmedimag.2019.101663
0895-6111/

features from single modality. Recently, with the development of biomedical imaging techniques, multi-modality based methods are promising in the field of medical image analysis since multi-modality information is naturally available in the data acquisition procedures of various clinical tasks. Different modalities can provide different views of brain structure or function and reveal different aspects of pathological changes related to AD. For example, structural MRI provides information related to the tissue type of brain, while the FDG-PET measures the cerebral metabolic rate for glucose. Numerous studies have shown that the complementary neuroimaging modalities can help to discover the hidden information which may be ignored by the single modality, and the fusion of the information from different modalities can enhance the diagnostic performance. Hence, multiple modalities are preferred to improve the accuracy of AD diagnosis (Liu et al., 2014, 2015; Suk et al., 2014; Zhu et al., 2014; Gray et al., 2013; Ahmed et al., 2017; Lei et al., 2017). For instance, Liu et al. (2015) and Suk et al. (2014) used two modalities including MRI and PET for the diagnosis of AD. Zhu et al. (2014) combined MRI, PET and cerebrospinal fluid (CSF) for the regression and classification of AD. Gray et al. (2013) used MRI, PET, CSF and categorical genetic information for AD/MCI classification.

Although the existing multi-modality based methods have achieved promising results, there are still some problems which may limit the classification performance. For neuroimaging, even after feature extraction, the feature dimension is relatively high compared to the sample size, and the subsequent classification performance may be poor because of the redundant or irrelative features. Therefore, feature selection which removes the redundant or irrelative features has become an important step in the diagnosis of AD. Some feature selection methods have been used for identifying the corresponding disease-related regions in AD. For example, Zhu et al. (2016) combined two subspace learning methods, namely, linear discriminant analysis and locality preserving projection to select features in neuroimaging. Jie et al. (2015) proposed a manifold regularized multitask feature learning method which uses multi-task learning and manifold based Laplacian regularization to preserve both the intrinsic relatedness among multiple modalities of data and the data distribution information in each modality and thus induces more discriminative features. Zu et al. (2016) proposed a label-aligned multi-task feature learning method which adds a new label-aligned regularization term to the objective function of standard multi-task feature selection to ensure that all multi-modal subjects with the same class labels should be close in the new feature-reduced space.

However, one disadvantage of the existing methods is that they only considered the pairwise relationships between subjects, while ignoring the high-order relationships which are a kind of important prior information for the learning task. In many real-world problems, relationships among the objects of interest are more complex than pairwise. Naively squeezing the complex relationships into pairwise ones inevitably leads to the information loss which can be expected valuable for our learning tasks (Zhou et al., 2007). Intuitively, modeling the high-order relationships among subjects can induce more discriminative features and further boost the performance of subsequent classification. In many applications, researchers use hypergraph to model the complex relationships among subjects, where a hyperedge can connect more than two vertices at the same time and capture the high-order structure. For example, Bu et al. (2010) adopted hypergraph for music recommendation. Hong et al. (2016) proposed to recovery human pose via hypergraph Laplacian.

In this paper, we propose a hypergraph based multi-task feature selection method where a hypergraph-based regularization is developed to explicitly depict the high-order relationship in each modality. Specifically, our proposed method contains three steps: (1) hypergraph construction, (2) hypergraph based multi-task feature learning, (3) multimodal classification. Specifically, we first construct a hypergraph in each modality by constructing multiple hyperedges that reflect the high-order relationships among subjects. Then we treat feature learning in each modality as a single learning task and formulate the multimodal classification as the multi-task learning (MTL) problem. MTL exploits the intrinsic task relatedness, based on which the information from each task can be shared across multiple tasks and thus facilitates the individual task learning. Specifically, the $\ell_{2,1}$ norm is introduced to select features jointly, which can guarantee the features of different modalities in the same brain regions are selected at the same time. Then, we add hypergraph-based regularization terms to the standard multi-task objective function. Finally, we use the multi-kernel support vector machine SVM to fuse the selected features from multimodal data for final classification. To validate our method, we conduct experiments on the ADNI dataset and the experimental results show the efficiency of the proposed method compared with the start-of-the-art methods.

**Table 1**
Characteristics of the subjects.

| Subjects | AD | EMCI | LMCI | NC |
|---|---|---|---|---|
| Number | 160 | 273 | 187 | 160 |
| Gender (M/F) | 95/65 | 153/120 | 108/79 | 96/65 |
| Age (mean ± std) | 75.18 ± 7.88 | 71.48 ± 7.12 | 73.86 ± 8.44 | 76.14 ± 6.53 |
| Education (mean ± std) | 15.86 ± 2.75 | 16.08 ± 2.62 | 16.38 ± 2.81 | 16.45 ± 2.62 |

## 2. Method

Fig. 1 illustrates the framework of the proposed method, which includes three main steps: hypergraph construction, feature selection and classification. In this section, we first introduce the dataset used in our experiments and then details of the proposed method will be given.

### 2.1. Subjects

In this paper, we use the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset for performance evaluation. The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and nonprofit organizations, with a $60 million five-year public private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

Imaging data from 831 ADNI participants with corresponding baseline MRI and FDG-PET data are collected, including 160 AD patients, 273 Early MCI (EMCI) patients, 187 Late MCI (LMCI) patients and 211 normal controls (NC). Table 1 lists the clinical and demographic information for the study population.

Briefly, NC participants have normal cognitive performance and no subjective or informant-based complaint of memory decline. EMCI participants have a memory concerns reported by the subject, memory function approximately 1 standard deviation below normal performance and their Mini-Mental State Examination (MMSE) (Risacher et al., 2015) total score is greater than 24. Besides a subjective memory concern, Clinical Dementia Rating (CDR) on LMCI subjects is 0.5 and Memory Box (MB) score is at least 0.5. MMSE score on AD is between 20 and 26 and CDR is 0.5 or 1.0.

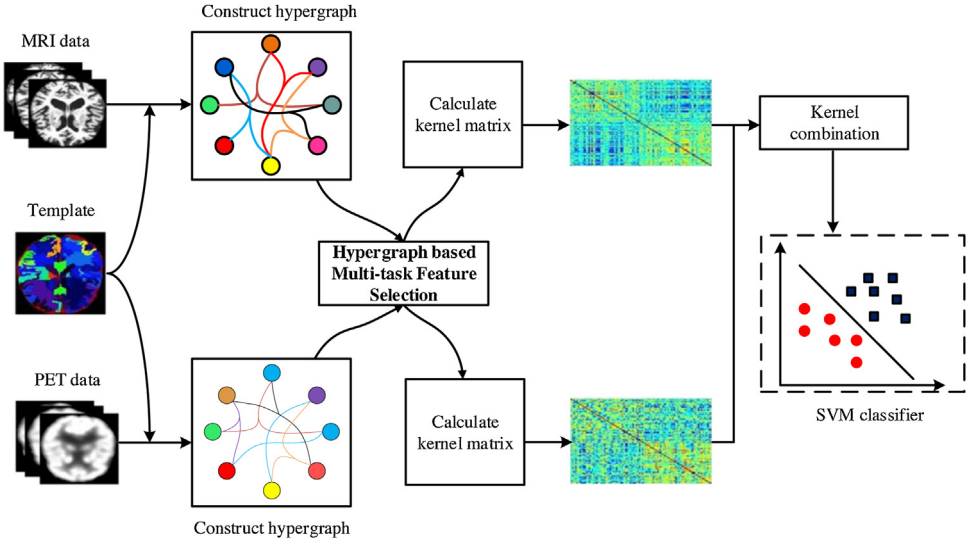



**Fig. 1.** The proposed classification framework. Our proposed method contains three steps: (1) hypergraph construction, (2) hypergraph based multi-task feature selection, (3) multimodal classification.

