Pattern Recognition 126 (2022) 108566

Contents lists available at ScienceDirect

Pattern Recognition

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

ASMFS: Adaptive-similarity-based multi-modality feature selection for classification of Alzheimer's disease

Yuang Shi[a], Chen Zu[b], Mei Hong[a], Luping Zhou[c], Lei Wang[d], Xi Wu[e], Jiliu Zhou[a,e], Daoqiang Zhang[f], Yan Wang[a,*]

[a] School of Computer Science, Sichuan University, Chengdu, China
[b] Risk Controlling Research Department, JD.com, Chengdu, China
[c] School of Electrical and Information Engineering, University of Sydney, Australia
[d] School of Computing and Information Technology, University of Wollongong, Australia
[e] School of Computer Science, Chengdu University of Information Technology, Chengdu, China
[f] School of Computer Science, Nanjing University of Aeronautics and Astronautics, Nanjing, China

ARTICLE INFO

Article history:
Received 14 February 2020
Revised 27 January 2022
Accepted 30 January 2022
Available online 2 February 2022

Keywords:
Multi-modality
Similarity learning
Feature selection
Alzheimer's disease

ABSTRACT

Multimodal classification methods using different modalities have great advantages over traditional single-modality-based ones for the diagnosis of Alzheimer's disease (AD) and its prodromal stage mild cognitive impairment (MCI). With the increasing amount of high-dimensional heterogeneous data to be processed, multi-modality feature selection has become a crucial research direction for AD classification. However, traditional methods usually depict the data structure using pre-defined similarity matrix as a priori, which is difficult to precisely measure the intrinsic relationship across different modalities in high-dimensional space. In this paper, we propose a novel multimodal feature selection method called Adaptive-Similarity-based Multi-modality Feature Selection (ASMFS) which performs adaptive similarity learning and feature selection simultaneously. Specifically, a similarity matrix is learned by jointly considering different modalities and at the same time, an efficient feature selection is conducted by imposing group sparsity-inducing $l_{2,1}$-norm constraint. Evaluated on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database with baseline MRI and FDG-PET imaging data collected from 51 AD, 43 MCI converters (MCI-C), 56 MCI non-converters (MCI-NC) and 52 normal controls (NC), we demonstrate the effectiveness and superiority of our proposed method against other state-of-the-art approaches for multi-modality classification of AD/MCI.

1. Introduction

Alzheimer's disease (AD) is a chronic neurodegenerative disease which is the main cause of dementia, leading to problems with language, disorientation, mood swings, bodily functions and, ultimately, death [1]. According to a recent report by Alzheimer's Association, about 5.3 million Americans have AD and 5.1 million of them are old people who are aged over 65 [2]. In 2050, the AD population will increase to beyond 100 million [3]. Although some therapies may temporarily improve symptoms of AD, there is no treatment that stops or reverses its progression so far. Hence, the early diagnosis of AD and its prodromal condition known as mild cognitive impairment (MCI) is highly essential for timely therapy. For the last decades, neuroimaging technique has proven to be a powerful tool to investigate the characteristics of neurodegenerative progression between AD and normal controls (NC), for instance, structural magnetic resonance imaging (MRI) for brain atrophy measurement [4], functional imaging (e.g., fluorodeoxyglucose positron emission tomography, FDG-PET) for hypometabolism quantification [6], and cerebrospinal fluid (CSF) for quantification of specific proteins [5,7].

In recent years, machine learning and pattern classification methods have been widely applied for the early diagnosis of AD based on single modality of biomarkers. For example, Lei et al. [8] proposed to build a framework based on longitudinal multiple time points data to predict clinical scores of AD. Liu et al. [9] developed an inherent structure-based multi-view learning method which utilizes the structure information of MRI data well. In addition to structural MRI, some researchers also used FDG-PET for AD or MCI classification [10]. However, these aforementioned methods tend to treat each modality of biomarkers as independent input without considering the intrinsic association among modali-

* Corresponding author.
E-mail address: wangyanscu@hotmail.com (Y. Wang).

https://doi.org/10.1016/j.patcog.2022.108566
0031-3203/

ties, which may result in suboptimal performance in predicting the progression of brain diseases.

In fact, different modalities of biomarkers can provide inherently complementary information for clinical diagnosis [12–15]. For example, structural MRI reveals patterns of gray matter atrophy, while FDG-PET measures the reduced glucose metabolism in the brain. It is reported that MRI and FDG-PET provide different sensitivity for memory prediction between disease and health [11]. As a result, many studies have used multimodal data to further improve the classification performance. For instance, Tong et al. [12] presented a multi-modality classification framework using nonlinear graph fusion to efficiently exploit the complementarity in the multi-modal data of PET and CSF. Hinrichs et al. [13] combined two modalities, i.e., MRI and PET, for classification of AD. Zhang et al. [14] further combined three modalities, i.e., MRI, FDG-PET and CSF, to classify AD/MCI from NC. Gray et al. [15] used MRI, FDG-PET, CSF and categorical genetic information for AD/MCI classification. These existing studies have suggested that different imaging modalities can provide different views of brain structure or function that might be overlooked by using a single modality. Thus, utilizing multiple modalities together to improve the accuracy in disease diagnosis becomes a sensible idea for researches.

Despite the promising performance of the above multi-modality classification methods, they all face the challenges of handling high dimensional features for the analysis. On the one hand, the curse of dimensionality, which tends to occur when there are insufficient training subjects versus large feature dimensions, limits the further performance improvement of existing methods. On the other hand, the high dimensional feature vectors usually contain some irrelevant and redundant features, which could possibly lead to the overfitting problem and hurt the generalization ability of the algorithm. Recently, in view of the capability of deep learning to automatically extract features, it has been widely applied in many medical feilds, including medical image reconstrcution [16–18], medical image segmentation [19,20], and radiation therapy [21,22]. As for AD classification, there still are some deep learning based methods works proposed to tackle the above challenges [23–26]. For instance, authors in [23–25] proposed convolutional neural networks (CNNs) for AD diagnosis on MRI and PET. Lin et al. [26] constructed 3D reversible generative adversarial networks (GANs) for AD classification. Although these deep-learning-based methods can achieve good performance, they always require large amounts of data to train the deep network, which is difficult to acquire in real practice. Moreover, in neuroimaging data analysis, features may correspond to brain regions with brain atrophy, pathological amyloid depositions or metabolic alterations. As black-box models, current deep-learning-based methods cannot explicitly interpret the clinical relevance of their intermediate features and the final predictions, which to some extent weakens the trust of physicians. Hence, explicit multi-modality feature selection methods still play an important role in clinical AD diagnosis.

However, there are two main challenges for feature selection in the multi-modality setting [27]. First, because the feature representations extracted from different modalities may have distinct distributions in a variety of feature spaces, it is challenging to integrate these discriminative features into a unified form of feature representation. Second, since various features from different modalities play distinctive roles in the classification task, how to evaluate each feature group and select the relevant features for the task remains a problem. Concentrating on the above challenges, several multi-modality feature selection methods have been developed in recent years [14,28–32]. A typical example is the multi-task feature selection (MTFS) proposed in [14], which selects common subset of relevant features from each modality. Based on MTFS, Liu et al. [28] proposed a multi-task feature selection method (IMTFS) to preserve the complementary inter-modality information. Different from MTFS, IMTFS imposes an inter-modality term, which can maintain the geometry structure of different modalities from the same subject. Also, a manifold regularized multi-task feature learning method (M2TFS) [29] was proposed to preserve the data distribution information by using a pre-defined similarity matrix to embed the manifold information into the feature selection procedure. Zhu et al. [31] proposed a multi-modality canonical feature selection (MCFS) method which considers the correlations between features of different modalities by projecting them into a canonical space determined by canonical correlation analysis. In these approaches, MTFS focuses on feature selection without considering the underlying data structure. IMTFS, M2TFS, and MCFS not only focus on feature selection, but also preserve the underlying data structure by modeling the relationship among subjects.

To better illustrate the differences among the aforementioned multi-modality feature selection methods, we plot the high-level overview of them in Fig. 1. As shown in Fig. 1(a) and (b), the traditional methods like MTFS and IMTFS obtain the selected features from the multimodal data directly, while the M2TFS employs a pre-defined similarity matrix to integrate the manifold information. Compared with the direct feature selection manner, the construction of similarity matrix can capture the relationship among all subjects within modalities, beneficial to reveal the underlying data structure to some extent. Nevertheless, in M2TFS, the neighbors and the similarity between the original high-dimensional data are usually obtained separately from each individual modality. This brings about two risks that may result in inaccurate similarity matrix: (1) considering the existence of noisy and redundant features, the relationship of subjects in high-dimensional space may not fully reveal the underlying data structure in the low-dimensional space after feature selection; (2) the similarity matrix is fixed before feature selection, meaning that the construction of the similarity matrix is performed separately from the feature selection task, which may lead to disturbed features in the classification stage. Given the performance of a mass of machine learning methods, such as Support Vector Machine (SVM), Locality Preserving Projection (LPP), K-means as well as other clustering algorithms, to a large extent, are highly dependent on the similarity between each pair of the subjects, the inaccurate similarity matrix could have negative impact on feature selection, thus potentially degrading the performance of AD classification.

To further boost the accuracy of the AD classification, this paper argues that the similarity should not be fixed but adaptive to change with the low-dimensional representation after feature selection. To this end, we propose a novel learning method termed as Adaptive-Similarity-based Multi-modality Feature Selection, or ASMFS for short, which is able to simultaneously capture the intrinsic similarity shared across different modality data, and select the most informative features. The high-level overview of our proposed method is shown in Fig. 1(c), where the similarity matrix and the feature selection procedure of the proposed method are updated alternatively, thus gradually improving the feature selection performance.