We aligned the preprocessed multimodal imaging data (MRI, FDG) to each participants with the same visit scan, and then created normalized gray matter density maps from MRI data in the standard Montreal Neurological Institute (MNI) space as $2 \times 2 \times 2\,\text{mm}^3$ voxels, and registered the FDG-PET scans into the same space by SPM software package (Friston and Ashburner, 2007). 116 ROI level measurements of mean gray matter densities, FDG-PET glucose utilization was further extracted based on the MarsBaR AAL atlas (Tzourio-Mazoyer et al., 2002). After the removal of cerebellum, the imaging measures on each modality (MRI and FDG) of 90 ROIs were used in our experiments. All the measures were pre-adjusted for age, gender, and education.

## 2.2. Related work

### 2.2.1. Hypergraph learning

In machine learning problems, we often assume that there are pairwise relationships between subjects, which can be naturally illustrated as a graph. In graph learning, a vertex represents a subject, and any two vertices that have some kinds of relationship are jointed together by an edge. The graph can be undirected or directed, depending on whether the pairwise relationships between subjects are symmetric or not.

However, in many practical problems, relationships among the subjects of interest are more complex than pairwise and cannot be illustrated by traditional graphs. Naively squeezing the complex relationships into pairwise ones inevitably lose information which can be expected valuable for our learning tasks (Zhou et al., 2007). Traditional methods which deal with pairwise relationships cannot properly model these high-order relationships. Thus, hypergraph learning has been proposed, and it has been widely used in various applications (Gao et al., 2012; Ji et al., 2014).

Different from traditional graph, a hypergraph can connect more than two vertices by a hyperedge. In other words, a hyperedge is a subset of vertices. It is obvious that a simple graph is a special kind of hypergraph with each edge only connecting two vertices. Hence, we can model complex relationships (i.e., high-order relationships) among subjects of our interest by using hypergraph (Zhou et al., 2007; Bu et al., 2010; Wang and Zhang, 2008). As in the left of Fig. 2, the vertices contained in one hyperedge are connected by the edges with the same color and each hyperedge can connect more than two vertices. Specially, when the hyperedge contains two vertices, it degenerates into the normal edge.

Let $G(V, E, a)$ represents a hypergraph, where $V$ is the set of vertices, $E$ is the set of hyperedges and $a$ is the set of hyperedge weights. Here, each hyperedge $e_i (i = 1, \ldots, N_e)$ is assigned a weight $a(e_i)$. For the hypergraph $G$, its incidence matrix $\mathbf{H}$, as shown in the right of Fig. 2, is used to represent the relationships between hyperedges and vertices. Specifically, the $(i, j)$th entry of matrix $\mathbf{H}$ denotes weather the $j$th hyperedge contains the $i$th vertex. The matrix $\mathbf{H}$ is defined as follows:

$$\mathbf{H}(v, e) = \begin{cases} 1, & \text{if } v \in e \\ 0, & \text{if } v \notin e \end{cases} \tag{1}$$

According to the incidence matrix $\mathbf{H}$, the degree of each vertex $v$ and each hyperedge $e$ are defined as follows respectively:

$$d(v) = \sum_{e \in E} w(e)\mathbf{H}(v, e) \tag{2}$$

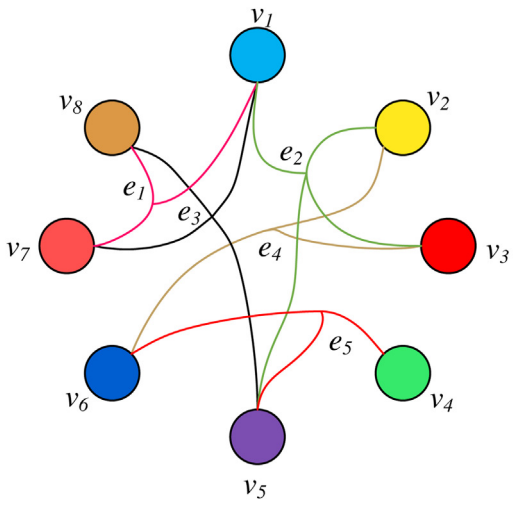
$$\delta(e) = \sum_{v \in V} \mathbf{H}(v, e) \tag{3}$$

In addition, let $\mathbf{D}_v$ and $\mathbf{D}_e$ denote the diagonal matrices containing the vertex and hyperedge degrees respectively. $\mathbf{A}$ denote the diagonal matrix containing the weights of hyperedges and $\mathbf{A}_{ii} = a(e_i)$.

In recent decades, there are several methods to compute the hypergraph Laplacian, which can be roughly divided into two categories. The first one is to build a simple graph in the original hypergraph, such as star expansion (Zien et al., 1999). The second one is to analogy simple graph Laplacian matrix to define the hypergraph Laplacian (Zhou et al., 2007). The above two methods are quite similar. In this paper, we adopt the method proposed in Zhou et al. (2007) to compute the hypergraph Laplacian matrix:

$$\mathbf{L}^h = \mathbf{I} - \Theta \tag{4}$$

$\mathbf{L}^h$ is the hypergraph Laplacian matrix, where $I$ is the identity matrix and $\Theta = \mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{A}\mathbf{D}_e^{-1}\mathbf{H}^T\mathbf{D}_v^{-1/2}$. The construction of the hypergraph is the most critical step in hypergraph learning. According to the method proposed in Zhou et al. (2007), we adopt the K nearest neighbor strategy to construct the hypergraph. Specifically, we first take each vertex as a center node and calculate the Euclidean distance between the center and other vertices. Then we construct a hypergraph by connecting the center and its $k$ nearest vertices.




| | $e_1$ | $e_2$ | $e_3$ | $e_4$ | $e_5$ |
|---|---|---|---|---|---|
| $v_1$ | 1 | 1 | 1 | 0 | 0 |
| $v_2$ | 0 | 1 | 0 | 1 | 0 |
| $v_3$ | 0 | 1 | 0 | 1 | 0 |
| $v_4$ | 0 | 0 | 0 | 0 | 1 |
| $v_5$ | 0 | 1 | 1 | 0 | 1 |
| $v_6$ | 0 | 0 | 0 | 1 | 1 |
| $v_7$ | 1 | 0 | 1 | 0 | 0 |
| $v_8$ | 1 | 0 | 0 | 0 | 0 |

H

**Fig. 2.** Hypergraph model. The left represents one example of hypergraph. The vertices represents different samples, and the vertices contained in one hyperedge are connected by the edges with the same color. The right represents the incidence matrix of the left hypergraph. The *entry*$(v_i, e_j)$ is set to 1 if $v_i$ is contained in $e_j$, and 0 otherwise.

Given $n$ subjects, we can construct $n$ hyperedges. In this paper, each hyperedge is assigned a weight valued 1, i.e., $a(e_i) = 1$. And we fix the value of $k$ as 7.