Additionally, it is commonly accepted that a large amount of real-world high-dimensional data actually lie on the low-dimensional manifolds embedded within a high-dimensional space [33]. Provided there is sufficient data (such that the manifold is well-sampled), we expect each data point and its neighbors to lie on or close to a locally linear patch of the manifold. We can characterize the local geometry of these patches by linear coefficients that are used to reconstruct each data point from its neighbors, which is called the manifold hypothesis [34]. Moreover, since neighborhood similarity is more reliable compared with the similarity retrieved from farther samples, preserving local neighborhood structure is of great help to construct an accurate similarity

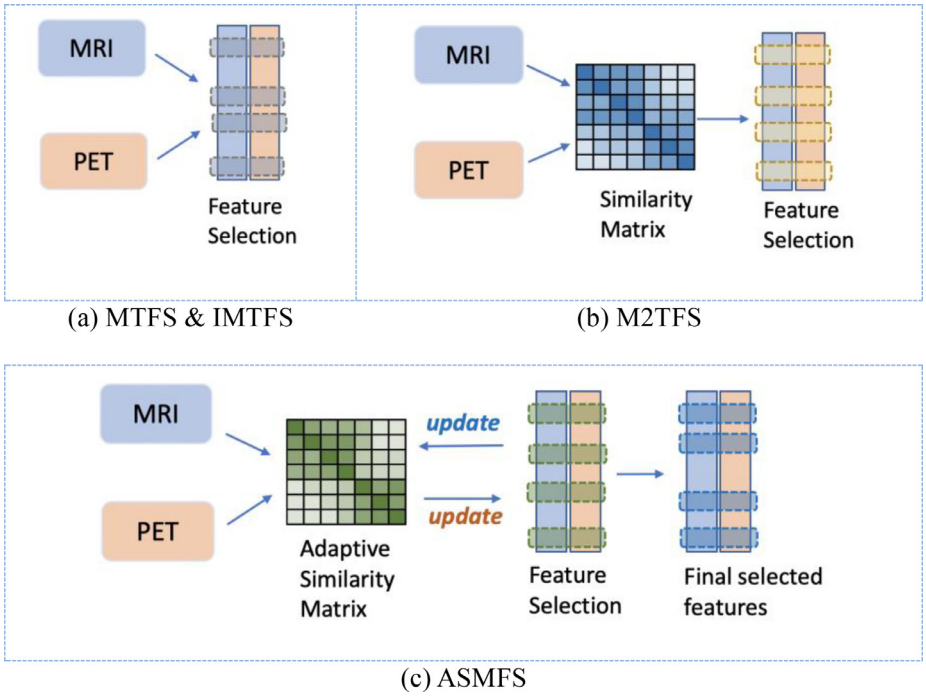


Fig. 1. Illustrations on three different multi-modality feature selection frameworks for AD classification.

matrix. Therefore, instead of updating the similarities between every data pair, we only consider the local neighborhood similarities of every subject.

Our contributions are as follows.

A novel multi-modality feature selection method named ASMFS is proposed to simultaneously perform similarity learning and feature selection. With the manifold hypothesis introduced, the similarity learning can derive a more accurate similarity matrix by preserving local structure information.

An adaptive learning strategy with regard to the similarity matrix is proposed to better depict the structure of data in low-dimensional space. In this manner, the similarity matrix is more informative, and thus helpful to select the discriminative features.

The similarity matrix is designed to be shared among different modality data collected from the same subject. By doing so, it can retrieve the collective information among multiple modalities as prior knowledge to further improve the performance of multi-modality feature selection.

Evaluated on the AD classification task with the MRI and FDG-PET data from the ADNI database, our proposed ASMFS is demonstrated to be effective and superior in identifying disease status and discovering the disease sensitive biomarkers compared with other feature selection methods.

The rest of this paper is organized as follows. Section 2 introduces our proposed multi-modality feature selection architecture and methodology. Experiments and experimental results are presented in Section 3. Finally, we discuss and conclude this paper in Section 4 and Section 5.

2. Method

2.1. Multi-modality feature selection with adaptive similarity learning

Fig. 2 gives the overview of the proposed method, which includes two major steps: (1) adaptive similarity learning and multi-modality feature selection, and (2) multimodal classification. In this section, we first introduce how to learn similarity measure from both single- and multi-modality data through adaptive similarity learning. Then, we show how to embed this similarity learning into our multi-modality feature selection framework. The selected features are eventually taken in a multi-kernel support vector machine (SVM) for disease classification [14].

2.1.1. Adaptive similarity learning

Let's consider the single modality scenario first. Suppose that in a d dimensional feature space, the data matrix of n subjects is denoted as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$. The subjects can be divided into c classes and the corresponding label vector is given as $\boldsymbol{y} = [y_1, y_2, ..., y_n]$. The similarity matrix $\boldsymbol{S}$ that indicates the similarity of data pairs can be constructed by two assumptions: 1) it is hoped that the similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be reflected by their Euclidean distance. If the distance $||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2$ between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is small, the similarity s_{ij} should be large, 2) if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to different classes, the similarity s_{ij} should be zero. In this section, we will first discuss two ideal cases according to the above two assumptions, and then propose our case with consideration of the two cases.

To begin with, we formulate the following objective to determine the similarities s_{ij} based on aforementioned assumptions:

$$\begin{aligned} &\min_{s_i} \sum_{j=1}^{n} x_i - x_j{}_2^2 s_{ij}, \\ &\text{s.t.} s_i^T 1 = 1,\ 0 \le s_{ij} \le 1 \\ &s_{ij} = 0,\ \text{if } y_i \ne y_j, \end{aligned} \tag{1}$$

where $\boldsymbol{s}_i \in \mathbb{R}^n$ is a vector of which the j-th entry is S_{ij} and 1 denotes a column vector with all the elements as one. However, by solving problem (1), it can be found that only one which is the closest neighbor to x_i has the similarity $s_{ij} = 1$, while the others are 0. In other words, it is a trivial solution.

Then, suppose the distance information is unavailable between subjects and the following problem is solved to estimate the simi-

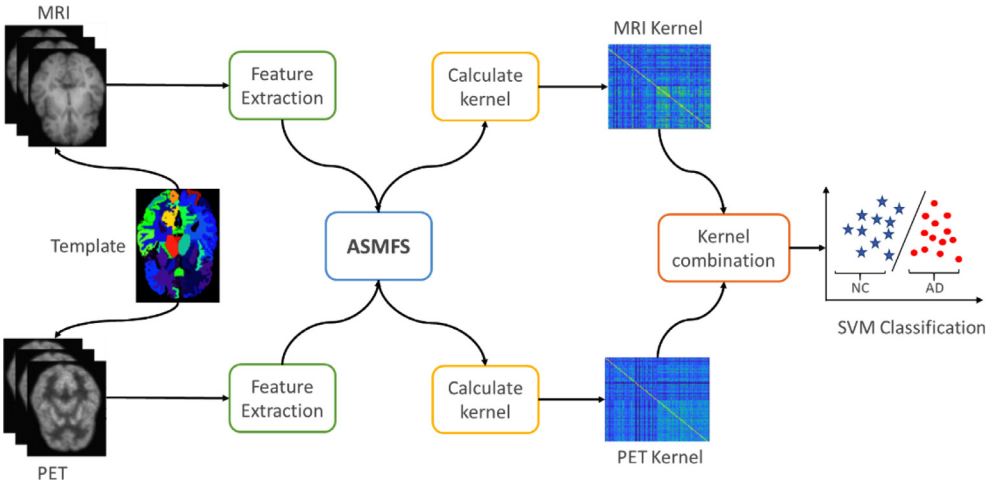


Fig. 2. Overview of multi-modality feature selection with adaptive similarity learning.

larities:

$$\begin{aligned} \min_{s_i} \quad & \sum_{j=1}^{n} s_{ij}^2, \\ \text{s.t} \quad & s_i^T 1 = 1,\ 0 \le s_{ij} \le 1, \\ & s_{ij} = 0,\ \text{if } y_i \ne y_j. \end{aligned} \tag{2}$$

The solution of $s_{ij} = \frac{1}{n}$ reveals that all the subjects will become the nearest neighbors of x_i with $\frac{1}{n}$ probability. The problem (2) can be actually regarded as the prior of the nearest neighbor probability when the pairwise subject distance is unknown. Considering problems (1) and (2) jointly, we solve the following objective to obtain the similarities s_{ij} :

$$\begin{aligned} \min_{s_i} \quad & \sum_{j=1}^{n} \left(\| x_i - x_j \|_2^2 \, s_{ij} + \alpha s_{ij}^2 \right), \\ \text{s.t} \quad & s_i^T 1 = 1,\ 0 \le s_{ij} \le 1, \\ & s_{ij} = 0,\ \text{if } \ y_i \ne y_j. \end{aligned} \tag{3}$$

The second term s_{ij}^2 can be regarded as a regularization term to avoid the trivial solution in problem (1) and α is the regularization parameter. It is notable that when $\alpha = 0$, Eq. (3) will degrade to Eq. (1) whose solution indicates that every subject only has one neighbor. On the other hand, when $\alpha = \infty$, the solution of Eq. (3) is the same as that of Eq. (2), suggesting that for each subject, all the other subjects become the nearest neighbors with the same similarity. Both of the extreme cases reveal that the regularization term of similarity learning is correlated with the number of subject's nearest neighbors. The problem (3) can be applied to calculate the similarities for each subject x_i. Consequently, in this paper we estimate the similarities for all subjects by solving the following problem:

$$\begin{aligned} \min_{\forall i, s_i} \quad & \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\| x_i - x_j \|_2^2 \, s_{ij} + \alpha s_{ij}^2 \right), \\ \text{s.t} \quad & s_i^T 1 = 1,\ 0 \le s_{ij} \le 1, \\ & s_{ij} = 0,\ \text{if } y_i \ne y_j. \end{aligned} \tag{4}$$

We can transform problem (4) to linearly constrained quadratic programming which can be solved by KKT conditions [35]. And the matrix $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ can be treated as a similarity matrix of n subjects.