#### 2.2.2. Multi-task feature selection

Multi-task learning (Jie et al., 2015; Zu et al., 2016; Shao et al., 2019) learns several related tasks at the same time and improves the performance by learning the relevant information between the multiple tasks. In recent years, multi-task learning has been widely used in many fields, such as medical diagnosis, biomedical information, text classification and so on (Cao et al., 2017; Gao et al., 2016; Zhang et al., 2017). In this paper, the feature selection of each modality in medical image is considered as a task. Suppose that we have $M$ learning tasks (i.e., M modalities). Denote $\mathbf{X}_m = [\mathbf{x}_m^1, \mathbf{x}_m^2, \ldots, \mathbf{x}_m^N]^T \in \mathbb{R}^{N\times d}$ as the training data matrix in the $m$th task (i.e., $m$th modality), where $\mathbf{x}_m^i$ is a column vector which represents the corresponding feature vector of the $i$th subject, $d$ is the dimension of features and $N$ is the number of subjects. Let $\mathbf{Y} = [y_1, y_2, \ldots, y_N]^T \in \mathbb{R}^N$ be the corresponding label vector of the $N$ subjects. And the value of $y_i$ is 1 or $-1$ (i.e., patient or normal control). It is worth noting that the label of different modalities from the same subject is the same. We use a linear function to fit the class label and thus the objective function of the multi-task feature selection model is as follows (Yuan and Lin, 2006):

$$\min_{\mathbf{W}} \frac{1}{2}\sum_{m=1}^{M} \|\mathbf{Y} - \mathbf{X}_m\mathbf{w}_m\|_2^2 + \lambda_1\|\mathbf{W}\|_{2,1} \tag{5}$$

where $\mathbf{w}_m \in \mathbb{R}^d$ is the vector of the regression coefficient associated with the $m$th modality. And the vectors of all the $M$ modalities form a weight matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M] \in \mathbb{R}^{d\times M}$. In Eq. (5), $\|\mathbf{W}\|_{2,1}$ is the $\ell_{2,1}$ norm of matrix $\mathbf{W}$, which is defined as follows: $\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{d} \|\mathbf{w}^i\|_2$, where $\mathbf{w}^i$ is the $i$th row of matrix $\mathbf{W}$ corresponding to the $i$th feature across different tasks. In Eq. (5), the optimal solution assigns a relatively large weight to the informative features and zero or a small weight to uninformative or less informative features, resulting in many rows are all zeroes for the characteristic of group sparsity. For feature selection, only those features with nonzero coefficients are kept. In other words, this norm combines multiple tasks and ensures that a small number of common features can be jointly selected across different tasks, considering the relatedness between different tasks. The parameter $\lambda_1$ is the regularization coefficient, which controls the relative weight of the two terms. It is worth noting that when the number of tasks equals to one, Eq. (5) will degenerate into the least absolute shrinkage and selection operator (LASSO) model (Tibshirani, 2011).

### 2.3. Hypergraph-based multi-task feature selection

#### 2.3.1. Hypergraph regularized multi-task model

Based on the hypergraph Laplacian matrix, we define the hypergraph regularization term as follows:

$$\Omega = \mathbf{w}^T\mathbf{X}^T\mathbf{L}^h\mathbf{X}\mathbf{w} \tag{6}$$

Intuitively, we want to preserve the structure information of data in the original feature space and the high-order structure information is reflected by the hypergraph Laplacian matrix $\mathbf{L}^h$. We construct a hypergraph in each modality to preserve the high-order structure of data, and define the objective function of hypergraph based multi-task feature selection as follows:

$$\begin{aligned}\min_{w} \frac{1}{2}\sum_{m=1}^{M} \|\mathbf{Y} - \mathbf{X}_m\mathbf{w}_m\|_2^2 + \lambda\|\mathbf{W}\|_{2,1} \\ + \kappa\mathbf{w}_1^T\mathbf{X}_1^T\mathbf{L}_1^h\mathbf{X}_1\mathbf{w}_1 + \mu\mathbf{w}_2^T\mathbf{X}_2^T\mathbf{L}_2^h\mathbf{X}_2\mathbf{w}_2\end{aligned} \tag{7}$$

where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2]$, $M = 2$. $\mathbf{L}_m^h$ is the hypergraph Laplacian matrix of the $m$th modality. In Eq. (7), the first term is the empirical risk error on the training set and the second term is the $\ell_{2,1} - norm$. The last two terms are the hypergraph regularization terms.

In our model, considering the relatedness across different tasks, we can not only jointly select the common features in the different modalities, but also preserve the high-order structure information in each modality by adding hypergraph regularization terms. The existing multi-modal feature selection algorithm only consider the pairwise relationships between subjects, while ignoring the high-order relationships among subjects and thus may lose some valuable information.

#### 2.3.2. Optimization

It is worth noting that the objective function defined in Eq. (7) includes both the group-sparsity and the hypergraph regularization, which cannot be solved by the existing sparse learning models. Fortunately, there are many methods that can be used to solve the objective function in Eq. (7), such as ADMM and APG algorithms. In this paper, we adopt APG algorithm to solve our problem (Nesterov,

2013; Chen et al., 2009). We divide objective function in Eq. (7) into smooth part

$$f(\mathbf{W}) = \frac{1}{2}\sum_{m=1}^{M}\|\mathbf{Y}-\mathbf{X}_m\mathbf{w}_m\|_2^2 + \kappa\mathbf{w}_1^T\mathbf{X}_1^T\mathbf{L}_1^h\mathbf{X}_1\mathbf{w}_1 + \mu\mathbf{w}_2^T\mathbf{X}_2^T\mathbf{L}_2^h\mathbf{X}_2\mathbf{w}_2 \tag{8}$$

and non-smooth term

$$g(\mathbf{W}) = \lambda\|\mathbf{W}\|_{2,1} \tag{9}$$

Then we use the following function to approximate $f(\mathbf{W})+g(\mathbf{W})$:

$$\Omega_l(\mathbf{W},\mathbf{W}^{(k)}) = f(\mathbf{W}^{(k)}) + \langle\mathbf{W}-\mathbf{W}^{(k)}, \nabla f(\mathbf{W}^{(k)})\rangle + \frac{1}{2}\|\mathbf{W}-\mathbf{W}^{(k)}\|_F^2 + g(\mathbf{W}) \tag{10}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\nabla f(\mathbf{W}^{(k)})$ is the gradient of $f(\mathbf{W})$ at point $\mathbf{W}^{(k)}$ of the $k$th iteration and $l$ is the step size. Finally, the update step of the APG algorithm is as follows:

$$\mathbf{W}^{(k+1)} = \underset{\mathbf{W}}{\operatorname{argmin}}\frac{1}{2}\|\mathbf{W}-\mathbf{U}^{(k)}\|_F^2 + \frac{1}{l}g(\mathbf{W}) \tag{11}$$

where $l$ can be obtained by line search, $\mathbf{U}^{(k)} = \mathbf{W}^{(k)} - \frac{1}{l}\nabla f(\mathbf{W}^{(k)})$. The key to APG algorithm is how to compute the update steps efficiently. It is reported that this problem can be decomposed into $d$ independent sub-problems, and the solution to these sub-problems can be easily obtained. In addition, according to the method proposed by Chen et al. (2009), we can use $\mathbf{Q}^{(k)}$ to calculate $\Omega_l(\mathbf{W},\mathbf{W}^{(k)})$, and $\mathbf{Q}^{(k)}$ is defined as follows:

$$\mathbf{Q}^{(k)} = \mathbf{W}^{(k)} + \eta_k(\mathbf{W}^{(k)} - \mathbf{W}^{(k-1)}) \tag{12}$$

where $\eta_k = \frac{(1-\gamma_{k-1})\gamma_k}{\gamma_{k-1}}$, $\gamma_k = \frac{2}{k+3}$. The convergence rate of this algorithm is $O(1/K^2)$, where $K$ is the maximum number of iterations. We also share the implementation of our method at the address https://github.com//shaoweinuaa//HM2TFS.