Now, we extend the above adaptive similarity learning to multi-modality case. The multi-modality data are denoted as $X_1, X_2, ..., X_M$, where M is the number of modalities. The data matrix of the m-th modality is defined as $X_m = [x_1^{(m)}, x_2^{(m)}, ..., x_n^{(m)}]$. For all the multi-modality data, we solve the following problem to obtain the similarity matrix S:

$$\begin{aligned} \min_{S} \quad & \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\sum_{m=1}^{M} \| x_i^{(m)} - x_j^{(m)} \|_2^2 \, s_{ij} + \alpha s_{ij}^2 \right), \\ \text{s.t} \quad & s_i^T 1 = 1,\ 0 \le s_{ij} \le 1, \\ & s_{ij} = 0,\ \text{if } \ y_i \ne y_j. \end{aligned} \tag{5}$$

Please note that different from traditional multi-modality methods which calculate the similarity for each modality separately, the similarity matrix S obtained in (5) is shared by different modality data. Thus, the similarities of these data in diverse modalities would be identical.

Then, we embed the adaptive similarity learning into multi-modality feature selection in order to learn the optimal neighborhood similarity for feature selection, thereby improving the performance of multi-modality classification by utilizing more discriminative information.

2.1.2. Multi-modality feature selection with adaptive similarity learning

To integrate the similarity learning problem (5) with multi-modality feature selection, the objective function of our proposed method is defined as:

$$\begin{aligned} \min_{W,S} & \sum_{m}^{M} \sum_{i}^{N} \| y_i - w_m^T x_i^m \|_2^2 + \mu \, \| W \|_{2,1} \\ & + \lambda \sum_{i}^{N} \sum_{k \in \{k | y_i = y_k\}} \left(\sum_{m}^{M} \| w_m^T x_i^m - w_m^T x_k^m \|_2^2 \, s_{ik} + \alpha s_{ik}^2 \right) \\ \text{s.t.} & \sum_{k}^{n} s_{ik} = 1, \\ & 0 \le s_{ik} \le 1, \end{aligned} \tag{6}$$

where $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M] \in \mathbb{R}^{\mathrm{d} \times \mathrm{M}}$ is the coefficient matrix, $\boldsymbol{w}_m \in \mathbb{R}^{\mathrm{d}}$ is the coefficient of the m-th modality. The $\ell_{2,1}$ norm of $\boldsymbol{W}$ is defined as $\| \boldsymbol{W} \|_{2,1} = \Sigma_i^d \sqrt{\sum_j^M w_{ij}^2}$, which can result in sparse rows of $\boldsymbol{W}$ to achieve feature selection. ASMFS considers different modalities of subjects into similarity construction. λ, μ and α are parameters to balance the terms in (6).

What is noteworthy is that by using $\| W \|_{2,1}$, features are selected in the same brain regions from different modalities. These

selected features are essentially different from each other. For instance, MRI and PET images of hippocampus describe different characteristics of this brain region from the perspective of space and function respectively. As discussed in Section 2.1.1, in this paper, we aim to select features by utilizing inherently complementary information from different modalities. So by using $\|W\|_{2,1}$, we can not only investigate how AD affects the same brain regions from different perspectives, but also reduce the negative impact on feature selection caused by noise from data sampling and preprocessing.

From Eq. (6), we can capture the accurate inherent similarity shared across different modality data and then use this structure information to guide feature selection. Specifically, by performing adaptive similarity learning and feature selection in an alternate manner, similarity matrix in every learning step is able to depict the neighborhood information of data points in current low-dimensional space and helps to select the most discriminative features in the next step. Then, the selected feature in the next step will in turn be used to update similarity matrix. Procedure described above will be repeated until they converge.

2.2. Optimization algorithm

The objective function (6) is optimized in an alternate manner. Specifically, we fix S and optimize W and then fix W and optimize S.

(1) Fix S and optimize W.

Removing the irrelative part to W from (6), we get the following objective:

$$\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) = \sum_{m=1}^{M}\sum_{i}^{n} \| y_i - \boldsymbol{w}_m^T \boldsymbol{x}_i^m \|_2^2 + \mu \| \boldsymbol{W} \|_{2,1} + \lambda \sum_{i=1}^{n} \sum_{k\in\{k|y_i=y_k\}} \left(\sum_{m=1}^{M} \| w_m^T x_i^m - w_m^T x_k^m \|_2^2 s_{ik} \right). \tag{7}$$

Inspired by [38], we solve (7) using the weighted and iterative method. When the row elements in W are nonezero, that is $w_{i,:} \neq 0, i = 1, 2, ..., d$, we take the derivative of $\|W\|_{2,1}$ in terms of $w_{i,j}$:

$$\frac{\partial \|W\|_{2,1}}{\partial w_{ij}} = \frac{w_{ij}}{\sqrt{\sum_p^M w_{ip}^2}} = 2 d_{ii} w_{ij}, \tag{8}$$

where we set $d_{ii} = \frac{1}{2} \| w_{i,:} \|_2^{-1}$.

Then, following equation can be derived from (8):

$$\frac{\partial \|W\|_{2,1}}{\partial W} = 2DW, \tag{9}$$

where $\boldsymbol{D} \in \mathbb{R}^{d\times d}$ is a diagonal matrix, the i-th diagonal element is d_{ii}.

Taking the antiderivative of (8):

$$\begin{aligned} \|W\|_{2,1} &= \sum\nolimits_{\mathbf{i},\mathbf{j}} d_{ii} w_{ij}^2 + c \\ &= \sum_i d_{ii} \left(\sum_j w_{ij}^2 \right) + c \\ &= \mathrm{Tr}\left(W^T D W\right) + c \end{aligned} \tag{10}$$

When D is fixed, taking the derivative of W in (7) is equivalent to doing so in the following objective:

$$\min_W \mathcal{L}(W) = \sum_{m=1}^{M}\sum_{i}^{n} \| y_i - w_m^T x_i^m \|_2^2 + \mu \mathrm{Tr}\left(W^T D W\right) + \lambda \sum_{i}^{n} \sum_{k\in\{k|y_i=y_k\}} \left(\sum_{m=1}^{M} \| w_m^T x_i^m - w_m^T x_k^m \|_2^2 s_{ik} \right). \tag{11}$$

Please note that, the analytical form of W can be obtained via solving (11), and therefore (11) substitutes (7) in our learning framework.

(2) Fix W and optimize S.

Removing the irrelative part to S from (6), we can get the following objective:

$$\begin{aligned} &\min_s \sum_{i=1}^{n} \sum_{k\in\{k|y_i=y_k\}} \left(\sum_{m=1}^{M} w_m^T x_i^m - w_m^T x_{k2}^{m2} s_{ik} + \alpha s_{ik}^2 \right) \\ &\text{s.t.} \sum_k^n s_{ik} = 1, \\ &0 \le s_{ik} \le 1. \end{aligned} \tag{12}$$

In Section 2.1.1, we assumed that if x_i and x_j belong to different classes, the similarity s_{ij} should be zero. So when $k \in \{k|y_i \neq y_k\}$, then $s_{ik} = 0$, which means $w^T{}_m x_i{}^m - w^T{}_m x^{m2}{}_{k2} S_{ik} = 0$. Hence, we only need to consider the similarity between subjects from the same class, i.e., s_{ik} when $k \in \{k|y_i = y_k\}$.

Since the similarity learning of one subject is independent with respect to the learning of the others, we can safely decompose the similarity of individual subject according to the objective from (12):

$$\begin{aligned} &\min_{s_i} \sum_{k\in\{k|y_i=y_k\}} \left(\sum_{m=1}^{M} w_m^T x_i^m - w_m^T x_{k2}^{m2} s_{ik} + \alpha s_{ik}^2 \right), \\ &\text{s.t.} \sum_k^n s_{ik} = 1, \\ &0 \le s_{ik} \le 1. \end{aligned} \tag{13}$$

By defining $d_{ik} = \sum_{m=1}^{M} ||w_m^T x_i^m - w_m^T x_k^m||_2^2$, (13) can be simplified to the following form:

$$\begin{aligned} &\min_{s_i} \textstyle\sum_{k\in\{k|y_i=y_k\}} \left(d_{ik} s_{ik} + \alpha_i s_{ik}^2 \right) \\ &= \min_{s_i} \textstyle\sum_{k\in\{k|y_i=y_k\}} \left(\alpha_i \left(s_{ik} + \frac{1}{2\alpha_i} d_{ik} \right)^2 - \frac{d_{ik}^2}{4\alpha_i} \right) \\ &= \min_{s_i} \textstyle\sum_{k\in\{k|y_i=y_k\}} \left(\alpha_i \left(s_{ik} + \frac{1}{2\alpha_i} d_{ik} \right)^2 \right) \\ &= \min_{s_i} \alpha_i \textstyle\sum_{k\in\{k|y_i=y_k\}} \left(s_{ik} + \frac{1}{2\alpha_i} d_{ik} \right)^2 \\ &= \min_{s_i} \alpha_i s_i + \frac{1}{2\alpha_i} d_{i2}^2 \end{aligned} \tag{14}$$

Accordingly, the following objective is formulated:

$$\begin{aligned} &\min_{s_i} s_i + \frac{1}{2\alpha_i} d_{i2}^2, \\ &\text{s.t.} \ \textstyle\sum_{k=1}^{n} s_{ik} = 1, \\ &\quad 0 \le s_{ik} \le 1. \end{aligned} \tag{15}$$

We can solve (15) by KKT conditions.

The above objective is a convex function which can be solved utilizing Lagrange method:

$$L(s_i, \eta, \beta) = \frac{1}{2} s_i + \frac{1}{2\alpha_i} d_{i2}^2 - \eta\left(s_i^T 1 - 1\right) - \beta^T s_i, \tag{16}$$

where $\beta \ge 0$, $\eta \ge 0$ are Lagrange multipliers and 1 denotes a column vector with all the elements as one. Taking the derivative of (16) with respect to s_i and setting it equal to 0, we have:

$$\begin{aligned} \frac{\partial L(s_i, \eta, \beta)}{\partial s_i} &= \frac{\partial}{\partial s_i}\left(\frac{1}{2} s_i + \frac{1}{2\alpha_i} d_{i2}^2 - \eta\left(s_i^T 1 - 1\right) - \beta^T s_i \right) \\ &= s_i + \frac{1}{2\alpha_i} d_i - \eta 1 - \beta = 0. \end{aligned} \tag{17}$$

Algorithm 1
Multi-modality feature selection with adaptive similarity learning.

Input:Multi-modality sample matrix $\{X^1, X^2, \ldots, X^M\}$ and label matrix $y = [y_1, y_2, \ldots, y_n]$.
Initial:$\lambda > 0, \mu > 0, K > 0, D = I$, initialize S by solving problem (5).
Repeat
1. Update W using Eq. (11);
2. Update D using Eq. (8);
3. Update S using Eq. (21);
Until converges
Output:W

According to KKT conditions, we can get following equations:

$$\sum_{k=1}^{n} \beta_k s_{ik}^* = 0 \tag{18}$$

$$\beta_k \geq 0, \text{ for } k = 1, \ldots, n, \tag{19}$$

where s_{ik}^* is the optimal solution, (18) and (19) are dual feasibility condition and complementary slackness condition in KKT conditions respectively. Then, (20) can be derived from (17), (18) and (19):

$$\begin{cases} s_{ik}^* = 0, & \beta_i > 0 \\ s_{ik}^* = -\frac{d_{ik}}{2\alpha_i} + \eta, & \beta_i = 0. \end{cases} \tag{20}$$

Hence, the optimal solution can be figured out from (20):

$$s_{ik}^* = \left(-\frac{d_{ik}}{2\alpha_i} + \eta\right)_+ = \max\left(-\frac{d_{ik}}{2\alpha_i} + \eta, 0\right). \tag{21}$$

Practically, as discussed in Section 1, keeping the local manifold structure of data is proved well effective [36,37] in feature selection. One can improve the classification performance with attention only to the local structure of data. Therefore, we expect to learn a sparse s_i. That is, only the nearest K neighbors of x_i have the opportunity to connect with x_i. Moreover, sparse similarity matrix learning is of great help to reduce the computational burden for the later processing.

Let us suppose that $d_{i1}, d_{i2}, \ldots, d_{in}$ are sorted from the lowest to the highest. Provided that the optimal solution s_i^* has only K non-zero elements, using (21), we know $s_{iK}^* > 0$ and $s_{i,K+1}^* \leq 0$. Hence, the following inequalities hold:

$$\begin{cases} s_{ik}^* = -\frac{d_{ik}}{2\alpha_i} + \eta > 0, & k \leq K \\ s_{ik}^* = -\frac{d_{ik}}{2\alpha_i} + \eta \leq 0, & k > K. \end{cases} \tag{22}$$

Substituting the constraint $\sum_{k=1}^{K} s_{ik}^* = 1$ into the (22), we have:

$$\begin{aligned} &\sum_{k=1}^{K} \left(-\frac{d_{ik}}{2\alpha_i} + \eta\right) = 1 \\ &\Rightarrow \eta = \frac{1}{K} + \frac{1}{2K\alpha_i} \sum_{k=1}^{K} d_{ik}. \end{aligned} \tag{23}$$

Plugging η into (22) leads to the constraint of α_i:

$$\frac{K}{2} d_{iK} - \frac{1}{2} \sum_{k=1}^{K} d_{ik} < \alpha_i \leq \frac{K}{2} d_{i,K+1} - \frac{1}{2} \sum_{k=1}^{K} d_{ik}. \tag{24}$$

Finally, we have the optimal a^*_1 derived from (21) and (24):

$$\alpha_i^* = \frac{K}{2} d_{i,K+1} - \frac{1}{2} \sum_{k=1}^{K} d_{ik}. \tag{25}$$

The procedure is summarized in Algorithm 1.

Table 1
Clinical and demographic information of the study population.

	AD	MCI-C	MCI-NC	NC
Subjects number	51	43	56	52
Age	75.2±7.4	75.8±6.8	74.7±7.7	75.3±5.2
Education	14.7±3.6	16.1±2.6	16.1±3.0	15.8±3.2
MMSE	23.8±2.0	26.6±1.7	27.5±1.5	29.0±1.2
CDR	0.7±0.3	0.5±0.0	0.5±0.0	0.0±0.0

2.3. Multi-kernel support vector machine

Multi-kernel support vector machine (MKSVM) [39] is adopted for classification after feature selection processing. First, we generate a kernel matrix $k^m(x_i^m, x_j^m) = \phi^m((x_i^m)^T(x_j^m))$ for each modality data after feature selection. And the M kernel matrices are linearly combined $k(x_i, x_j) = \sum_{m=1}^{M} \beta_m k^m(x_i^m, x_j^m)$ where $\sum_{m=1}^{M} \beta_m = 1, \beta_m \geq 0$. Then, MKSVM is trained by the selected kernel combination weights and linear kernels. It is notable that in our experiments, the optimal β_m is determined via a coarse-grid search through cross-validation on the training set. Finally, when an unseen subject comes, the trained MKSVM model is able to predict the category of the new subject by the following decision function:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} y_i \alpha_i \sum_{m=1}^{M} \beta_m k^m\left(x_i^m, x^m\right) + b\right). \tag{26}$$

In particular, we adopt a linear SVM as the classifier because it intrinsically uses a feature weighting mechanism, i.e., the absolute values of components in the normal vector of SVM's hyperplane can be regarded as weights on features [40]. In this way, we can rank the features according to their averaged SVM weights. Then, the most discriminative brain regions can be determined by these ranked features. Detailed discussion about feature selection result is presented in Section 3.3.

3. Experiments

3.1. Dataset and settings

Dataset: The data involved in this paper are obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.usc.edu). The ADNI was launched by a wide range of academic institutions and private corporations and the subjects were collected from approximately 200 cognitively normal older individuals with 3 years of follow-up, 400 MCI patients with 3 years of follow-up, and 200 early AD patients with 2 years of follow-up across the United States and Canada.

In this paper, subjects with all corresponding MRI and PET baseline data are included. This yields a total of 202 subjects. In particular, it includes 51 AD patients, 99 MCI patients and 52 NC. The MCI patients were divided into 43 MCI converters (MCI-C) who have progressed to AD with 18 months and 56 MCI non-converters (MCI-NC) whose diagnoses have still remained stable within 18 months. A detailed description on acquiring MRI, PET data from ADNI can be found at [14]. Table 1 lists the clinical and demographic information of the study population.

In this study, image pre-processing is performed for all MRI and PET images following the same procedures as in [14,41,42]. Specifically, N3 algorithm [43] is employed to correct the intensity inhomogeneity after anterior commissure-posterior commissure correlation performed. For MRI data, the gray matter (GM) is segmented by FAST [44] and then the GM tissue volume of each region obtained according to a 93 manual labels template is chosen as a

Table 2
Comparison of different methods for AD vs. NC classification.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 Score	AUC
SVM	88.24±0.0972	91.07±0.1155	85.57±0.1591	88.61±0.0925	0.9471±0.0007
lassoSVM	90.90±0.0873	90.60±0.1240	91.23±0.1233	90.71±0.0900	0.9460±0.0007
MKSVM	91.87±0.0875	92.30±0.1249	91.63±0.1160	91.68±0.0927	0.9526±0.0007
lassoMKSVM	92.33±0.0739	93.47±0.1030	91.30±0.1261	92.41±0.0726	0.9534±0.0007
MTFS	92.52±0.0816	93.77±0.1115	91.37±0.1213	92.50±0.0846	0.9541±0.0007
M2TFS	95.00±0.0707	94.67±0.1009	95.40±0.0826	94.85±0.0740	0.9636±0.0006
ASMFS	96.76±0.0545	96.10±0.0836	97.47±0.0660	96.63±0.0573	0.9703±0.0006

feature. After alignment to the respective MRI image, the average intensity of each ROI in the PET image is calculated as a feature. Therefore, there are totally 93 features for each MRI image and 93 features for each PET image.