#### 2.3.3. Multi-kernel support vector machine

Following Zhang et al. (2011) and Shao et al. (2015), we adopted the multi-kernel support vector machine for final classification. Given the training set, the kernel function of the $m$th modality for subjects $\mathbf{x}_m^i$ and $\mathbf{x}_m^j$ is $k_m(\mathbf{x}_m^i,\mathbf{x}_m^j) = \phi_m(\mathbf{x}_m^i)^T\phi(\mathbf{x}_m^j)$. Linear combination kernel $k(\mathbf{x}^i,\mathbf{x}^j) = \sum_{m=1}^{M}\beta_m k_m(\mathbf{x}_m^i,\mathbf{x}_m^j)$ is adopted as a mixed kernel for fusing information from different modalities. $\beta_m$ is the combination coefficient. In this paper, the SVM classifier can be solved by using LIBSVM toolbox (Chang and Lin, 2011). To find the optimal values of $\beta_m$, we use a coarse-grid search in range [0, 1] through cross-validation on the training samples.

## 3. Experiments and results

To validate the effectiveness of our proposed method, we perform experiments in different scenarios, including AD vs. NC, LMCI vs. NC and EMCI vs. LMCI. Classification performance is accessed on the MRI and FDG-PET modalities from 831 ADNI participants. In our experiments, 10-fold cross-validation strategy is adopted to evaluate the classification performance. Specifically, the whole set is equally divided into 10 subsets. For each cross-validation, we take 9 subsets as the training set and the remaining one as the test set. The process is independently repeated 10 times to eliminate the bias introduced by randomly dividing dataset. We evaluate the performance by computing the classification accuracy (ACC), the sensitivity (SEN), the specificity (SPE) as well as the area under receiver operating characteristic curve (AUC). The classification accuracy (ACC) is computed as the proportion of subjects that are correctly classified among all subjects. The sensitivity (SEN) represents the proportion of patients that are correctly predicted and the specificity (SPE) measures the proportion of normal controls that are correctly classified.

For classification, we implement the linear SVM based on the LIBSVM library (Chang and Lin, 2011) with the default parameters. The optimal value of parameter $\lambda$, $\kappa$, $\mu$ and the weights $\beta_m$ in multi-kernel classifier are determined by another 10-fold cross-validation on the training set.

Our method is compared with three existing multi-kernel based methods, including multi-kernel method (Zhang et al., 2011) without performing feature selection (denote as Baseline), multi-kernel method with LASSO (Tibshirani, 2011) feature selection performed on each single modality (denote as SMFS) and multi-kernel method with manifold regularization (denote as M2TFS) proposed in Jie et al. (2015). In addition, we also compare our method with the $t$-test based feature selection method, in which we firstly concatenate the multi-modal data (i.e., 90 features from MRI-VBM and 90 features from FDG-PET) into a 180-dimensional vector, and then perform $t$-test to sequentially evaluate whether the values of each feature is significantly different (i.e., $p$-value $< 0.05$) from the values of the same feature across different classes (i.e., AD vs. NC, LMCI vs. NC, EMCI vs. LMCI). Finally, we feed all the significantly different features to the SVM classifier for final classification. Similarly, we also perform the LASSO (Tibshirani, 2011) feature selection algorithm on the concatenated 180 dimensional vector, donated as LASSO. For model selection, the regularization parameters for all methods are chosen from the range $\{10^{(-10)}, 10^{(-9)}, \ldots, 10^{0}\}$. The detailed experimental results are summarized in Table 2. Fig. 3 plots the ROC curves of all the methods.

### 3.1. Classification performance

The classification results of AD vs. NC are displayed in Table 2 and Fig. 3. As observed, our proposed method consistently outperformed the other methods in three scenarios. Specifically, our proposed method achieved the accuracy of 92.51%, 82.53%, 75.48% in AD vs. NC, LMCI vs. NC, EMCI vs. LMCI, respectively. The accuracy achieved by other methods are 90.70%, 77.93%, 72.20% respectively. In addition, the AUC of our proposed method are 0.95, 0.80, 0.71 respectively, indicating our method has better diagnostic performance. Also, in most cases, our proposed method achieved higher sensitivity than other methods. High sensitivity is very important for disease diagnosis. It is worth noting that in our experiments, the difference between sensitivity and specificity is large, e.g., a relatively high sensitivity but low specificity for each method. In medical diagnosis, the cost is different for misclassifying a patient into a normal control and misclassifying a normal control into a patient. Obviously, the cost of the former is higher and may delay the treatment. So large sensitivity and low specificity are advantageous for medical diagnosis.

### 3.2. Brain regions involved in classification

It is of great importance to identify the brain regions which are related to the disease. As the brain regions selected in each cross-validation are not the same, we count the top 10 most frequently selected regions in the classification of LMCI and NC as the most discriminative features. The top 10 regions are *hippocampus_left*, *amygdala*, *supp_motor_area_left*, *thalamus_left*, *parietal_inf_left*, *temporal_pole_sup_left*, *rectus_right*, *insula_right*, *cuneus_left*, *frontal_inf_orb_right*. Fig. 4 plots these regions in the template space. As can be seen from Fig. 4, most of the selected regions, such as *hippocampus*, *amygdala*, *temporal pole*, are consistent with the previous studies. It is reported that the *hippocampus* is closely related to long-term memory (Wolf et al.,

**Table 2**
Classification performance of different methods

| Method | AD vs. NC | | | | LMCI vs. NC | | | | EMCI vs. LMCI | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | SEN | SPE | AUC | ACC | SEN | SPE | AUC | ACC | SEN | SPE | AUC |
| *t*-test | 83.23 | 84.64 | 81.37 | 0.89 | 67.76 | 60.75 | 73.98 | 0.73 | 65.22 | 75.20 | 50.64 | 0.65 |
| Baseline | 87.84 | 90.71 | 84.06 | 0.91 | 71.23 | 63.42 | 78.15 | 0.76 | 69.27 | 88.25 | 41.24 | 0.66 |
| Lasso | 87.19 | 92.43 | 80.47 | 0.91 | 73.24 | 71.48 | 75.16 | 0.75 | 70.53 | 89.65 | 42.64 | 0.70 |
| SMFS | 88.25 | 92.68 | 83.61 | 0.92 | 75.60 | 67.96 | 82.36 | 0.75 | 71.07 | **90.73** | 42.40 | 0.67 |
| M2TFS | 90.70 | 93.79 | 86.63 | 0.93 | 77.93 | 68.12 | **86.63** | 0.76 | 72.20 | 90.21 | 45.88 | 0.67 |
| Hypergraph | **92.51** | **94.08** | **90.44** | **0.95** | **82.53** | **86.09** | 78.55 | **0.80** | **75.48** | 83.84 | **63.26** | **0.71** |

Here, the best classification results are displayed with bold fonts.