Validation: Z-score normalization $f_i' = (f_i - \bar{f}_i)/\sigma_i$ is performed on every feature f_i from MRI images and PET images separately, where $\bar{f}_i$ and σ_i respectively represent the mean and the standard deviation of the i-th feature from the training set. It is noted that we performed Z-score normalization on test set with $\bar{f}_i$ and σ_i calculated from the training set.

Performance measurements including accuracy (ACC), sensitivity (SEN), specificity (SPE), area under receiver operating characteristic (ROC) curve (AUC) and F1 Score are utilized in the experiments to quantify the classification performance of different methods. The 10-fold cross-validation strategy is adopted due to the limited subjects. Specifically, the whole set of subject samples are equally partitioned into 10 subsets, from which 9 subsets were randomly selected for training and the remaining subset for testing. The above procedure was repeated 10 times to avoid any bias caused by the partition.

Hyper-parameters: In our method, there are three hyper-parameters, i.e., the sparsity regularization coefficient μ, the adaptive similarity learning regularization coefficient λ, and the number of neighbors K. The above parameters are determined based on the training samples by 10-fold cross-validation. To be specific, in each 10-fold cross-validation used to compute the classification performance, we perform another 10-fold cross-validation on the training samples to determine the optimal values for these parameters. λ, μ and K are searched in the range $\{0.1, 5, 20, 60, 100\}$, $\{0, 5, 10, 15, 20\}$ and $\{1, 3, 5, 7, 9\}$, respectively. Moreover, in multi-kernel SVM with a linear kernel, C is set as 1 and the kernel combination coefficients β_{MRI}, β_{PET} are chosen from 0.1 to 1.0 with step 0.1 and constrained with $\beta_{MRI}+\beta_{PET}=1$and β_{MRI}, $\beta_{PET}0$.

3.2. Classification results

In order to assess the classification performance, the proposed method is compared with six existing multimodal classification methods including (1) standard SVM with linear kernel (denoted as SVM) [40], (2) standard SVM with linear kernel and LASSO feature selection (denoted as lassoSVM) [45], (3) multi-kernel SVM (denoted as MKSVM) [39], (4) multi-kernel SVM with LASSO feature selection performed independently on single modality (denoted as lassoMKSVM) [55], (5) multi-kernel SVM using multi-modal feature selection method (denoted as MTFS) [46], and (6) multi-kernel SVM with manifold regularized multitask feature learning (denoted as M2TFS) [29]. Please note that all methods above use the same SVM classifier.

3.2.1. AD vs. NC classification

Table 2 lists the results on the AD vs. NC classification task produced by our proposed method (ASMFS) and other six SVM-based methods. The standard deviations are given and the best results are denoted in bold in Table 2. As observed, our proposed method (ASMFS) consistently achieves the best performance compared with other methods. Specifically, ASMFS achieves the accuracy of 96.76%, the sensitivity of 96.1%, the specificity of 97.47%, the AUC of 0.9703 and the F1 Score of 96.63. The lowest accuracy of 88.24% is achieved by SVM since that SVM does not perform feature selection but directly uses the raw feature vectors for classification. lassoSVM achieves better performance than SVM because it adopts LASSO as feature selection which can remove redundant features and noise. Comparing the results between MKSVM and SVM, it can be found that utilizing complementary information from different modalities greatly promotes the accuracy (3%), indicating the necessity of jointly considering multiple modalities for AD classification. By incorporating LASSO to MKSVM for feature selection, the accuracy of lassoMKSVM is further boosted to 92.33%. Among the three multi-modality feature selection methods, i.e., MTFS, M2TFS and ASMFS, we can find that over 95% accuracy is achieved by M2TFS and ASMFS, indicating that maintaining inter-modality information is effective for feature selection. Furthermore, compared with M2TFS which adopts fixed similarity, the better performance achieved by ASMFS indicates that adaptive similarity is of great help to depict more accurate data distribution after feature selection. Moreover, our method achieves both the highest sensitivity and specificity, indicating that our method rarely overlooks an AD patient or misclassifies a normal individual as the diseased. Besides, ASMFS keeps the lowest standard deviation, indicating the proposed method is more stable.

Fig. 3 plots the corresponding ROC curves of all methods for AD vs. NC classification, from which we can see the proposed method obtains the best performance with the largest AUC with high true positive rate (TPR) at low false positive rate (FPR) when compared with other methods.

3.2.2. MCI vs. NC classification

Table 3 shows the performance of our method compared with other six SVM-based methods on the MCI vs. NC classification task. As observed, the proposed method gets the best performance in accuracy, specificity, F1 Score and AUC, while M2TFS achieves the best sensitivity of 86.73%. Nevertheless, ASMFS is only 0.75% lower than M2TFS. Note that the accuracy of M2TFS and ASMFS is much higher than other methods. This is probably because the selected structural features play a pivotal role in the improvement of classification performance.

In addition, the results listed in Table 3 are generally lower than those in AD vs. NC classification. It is because the changes occurring in the brain of MCI patients are less than those of AD patients. For instance, MCI patients have much less contraction in the hippocampus than AD patients. Hence, MCI classification is a more challenging task. Fig. 4 plots the corresponding ROC curves which reflect the classifier performance of the above methods. Similarly, our method still surrounds larger area than other methods.

3.2.3. MCI-C vs. MCI-NC classification

The classification results for MCI-C vs. MCI-NC are shown in Table 4. As observed, our proposed method achieves the best clas-

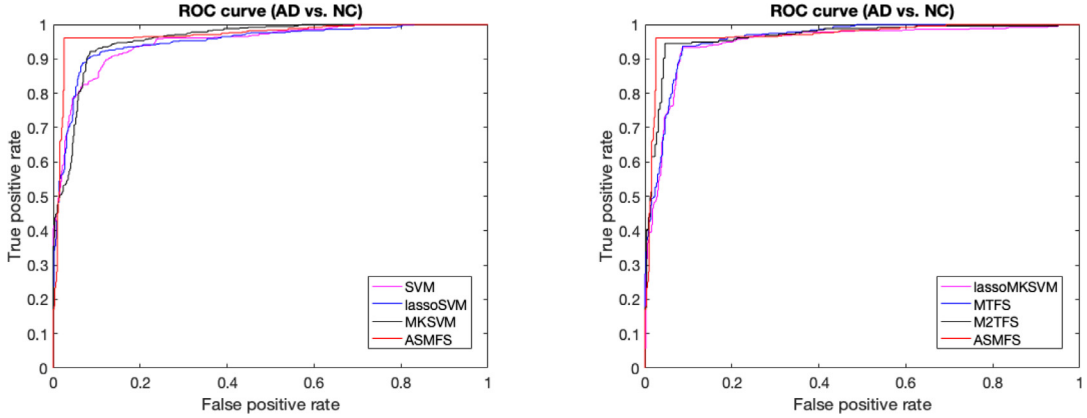


Fig. 3. ROC curves of seven multi-modality based methods for classification of AD vs. NC.

Table 3
Comparison of different methods for MCI vs. NC classification.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 Score	AUC
SVM	70.62±0.1035	84.03±0.1176	45.20±0.2111	81.04±0.0599	0.7463±0.0013
lassoSVM	73.40±0.1167	81.62±0.1358	58.00±0.2141	79.78±0.0960	0.7852±0.0013
MKSVM	73.17±0.0983	80.69±0.1141	59.00±0.2189	79.62±0.0762	0.7276±0.0014
lassoMKSVM	74.19±0.0894	86.57±0.1098	50.70±0.2703	81.44±0.0647	0.7539±0.0012
MTFS	74.86±0.0911	82.19±0.1135	61.07±0.2066	80.91±0.0716	0.7296±0.0014
M2TFS	78.97±0.0766	86.73±0.1070	64.53±0.2515	84.35±0.0561	0.7526±0.0014
ASMFS	80.73±0.0950	85.98±0.1081	70.90±0.2135	85.30±0.0738	0.7875±0.0014

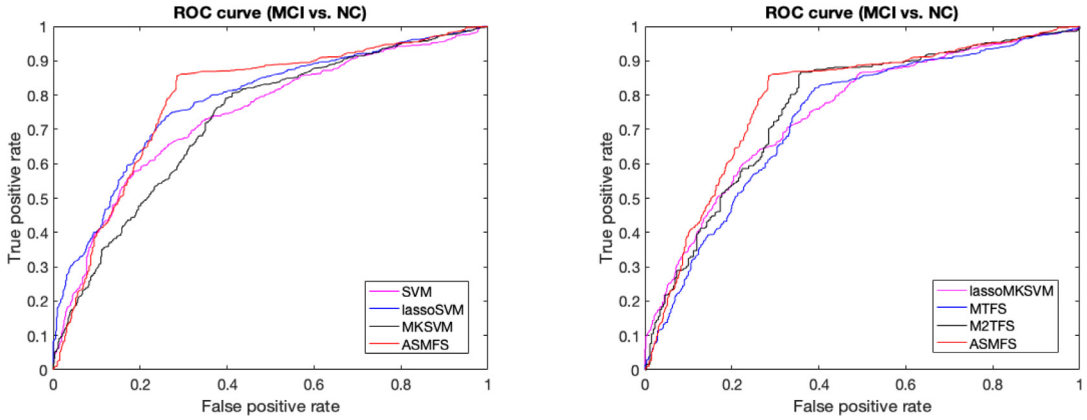


Fig. 4. ROC curves of seven multi-modality based methods for classification of MCI vs. NC.

Table 4
Comparison of different methods for MCI-C vs. MCI-NC classification.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 Score	AUC
SVM	56.45±0.1338	31.55±0.2126	75.90±0.2024	36.21±0.2195	0.6341±0.0017
lassoSVM	58.76±0.1394	48.75±0.2422	66.43±0.2127	48.69±0.1972	0.5830±0.0017
MKSVM	58.80±0.1206	54.45±0.2293	62.43±0.2202	51.74±0.1625	0.5753±0.0017
lassoMKSVM	61.73±0.1369	51.10±0.2469	70.23±0.2109	51.67±0.2032	0.6086±0.0018
MTFS	63.52±0.1220	59.65±0.2514	66.70±0.2108	56.63±0.1762	0.5894±0.0017
M2TFS	67.53±0.1059	54.50±0.2629	77.47±0.1873	55.84±0.2182	0.6647±0.0017
ASMFS	69.41±0.1194	65.30±0.2151	72.83±0.1811	63.98±0.1485	0.6534±0.0017

sification accuracy of 69.41%, sensitivity of 65.3% and F1 Score of 63.98, while the best specificity and AUC are obtained by M2TFS, 4.64% and 0.0113 higher than our method, respectively. Despite so, our method significantly surpasses M2TFS by 10.8% sensitivity and 8.14% F1 score, proving that our method is still superior to other methods.

Moreover, we have a similar observation to the MCI vs. NC classification that the accuracy of M2TFS and ASMFS is much higher than that of other methods, which well demonstrates the importance of preserving structural information. Meanwhile, compared with the results on the classifications of AD vs. NC and MCI vs. NC, the results on the MCI-C vs. MCI-NC classification are generally lower. It can be explained that the differences between MCI-C and MCI-NC are small and the early symptoms of MCI are also similar for MCI-C and MCI-NC. The ROC curves in Fig. 5 also indicate the superiority of our method compared with others.

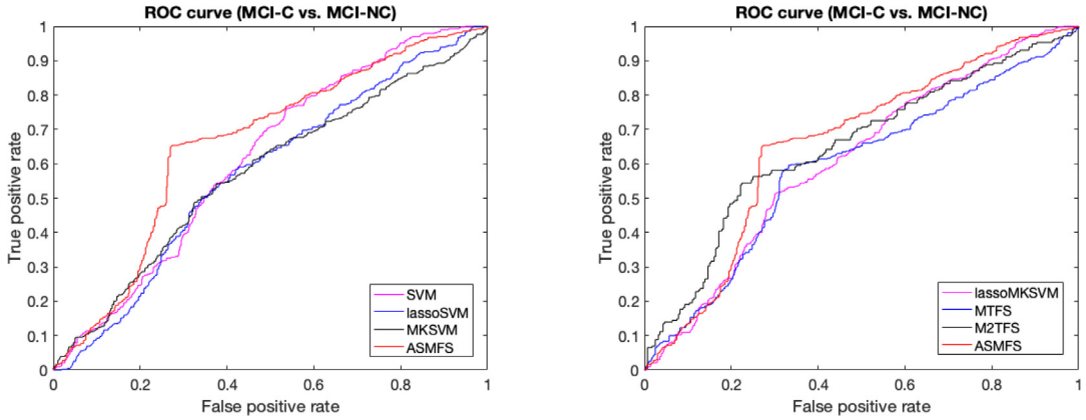


Fig. 5. ROC curves of seven multi-modality based methods for classification of MCI-C vs. MCI-NC.

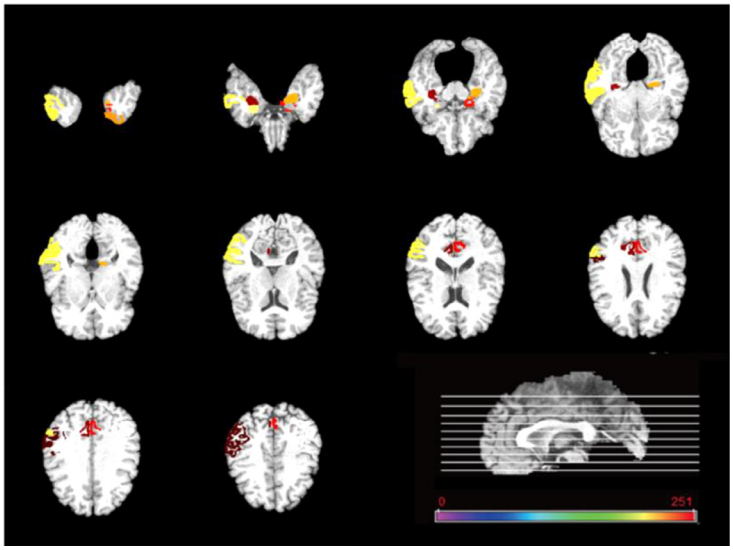


Fig. 6. Top 10 ROIs selected by the proposed method for AD vs. NC.

Table 5
Top 10 ROIs selected by the proposed method and M2TFS for AD vs. NC classification. Different features selected by ASMFS and M2TFS are highlighted in bold.

	Selected ROIs of ASMFS	Selected ROIs of M2TFS
1	hippocampal formation left	angular gyrus right
2	precuneus left	cingulate region left
3	middle temporal gyrus right	precuneus right
4	inferior temporal gyrus right	middle frontal gyrus left
5	angular gyrus right	precuneus left
6	angular gyrus left	precentral gyrus left
7	precuneus right	temporal pole left
8	uncus left	hippocampal formation left
9	amygdala right	uncus left
10	lateral occipitotemporal gyrus left	middle temporal gyrus right

Table 6
Top 10 ROIs selected by the proposed method and M2TFS for MCI vs. NC classification. Different features selected by ASMFS and M2TFS are highlighted in bold.

	Selected ROIs by ASMFS	Selected ROIs by M2TFS
1	angular gyrus left	cuneus left
2	hippocampal formation left	precuneus left
3	entorhinal cortex left	temporal pole left
4	cuneus left	entorhinal cortex left
5	amygdala right	hippocampal formation left
6	precuneus left	angular gyrus left
7	temporal pole left	hippocampal formation right
8	occipital pole left	occipital pole right
9	hippocampal formation right	occipital pole left
10	parahippocampal gyrus left	amygdala right

3.3. Feature selection results

The discriminability of brain regions is ranked by the regression coefficient W. Figs. 6–8 show the top 10 brain regions selected by ASMFS in the classification of AD vs. NC, MCI vs. NC and MCI-C vs. MCI-NC, respectively. To better illustrate the superiority of ASMFS, we also compare the top 10 selected brain regions of ASMFS with those of M2TFS as reported in [29]. The results are listed in Tables 5–7 in the AD, MCI and MCI-C classifications, respectively.

For the AD vs. NC classification, brain regions such as hippocampus, precuneus, uncus and temporal gyrus are found sensitive to AD by ASMFS. Simultaneously, the brain regions, for instance, hippocampus and amygdala, are also selected in the MCI vs. NC classification task. There have been several studies that have shown the association between these brain regions and AD. For example, the studies in [47,48] suggest that the hippocampus is responsible for short-term memory, and in the early stage of Alzheimer's disease also known as MCI, hippocampus begins to be destroyed, which directly results in the decline of short-

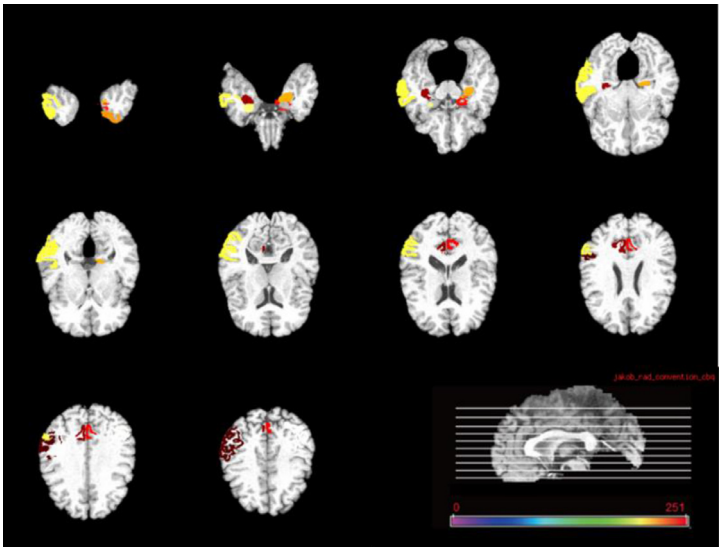


Fig. 7. Top 10 ROIs selected by the proposed method for MCI vs. NC.

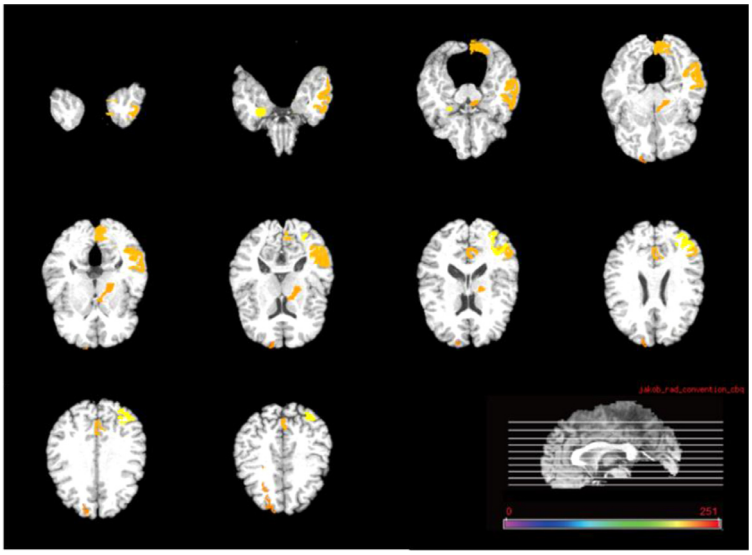


Fig. 8. Top 10 ROIs selected by the proposed method for MCI-C vs. MCI-NC.