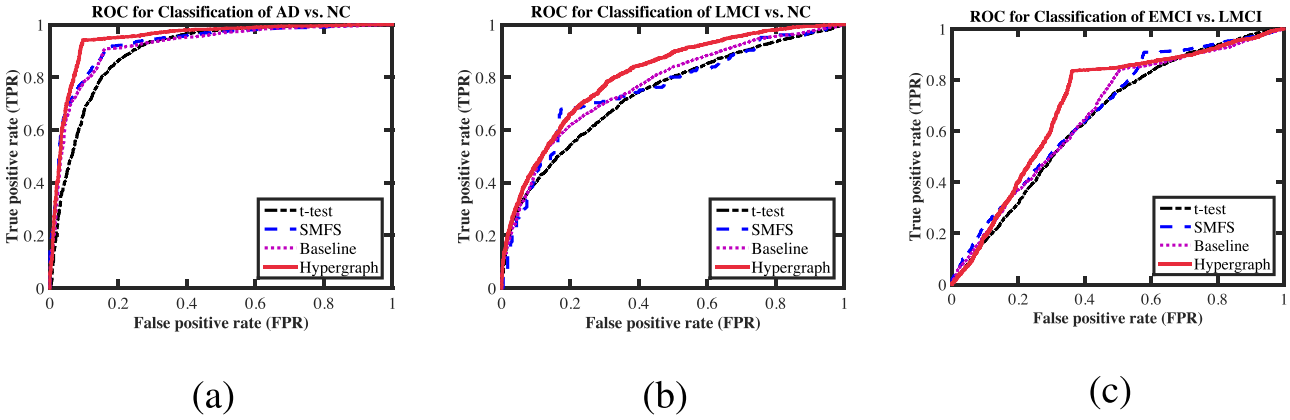



**Fig. 3.** The ROC Curves of the methods: (a) the classification of AD vs. NC, (b) the classification of LMCI vs. NC, (c) the classification of EMCI vs. LMCI. The horizontal axis represents the false positive rate; the vertical axis represents the true positive rate. The area under the curve (AUC) indicates the diagnosis power.

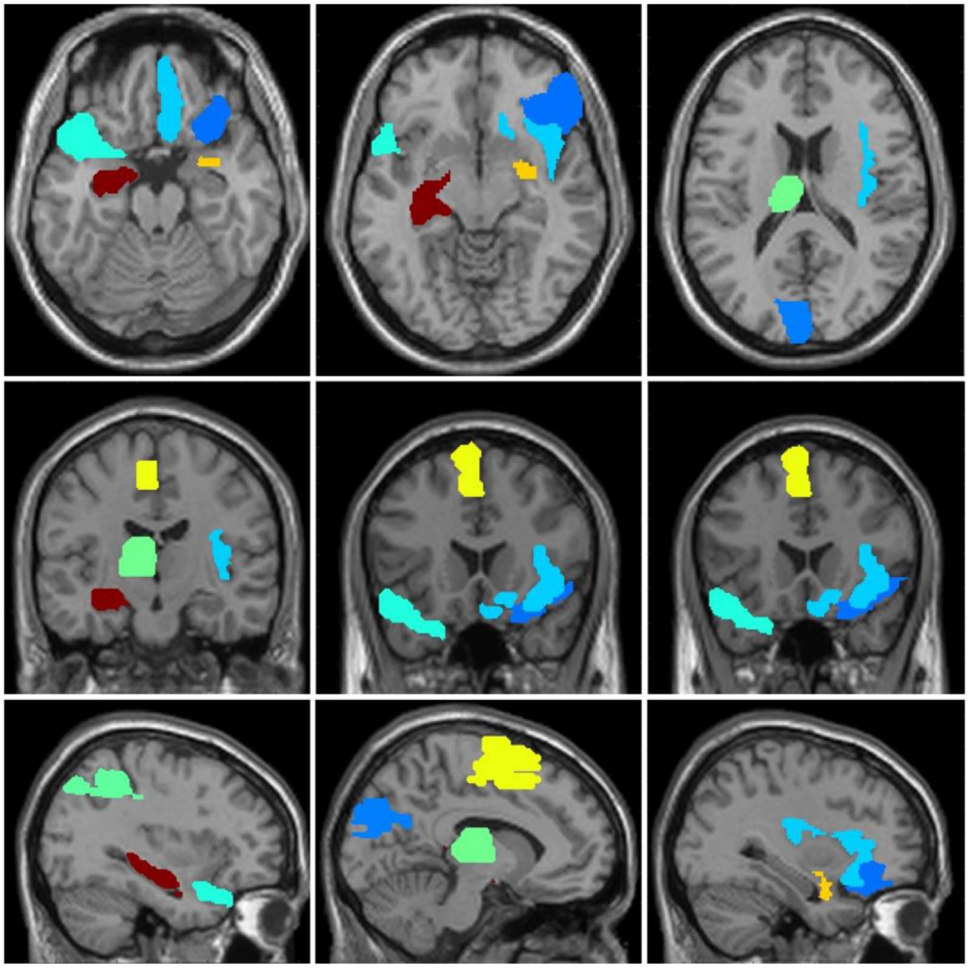


**Fig. 4.** Top 10 ROIs selected by the proposed method for LMCI diagnosis.

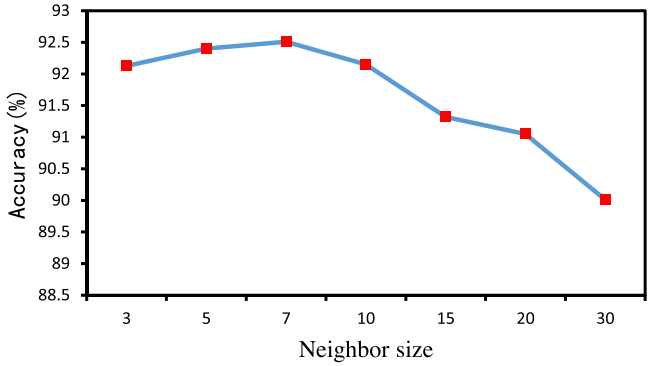



**Fig. 5.** The classification results on different neighbor size of hyperedges. The horizontal axis represents the neighbor size; the vertical axis represents the classification accuracy for AD diagnosis.

2003; Poulin et al., 2011) and the *amygdala* is related to AD. In addition, we also find it that the selected supplementary motor areas (i.e., *supp_motor_area_left*) and the thalamus (i.e., *thalamus_left*) are parts of the sensorimotor network, which has been suggested to be associated to the cognition, information processing speed, mood, and gait speed (Rosano et al., 2012). The selected insula and cuneus are parts of the default mode network, and many studies suggest that the abnormal topological organization of insula and cuneus in the default mode network are related to MCI (Brugere et al., 2018; Esposito et al., 2013; Grothe et al., 2016). Moreover, the inferior frontal gyrus (i.e., *frontal_inf_orb_right*) is one of the most complex areas in the human brain, and the study in Waskom and Wagner (2017) verifies that it is involved in memory functions via its diffuse anatomical networks. Last but not least, our model also could identify the region of rectus gyrus that is located on the orbital surface of the frontal lobe. The study in Wee and Yap (20111) indicate that the morphology changes in rectus gyrus will affect the progression of MCI.