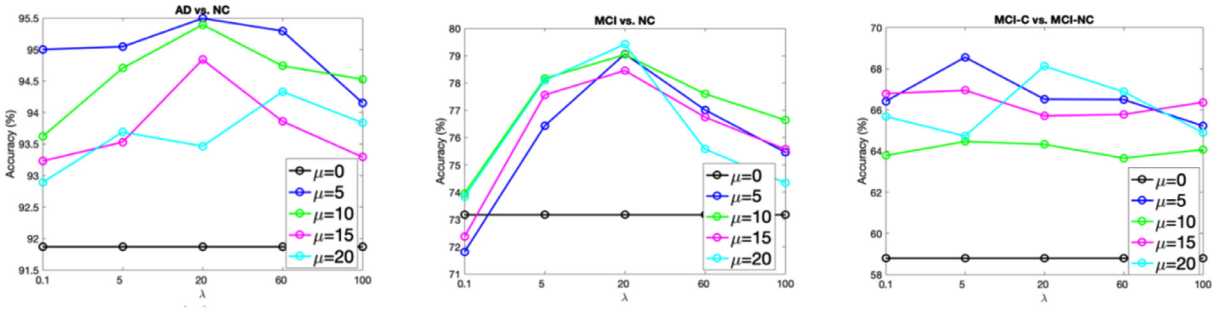
Table 7
Top 10 ROIs selected by the proposed method and M2TFS for MCI-C vs. MCI-NC classification. Different features selected by ASMFS and M2TFS are highlighted in bold.

	Selected ROIs of ASMFS	Selected ROIs of M2TFS
1	precuneus left	superior frontal gyrus right
2	perirhinal cortex left	caudate nucleus right
3	anterior limb of internal capsule left	postcentral gyrus left
4	middle temporal gyrus left	inferior frontal gyrus right
5	superior frontal gyrus right	precuneus left
6	amygdala right	frontal lobe WM right
7	lingual gyrus left	middle frontal gyrus left
8	middle occipital gyrus left	cingulate region left
9	hippocampal formation right	middle occipital gyrus left
10	fornix left	middle frontal gyrus right

term memory and disorientation. The study in [57] implies that amygdala is responsible for managing basic emotions such as fear and anger. The damage for amygdala caused by MCI/AD can lead to paranoia and anxiety. Accordingly, the selected regions by the proposed method are indeed effective for AD diagnosis. Compared with ASMFS, however, M2TFS attaches less importance to hippocampus and temporal gyrus for AD vs. NC and MCI vs. NC classifications. And amygdala is ignored by M2TFS in the AD vs. NC classification task. What's more, because of the subtle differences between MCI-C and MCI-NC, it is reasonable to observe that most of the features selected by M2TFS in Table 7 are not as discriminative as the features obtained in AD/MCI classifications. Nevertheless, the top 10 features from ASMFS are still close to the features selected in the AD and MCI classification task, including hippocampus, precuneus, temporal gyrus and amygdala. This observation indicates that ASMFS can still select distinct features in the MCI-C vs. MCI-NC task, demonstrating the effectiveness and robustness of our method.

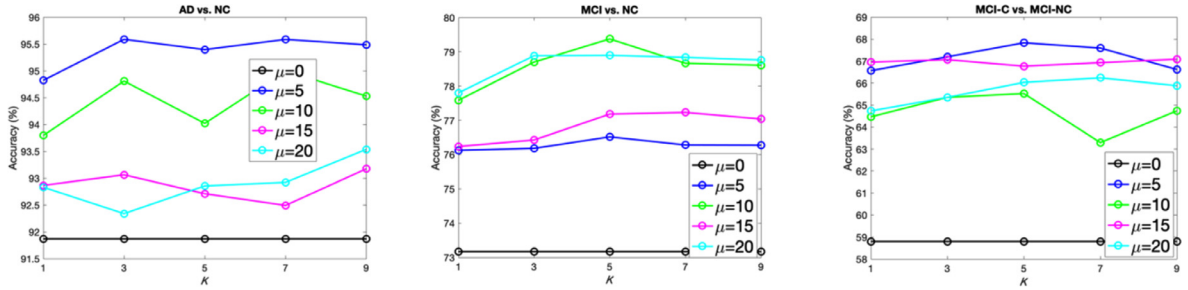
3.4. Effect of hyper-parameters

Regularization parameters: In ASMFS, there are three hyper-parameters, i.e., λ, μ and K. Specifically, the adaptive similarity learning regularization coefficient λ and the group sparsity regularization coefficient μ control the relative contribution of those regularization terms. K is the number of neighbors in adaptive simi-



(a) AD vs. NC. (b) MCI vs. NC. (c) MCI-C vs. MCI-NC.

Fig. 9. The classification accuracy with respect to regularization parameters λ and μ on (a) AD vs. NC classification task, (b) MCI vs. NC classification task, and (c) MCI-C vs. MCI-NC classification task.



(a) AD vs. NC. (b) MCI vs. NC. (c) MCI-C vs. MCI-NC.

Fig. 10. The classification accuracy with respect to regularization parameters K and μ on (a) AD vs. NC classification task, (b) MCI vs. NC classification task, and (c) MCI-C vs. MCI-NC classification task.

larity learning. As aforementioned, the above parameters are determined by another 10-fold cross-validation. λ, μ and K are searched in the range $\{0.1, 5, 20, 60, 100\}$, $\{0, 5, 10, 15, 20\}$ and $\{1, 3, 5, 7, 9\}$, respectively. It is worth mentioning that when $\mu = 0$, the group sparsity will cease to work. Although inherent similarity information can still be captured after optimizing the objective function, the most discriminative features cannot be selected without group sparsity. Then, all the features are retained for the subsequent classification. Hence, the subsequent classification method in this scenario degenerates to a multi-modality classification method without feature selection which is the same as the method proposed in [14].

Fig. 9 shows the classification results with regard to different values of λ and μ when K is fixed to 5. The X-axis indicates λ, Y-axis indicates classification accuracy and the curves with different colours represent different values of μ ranging among $\{0, 5, 10, 15, 20\}$, respectively. As observed, with the increase of λ from 0.1 to 20, the curves corresponding to different values of μ show a rising trend, whereas the downtrend of accuracy is observed when λ is larger than 20. Through analysis, we believe that λ conducts a less effective guide when it is relatively small because λ almost performs no constraint on the item of adaptive similarity learning.

Besides, as can be seen, when λ is fixed, μ has a greater impact on classification accuracy than λ, which is because μ affects the sparsity of W and determines the number of discriminative features. Also, as we can see from Fig. 9(a), when $\mu = 0$ which suggests that no feature is selected, the corresponding accuracy curve lies below other curves. The similar phenomenon can be seen in Fig. 9(b), (c) as well. Such result demonstrates the effectiveness of feature selection.

Fig. 10 shows the classification results with different values of K and μ when λ is fixed to 20. As observed, the classification performance with feature selection ($\mu = 5, 10, 15, 20$) is better than that without feature selection ($\mu = 0$). Most of the curves reach their peak when $K = 5$ but go down when K is beyond 5. Such result suggests that maintaining the local manifold structure of data helps to select discriminative features. What's more, Fig. 10 shows the similar profile of curves with different values of μ when compared with Fig. 9. According to the above experimental results, we determine the hyper-parameters μ, λ, and K as 5, 20, 5, respectively.

3.5. Adaptive similarity learning

To better demonstrate the reasonability of the similarity matrix shared by all modalities in adaptive similarity learning, we further conduct a comparison experiment. Specifically, in contrast to the proposed method with shared similarity matrix (denoted as "ASMFS-shared"), we additionally devise a method which learns similarity matrix in different modalities separately (denoted as "ASMFS-separated"). Then we compare these two methods on the AD vs. NC, MCI vs. NC and MCI-C vs. MCI-NC classification tasks, respectively. The experimental settings are the same to those in Section 3.1.

As shown in Table 8, compared with ASMFS-shared, the accuracies of ASMFS-separated degrade by 2.83%, 2.45%, and 1.79% for AD, MCI and MCI-C classification, respectively. Such results are in line with our assumption that the different modality data collected from the same subject share the similar intrinsic characteristics which essentially show the impact of AD to brain. By learning the relationship between samples across diverse modalities, we can bridge these modalities to capture the inherently complemen-

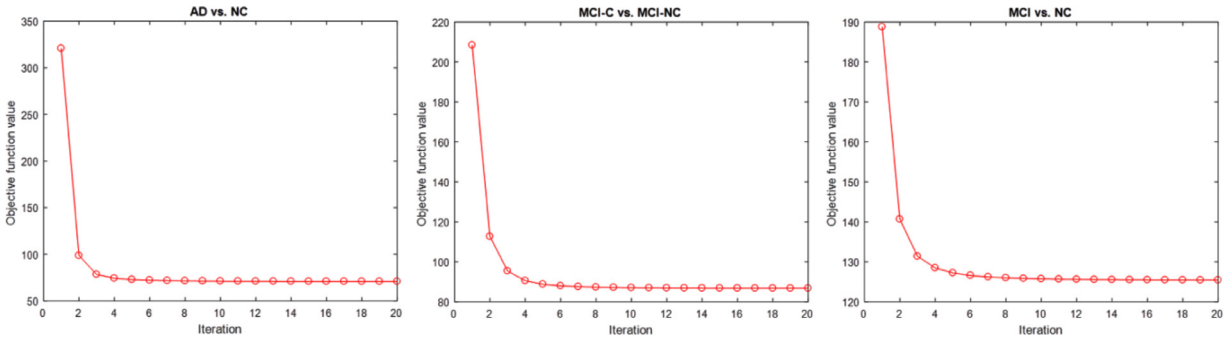


Fig. 11. Algorithm convergence of ATMFS.