## 4. Discussion

In this paper, we propose a new multi-modality based classification framework, that is, hypergraph based multimodal classification. Our proposed method includes three steps, hypergraph construction, hypergraph based multi-task feature selection and multi-kernel classification. To validate the effectiveness our proposed method, three sets of experiments are performed on the 831 subjects of ADNI dataset. The results show that the proposed method can not only take advantage of the multi-modality complementary information for AD classification but also help in discovering the disease-related biomarkers for understanding the pathological mechanism of AD.

### 4.1. The influence of neighbor size

As mentioned in Section 2.2.2, we use the KNN method to construct hypergraph. Specifically, we choose $k$ nearest vertices to generate a hyperedge. Now we discuss the influence of the neighbor size in KNN of our proposed method. The result is shown in Fig. 5, where the neighbor size varies in range {3, 5, 7, 10, 15, 20, 30}. As can be seen from Fig. 5, when the neighbor size is 7, the result is best. But when the neighbor size is larger than 7, the accuracy decreases rapidly. The underlying reason may be that when the neighbor size is large, it depicts the global structure rather than the local structure of the subjects. And if the neighbor size is large, the hyperedges may contain subjects of different classes, which may not be good to reflect the true data structure.

### 4.2. Hypergraph learning

As a generalization of simple graphs, hypergraphs are typically used to depict the high-order and multiple relations among subjects. Compared with the simple graph, where an edge can only connect two vertices, hypergraph conveys more information through a set of hyperedges that connects more than two vertices at the same time. In this way, hypergraph structure can capture high-order relationships among different subjects.

Hypergraph learning has been used in various applications, such as image and object retrieval and recognition. A few recent studies has incorporated hypergraph learning into medical diagnosis (Liu et al., 2017a; Gao et al., 2015). For example, Liu et al. (2017b) proposed a hypergraph regularized sparse feature learning method. Specifically, they construct a hypergraph on the ADNI data, and compute hypergraph Laplacian term. Then they introduce the hypergraph Laplacian term into the Lasso framework to select features for the subsequent classification. It is worth noting that the feature selection step in Liu et al. (2017b) ignores the potential relationship among different modalities. Liu et al. (2017a) proposed a multimodal centralized hypergraph learning to identify MCI (mild cognitive impairment) and NC (normal controls). They use four MR sequences to construct hypergraph separately and then use centralized hypergraph learning to generate a centralized relevance matrix for each imaging modality. Classification of a subject is determined by the corresponding value in the final fused relevance matrix. However, they do not consider the common information shared by different modalities.

In contrast, our proposed hypergraph-based multi-task learning method not only utilize the high-order relationship among subjects but also consider the complementary information provided by different modalities. Our proposed method is evaluated on ADNI database using the MRI and FDG-PET data for three scenarios of classification including AD vs. NC, LMCI vs. NC, LMCI vs. EMCI. The experimental results demonstrate the effectiveness of our proposed method.

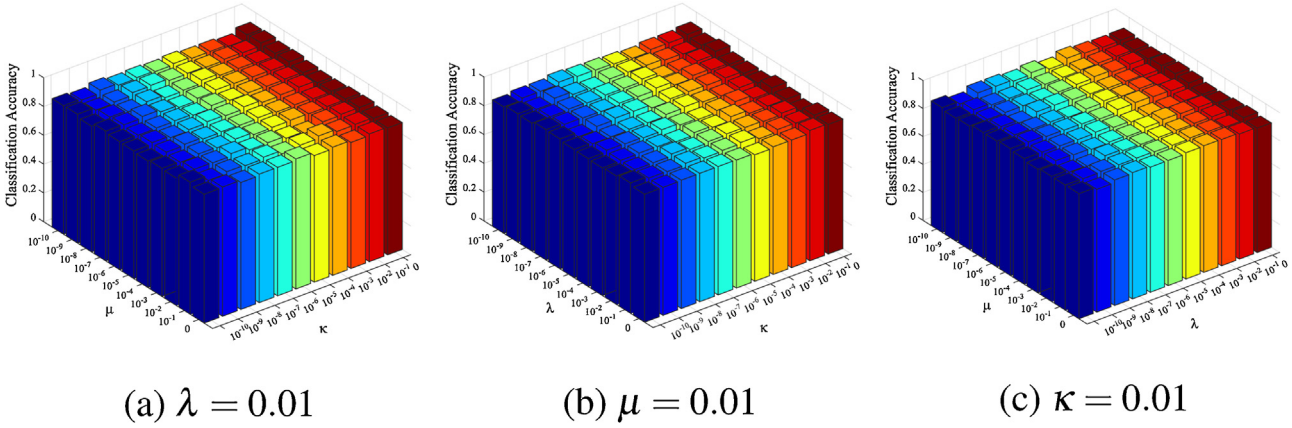
### 4.3. Multimodal classification

Recent studies on the diagnosis of AD and MCI have shown that different modalities can provide complementary information for helping identifying AD (Landau et al., 2010; Du et al., 2007; Morris et al., 2001). It has been reported that information fusion from different modalities can enhance the diagnosis performance. Different approaches have been proposed to fuse biomarkers of different modalities to produce more powerful classifiers (Zhang et al., 2011; Hinrichs et al., 2011; Young et al., 2013). A straightforward way to combine multimodal data is to concatenate features obtained from different modalities into a long vector for each subject. For example, Kohannim et al. (2010) concatenated features from multiple modalities and then a support vector machine was used for the classification of AD. Another way to fuse the different modalities is the kernel-based approach, such as multi-kernel learning (Zhang et al., 2011). An individual kernel matrix is calculated for each modality and a combined kernel matrix is obtained by a linear combination of them. Several results have shown the latter can achieve better performance than the former. Simple concatenation of different modalities cannot optimally integrate the multimodal data (Zhang et al., 2011) and may have bias introduced by the modality which has more features.

To evaluate the effectiveness of multi-modality classification, we use single modality for comparison. We performed experiments on MRI and PET data respectively, using our proposed classification framework. The corresponding results are shown in Table 3. As we can see, our proposed method with two modalities data achieved better performance than the single modality-based methods. The

**Table 3**
Comparison on classification results of different modalities

| Method | AD vs. NC | | | | LMCI vs. NC | | | | EMCI vs. LMCI | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | SEN | SPE | AUC | ACC | SEN | SPE | AUC | ACC | SEN | SPE | AUC |
| PET | 87.33 | 90.90 | 82.63 | 87.95 | 65.61 | 79.19 | 48.01 | 60.19 | 67.61 | 81.17 | 47.81 | 64.04 |
| MRI | 88.30 | 90.09 | 85.94 | 91.15 | 69.14 | 79.34 | 55.92 | 66.24 | 68.52 | **84.83** | 44.71 | 63.53 |
| MRI+PET | **92.51** | **94.08** | **90.44** | **94.57** | **82.53** | **86.09** | **78.55** | **80.32** | **75.48** | 83.84 | **63.26** | **70.34** |




(a) $\lambda = 0.01$ (b) $\mu = 0.01$ (c) $\kappa = 0.01$

**Fig. 6.** Classification accuracy of AD vs.NC classification with respect to different parameter values. We fix one parameter to 0.01 respectively and vary the other two in range $\{10^{(-10)}, 10^{(-9)}, \ldots, 10^{0}\}$.The *X*-axis and *Y*-axis represent the diverse value of parameters and the *Z*-axis represents the classification accuracy for AD diagnosis.

results further validate the multimodal data contain complementary information and can achieve better classification performance than single modality-based methods.

### 4.4. The effect of regularization parameters

In our method, there are three regularization parameters to be tuned, i.e., $\lambda, \kappa, \mu$, which balance the relative contribution of the group-sparsity regularizer and two hypergraph regularizer calculated on the two modalities respectively. In this sub-section, we investigate the effects of the regularization parameters on the classification performance. Specifically, we first fix the value of $\lambda$ to 0.01 and vary $\kappa$ and $\mu$ in range $\{10^{(-10)}, 10^{(-9)}, \ldots, 10^{0}\}$. Then we fix $\kappa$ as 0.01 and vary $\lambda$ and $\mu$ in range $\{10^{(-10)}, 10^{(-9)}, \ldots, 10^{0}\}$. Finally, we fix the value of $\mu$ to 0.01 and vary $\kappa$ and $\lambda$ in $\{10^{(-10)}, 10^{(-9)}, \ldots, 10^{0}\}$.The corresponding results are shown in Fig. 6(a)–(c) respectively. As we can see, our proposed method slightly fluctuates when varying the parameters $\lambda, \kappa, \mu$, indicating our method is not particularly sensitive to the parameter values.

### 4.5. Limitations

The current study has some limitations. First, we fixed the weights of all hyperedges to 1 without considering the importance of different hyperedges. Maybe the patterns in some hyperedges are more important than others for the final classification task. Second, we only investigated the binary classification problems and did not test the performance of multiclass classification. Although the multi-class classification is more difficult than binary classification, it is more important to diagnose the different stage of the disease. Finally, we only considered the MRI and FDG-PET images with the same number of features. But there are some other modalities with different number of features which may also carry important information that can further improve the accuracy of the classification.

## 5. Conclusion

In summary, this paper proposed a novel multi-task learning method to jointly select features from multimodal neuroimaging data for AD/MCI classification. MTL captures the intrinsic task relatedness, based on which the information from each task can be shared across multiple tasks. By introducing the hypergraph based regularization term into the multi-task learning framework, the proposed method can utilize the high-order relationships among the subjects to seek the most discriminative brain regions related to the disease. The experiment results on ADNI dataset demonstrate the effectiveness of our methods. In this study, we fix the weights of all hyperedges to 1, while the importance of hyperedges may be different. It is interesting to optimize the weights of hyperedges, which will be our future work.

## Conflict of interest

The authors of this manuscript have nothing to disclose.

## Acknowledgements

This work was supported by the the National Natural Science Foundation of China (Nos. 61902183, 61876082, 61861130366, 61732006) and National Key Research and Development Program of China (Nos. 2018YFC2001600, 2018YFC2001602), the Royal Society-Academy of Medical Sciences Newton Advanced Fellowship (No. NAFR1180371), and China Postdoctoral Science Foundation funded project (No. 2019M661831)

## References

Ahmed, O.B., Benois-Pineau, J., Allard, M., Catheline, G., Amar, C.B., Initiative, A.D.N., 2017. Recognition of Alzheimer's disease and mild cognitive impairment with multimodal image-derived biomarkers and multiple kernel learning. Neurocomputing 220, 98–110.

Association, A., 2012. 2012 Alzheimer's disease facts and figures. Alzheimer's Dementia 8 (2), 131–168.

Association, A., 2013. 2013 Alzheimer's disease facts and figures. Alzheimer's Dementia 9 (2), 208–245.
Brugere, I., Gallagher, B., Berger-Wolf, T., 2018. Network structure inference, a survey: motivations, methods, and applications. ACM Comput. Surv. 51 (1), 24–30.
Bu, J., Tan, S., Chen, C., Wang, C., Wu, H., Zhang, L., He, X., 2010. Music recommendation by unified hypergraph: combining social media information and music content. Proceedings of the 18th ACM International Conference on Multimedia, 391–400.
Cao, P., Shan, X., Zhao, D., Huang, M., Zaiane, O., 2017. Sparse shared structure based multi-task learning for MRI based cognitive performance prediction of Alzheimer's disease. Pattern Recogn. 72, 219–235.
Chang, C.-C., Lin, C.-J., 2011. Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3), 27.
Chen, X., Pan, W., Kwok, J.T., Carbonell, J.G., 2009. Accelerated gradient method for multi-task sparse learning problem. Ninth IEEE International Conference on Data Mining, ICDM'09, 746–751.
Chetelat, G., Desgranges, B., De La Sayette, V., Viader, F., Eustache, F., Baron, J.-C., 2003. Mild cognitive impairment: can FDG-PET predict who is to rapidly convert to Alzheimer's disease? Neurology 60 (8), 1374–1377.
Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., 2011. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol. Aging 32 (12), 2322.e19–2322.e27.
Desikan, R.S., Cabral, H.J., Hess, C.P., Dillon, W.P., Glastonbury, C.M., Weiner, M.W., Schmansky, N.J., Greve, D.N., Salat, D.H., Buckner, R.L., 2009. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. Brain 132 (8), 2048–2057.
Du, A.-T., Schuff, N., Kramer, J.H., Rosen, H.J., Gorno-Tempini, M.L., Rankin, K., Miller, B.L., Weiner, M.W., 2007. Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. Brain 130 (4), 1159–1166.
Esposito, R., Mosca, A., Valentina, P., 2013. Characterization of resting state activity in MCI individuals. J. Life Environ. Sci. 1 (1), 134–140.
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 (1), 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 (2), 277–285.
Foster, N.L., Heidebrink, J.L., Clark, C.M., Jagust, W.J., Arnold, S.E., Barbas, N.R., DeCarli, C.S., Scott Turner, R., Koeppe, R.A., Higdon, R., 2007. FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease. Brain 130 (10), 2616–2635.
Friston, K., Ashburner, J., 2007. Voxel-Based Morphometry. Academic Press, London, UK, pp. 92–98, Book Section Ch07.
Gao, Y., Wang, M., Tao, D., Ji, R., Dai, Q., 2012. 3-d object retrieval and recognition with hypergraph analysis. IEEE Trans. Image Process. 21 (9), 4290–4303.
Gao, Y., Wee, C.-Y., Kim, M., Giannakopoulos, P., Montandon, M.-L., Haller, S., Shen, D., 2015. MCI identification by joint learning on multiple MRI data. International Conference on Medical Image Computing and Computer-Assisted Intervention, 78–85.
Gao, Y., Shao, Y., Lian, J., Wang, A.Z., Chen, R.C., Shen, D., 2016. Accurate segmentation of CT male pelvic organs via regression-based deformable models and multi-task random forests. IEEE Trans. Med. Imaging 35 (6), 1532–1543.
Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.-S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. NeuroImage 47 (4), 1476–1486.
Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D., Initiative, A.D.N., 2013. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. NeuroImage 65, 167–175.
Grothe, M., Teipel, S., Initiative, A.D.N., 2016. Spatial patterns of atrophy, hypometabolism, and amyloid deposition in Alzheimer's disease correspond to dissociable functional brain networks. Human Brain Mapp. 37 (1), 35–53.
Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M.K., Johnson, S.C., the Alzheimer's Disease Neuroimaging Initiative, 2009. Spatially augmented lpboosting for ad classification with evaluations on the ADNI dataset. NeuroImage 48 (1), 138–149.
Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., the Alzheimer's Disease Neuroimaging Initiative, 2011. Predictive markers for ad in a multi-modality framework: an analysis of MCI progression in the ADNI population. NeuroImage 55 (2), 574–589.
Hong, C., Chen, X., Wang, X., Tang, C., 2016. Hypergraph regularized autoencoder for image-based 3d human pose recovery. Signal Process. 124, 132–140.
Ji, R., Gao, Y., Hong, R., Liu, Q., Tao, D., Li, X., 2014. Spectral-spatial constraint hyperspectral image classification. IEEE Trans. Geosci. Rem. Sens. 52 (3), 1811–1824.
Jie, B., Zhang, D., Cheng, B., Shen, D., 2015. Manifold regularized multitask feature learning for multimodality disease classification. Human Brain Mapp. 36 (2), 489–507.
Kohannim, O., Hua, X., Hibar, D.P., Lee, S., Chou, Y.-Y., Toga, A.W., Jack, C.R., Weiner, M.W., Thompson, P.M., Initiative, A.D.N., 2010. Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiol. Aging 31 (8), 1429–1442.
Landau, S., Harvey, D., Madison, C., Reiman, E., Foster, N., Aisen, P., Petersen, R.C., Shaw, L., Trojanowski, J., Jack, C., 2010. Comparing predictors of conversion and decline in mild cognitive impairment. Neurology 75 (3), 230–238.
Lei, B., Yang, P., Wang, T., Chen, S., Ni, D., 2017. Relational-regularized discriminative sparse learning for Alzheimer's disease diagnosis. IEEE Trans. Cybern. 47 (4), 1102–1113.
Liu, M., Zhang, D., Shen, D., 2014. Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. Human Brain Mapp. 35 (4), 1305–1319.
Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., Fulham, M.J., 2015. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans. Biomed. Eng. 62 (4), 1132–1140.
Liu, M., Zhang, J., Yap, P.-T., Shen, D., 2017a. View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data. Med. Image Anal. 36, 123–134.
Liu, M., Zhang, J., Guo, X., Cao, L., 2017b. Hypergraph regularized sparse feature learning. Neurocomputing 237, 185–192.
Morris, J.C., Storandt, M., Miller, J.P., McKeel, D.W., Price, J.L., Rubin, E.H., Berg, L., 2001. Mild cognitive impairment represents early-stage Alzheimer disease. Arch. Neurol. 58 (3), 397–405.
Nesterov, Y., 2013. Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer Science & Business Media, New York, USA.
Orrú, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A., 2012. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci. Biobehav. Rev. 36 (4), 1140–1152.
Poulin, S.P., Dautoff, R., Morris, J.C., Barrett, L.F., Dickerson, B.C., the Alzheimer's Disease Neuroimaging Initiative, 2011. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. Psychiatry Res.: Neuroimaging 194 (1), 7–13.
Risacher, S.L., Kim, S., Nho, K., Foroud, T., Shen, L., Petersen, R.C., Jack, C.R., Beckett, L.A., Aisen, P.S., Koeppe, R.A., 2015. APOE effect on Alzheimer's disease biomarkers in older adults with significant memory concern. Alzheimer's Dementia 11 (12), 1417–1429.
Rosano, C., Bennett, D., Newman, A., 2012. Patterns of focal gray matter atrophy are associated with bradykinesia and gait disturbances in older adults. J. Gerontol. Ser. A: Biomed. Sci. Med. Sci. 67 (1), 957–962.
Shao, W., Liu, M., Zhang, D., 2015. Human cell structure-driven model construction for predicting protein subcellular location from biological images. Bioinformatics 32 (1), 114–121.
Shao, W., Liu, M., Shen, H., Zhang, D., 2019. An organelle correlation-guided feature selection approach for classifying multi-label subcellular bio-images. IEEE/ACM Trans. Comput. Biol. Bioinformat. 15 (3), 828–838.
Sui, J., Adali, T., Yu, Q., Chen, J., Calhoun, V.D., 2012. A review of multivariate methods for multimodal fusion of brain imaging data. J. Neurosci. Methods 204 (1), 68–81.
Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A.D.N., 2014. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage 101, 569–582.
Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. J. R. Stat. Soc.: Ser. B Stat. Methodol. 73 (3), 273–282.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15 (1), 273–289.
Wang, F., Zhang, C., 2008. Label propagation through linear neighborhoods. IEEE Trans. Knowl. Data Eng. 20 (1), 55–67.
Waskom, M., Wagner, A., 2017. Distributed representation of context by intrinsic subnetworks in prefrontal cortex. Proc. Natl. Acad. Sci. U.S.A. 114 (1), 2030–2035.
Wee, C., Yap, P., 20111. Enriched white matter connectivity networks for accurate identification of MCI patients. NeuroImage 54 (1), 1812–1822.
Wolf, H., Jelic, V., Gertz, H., Nordberg, A., Julin, P., Wahlund, L., 2003. A critical discussion of the role of neuroimaging in mild cognitive impairment. Acta Neurol. Scand. 107 (s179), 52–76.
Ye, J., Wu, T., Li, J., Chen, K., 2011. Machine learning approaches for the neuroimaging study of Alzheimer's disease. Computer 44 (4), 99–101.
Young, J., Modat, M., Cardoso, M.J., Mendelson, A., Cash, D., Ourselin, S., the Alzheimer's Disease Neuroimaging Initiative, 2013. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. NeuroImage: Clinical 2, 735–745.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc.: Ser. B Stat. Methodol. 68 (1), 49–67.
Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., the Alzheimer's Disease Neuroimaging Initiative, 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. NeuroImage 55 (3), 856–867.
Zhang, H., Xiao, L., Wang, Y., Jin, Y., 2017. A Generalized Recurrent Neural Architecture for Text Classification With Multi-Task Learning. arXiv:1707.02892.

Zhou, D., Huang, J., Schölkopf, B., 2007. Learning with hypergraphs: clustering, classification, and embedding. Advances in Neural Information Processing Systems, 1601–1608.
Zhu, X., Suk, H.-I., Shen, D., 2014. A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. NeuroImage 100, 91–105.
Zhu, X., Suk, H.-I., Lee, S.-W., Shen, D., 2016. Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. IEEE Trans. Biomed. Eng. 63 (3), 607–618.
Zien, J.Y., Schlag, M.D., Chan, P.K., 1999. Multilevel spectral hypergraph partitioning with arbitrary vertex sizes. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 18 (9), 1389–1399.
Zu, C., Jie, B., Liu, M., Chen, S., Shen, D., Zhang, D., Initiative, A.D.N., 2016. Label-aligned multi-task feature learning for multimodal classification of Alzheimer's disease and mild cognitive impairment. Brain Imaging Behav. 10 (4), 1148–1159.