Table 8
Classification accuracy of different similarity learning mechanisms for AD classification.

Method	AD vs. NC	MCI vs. NC	MCI-C vs. MCI-NC
ASMFS-separated	93.93%	78.28%	67.62%
ASMFS-shared	96.76%	80.73%	69.41%

tary information from them, thus boosting the performance of AD classification. Particularly, we can observe a slighter degradation of ASMFS-separated on the MCI-C vs. MCI-NC classification task, which can be explained that the differences between MCI-C and MCI-NC are subtle, making it a challenging task to distinguish MCI-C from MCI-NC.

3.6. Algorithm convergence and robustness

Algorithm convergence: The optimization is performed in an alternative manner. The convergence of this optimization algorithm is guaranteed by the proof provided by [38,56]. We also conduct an experiment to investigate the convergency of the proposed method. As we can observe from Fig. 11, the proposed optimization algorithm has a good convergence, since the algorithm has basically converged after about 10 iterations on all the three tasks. Besides, the average training time for the proposed model to converge is less than 5 s, which is obtained by 10 runs.

Algorithm robustness: In order to evaluate the robustness of the proposed method, we add zero-mean Gaussian white noise with variance of 0.01, Poisson noise and Salt and Pepper noise with noise density of 0.05 to the input data, respectively, and compare the accuracy of our method with that of MTFS and M2TFS on the AD vs. NC classification task. Moreover, we also measure the degradation ratio to investigate how much the noise influences the classification accuracy.

Table 9 shows the accuracy and the degradation ratio (denoted as Deg. Ratio) of classification for AD vs. NC under three different kinds of noise. As can be seen, our proposed ASMFS successfully avoids being misguided by all kinds of noise, achieving the highest accuracies of 95.56%, 96.31% and 94.76% with the lowest degradation ratios of 1.24%, 0.46% and 2.06% under Gaussian white noise, Poisson noise and Salt and Pepper noise, respectively, demonstrating the robustness of our method. On the contrary, MTFS and M2TFS are perturbed significantly by noise with accuracies below 90%. Accordingly, it is easy to infer that our ASMFS, which adopts the adaptive similarity learning, can well handle the noise, proving the robustness of our method.

3.7. Comparison with the state-of-the-art methods

We further compare the results achieved by our ASMFS with (1) the results achieved by several traditional machine learning based works reported in the literature studying multi-modality feature selection, including the works in [15,28,29,31,53–55], and (2) the results achieved by several recent state-of-the-art deep learning based methods for AD classification, including the works in [23–25] which used CNN for AD diagnosis and [26] which applied generative model GAN to classification task. The details of each method and the corresponding results are listed in Table 10. As observed, our proposed method achieves the best performance with 96.76% accuracy on AD vs. NC classification. Compared with the method in [53] which gains the second-best performance, our method still boosts the accuracy by 0.81%. In addition, compared with [28,29,31,53] using the same amount of data (i.e., 51 AD + 99 MCI + 52 NC), our method still yields the best performance with 80.73% accuracy on MCI vs. NC classification and the comparable performance on MCI-C vs. MCI-NC classification. For those deep-learning-based methods, although they generally produce higher accuracy than our method on MCI-C vs. MCI-NC, such results can be attributed to the more MCI and NC data involved in their training stage. In summary, we can draw a conclusion that our proposed ASMFS is more effective and efficient under the same condition.

4. Discussion

Multi-modality learning, a recently developed technique in machine learning field which can jointly learn multiple modalities via a shared representation, has been successfully used across many applications of machine learning, from natural language processing [49] and speech recognition [50] to computer vision [51] and drug discovery [52]. Recently, multi-modality learning has been introduced into medical imaging field. However, the problem of small number of subjects and high feature dimensions limits further performance improvement of the multimodal classification methods. Our work aims to provide a novel multi-modality feature selection method which not only reduces irrelevant and redundant features but also considers the local similarity across different imaging modalities. Although the idea of jointly selecting features from multi-modality neuroimaging data has been seen in previous studies [30,46,53], these methods do not consider the potential relationship across different modalities. Besides, underlying data structure in the low-dimensional space may not be revealed in these methods since the neighbors and similarity of the original high-dimensional data are obtained separately from each individual modality.

In this paper, we apply an adaptive similarity learning method to address the above issues. The similarity measured from single- and multi-modality data is learned with the change of low-dimensional representation after feature selection. As can be observed from the experimental results in Section 3.2, our multi-modality feature learning method which adopts adaptive similarity learning method shows better performance than those with fixed

Table 9
Classification accuracy and degradation ratio for AD vs. NC under different noises.

Noise	ASMFS		MTFS		M2TFS	
	Accuracy	Deg. Ratio	Accuracy	Deg. Ratio	Accuracy	Deg. Ratio
Gaussian	95.56%	1.24%	86.31%	6.71%	86.02%	9.45%
Poisson	96.31%	0.46%	87.14%	5.81%	86.99%	8.43%
Salt and Pepper	94.76%	2.06%	86.60%	6.40%	87.18%	8.23%

Table 10
Comparison of classification accuracy of different multi-modality methods.

Methods		Subjects	Modalities	AD vs. NC	MCI vs. NC	MCI-C vs. MCI-NC
Traditional machine learning based methods	Hinrichs et al. [54]	48 AD + 66 NC	MRI + PET	87.6%	–	–
	Huang et al. [55]	49 AD + 67 NC	MRI + PET	94.3%	–	–
	Gray et al. [15]	37 AD + 75 MCI + 35 NC	MRI + PET + CSF + genetic	89.0%	74.6%	58.0%
	Jie et al. [29]	51 AD + 99 MCI + 52 NC	MRI + PET	95.03%	79.27%	68.94%
	Liu et al. [28]	51 AD + 99 MCI + 52 NC	MRI + PET	94.37%	78.80%	67.83%
	Zu et al. [53]	51 AD + 99 MCI + 52 NC	MRI + PET	95.95%	80.26%	69.78%
	Zhu et al. [31]	51 AD + 99 MCI + 52 NC	MRI + PET	95.50%	79.70%	71.20%
Deep learning based methods	Lin et al. [23]	93 AD + 204 MCI + 101 NC	MRI + PET	88.79%	–	73.04%
	Huang et al. [24]	465 AD + 567 MCI + 480 NC	MRI + PET	90.10%	82.58%	72.22%
	Liu et al. [25]	93 AD + 204 MCI + 101 NC	MRI + PET	93.26%	74.34%	–
	Lin et al. [26]	362 AD + 416 MCI + 308 NC	MRI + PET	89.26%	–	72.84%
	Proposed	51 AD + 99 MCI + 52 NC	MRI + PET	96.76%	80.73%	69.41%

similarity based methods, thus the superiority of adaptive similarity learning for feature selection is fully demonstrated. Besides, we list the top 10 brain regions selected by our method and M2TFS in Section 3.3. The results show that our method can identify the brain regions with high clinical relevance. In contrast, the brain regions selected by M2TFS are less effective. Through the experiments in Section 3.4, we can prove that keeping the local manifold structure of data is beneficial to feature selection. To validate the effectiveness of adaptive similarity learning, we construct another framework which learns similarity matrix in different modalities separately in Section 3.5, and compare it with our method. The results prove the positive effect of our proposed adaptive similarity learning on the multi-modality AD classification task. Finally, we investigate the convergence and robustness of our method in Section 3.6 and demonstrate the superiority of our method against several state-of-the-art AD classification methods in Section 3.7.

5. Conclusion

This paper proposes a novel Adaptive-Similarity-based Multi-modality Feature Selection (ASMFS) method for AD classification. Different from previous methods, our proposed ASMFS considers the similarity among multi-modality data and enables a joint learning of feature selection and similarity learning. Specifically, on the one hand, an adaptive learning strategy is developed to help the proposed ASMFS capture the intrinsic data structure of different modality data in the low-dimensional space, thus obtaining the more informative similarity matrix and more discriminative features. On the other hand, the proposed method can fully explore the relationships across modalities and subjects through mining and fusing discriminative features from multi-modality data for AD/MCI classification. Experimental results on the ADNI database demonstrate that our proposed method outperforms the state-of-the-art methods with respect to multimodal classification of AD/MCI.

Despite that we have demonstrated the superiority and effectiveness of our method in AD classification, there are several limitations which should be further considered in future studies. First, in this paper, we only consider two-class classification problems (i.e., AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC), and do not test the ability of our proposed method for the multi-class classification of AD, MCI and NC. Although multi-class classification is more challenging than two-class classification, it is more practical in the clinic since it can diagnose different stages of dementia, which is helpful for doctors to suit the remedy to the case. Second, the proposed method requires the same number of features from different modalities. However, the feature number could be variable in other modality data in the ADNI database, such as CSF and genetic data, which may limit the further application of these modalities in our method. We believe that exploring a method adaptive to the number of features could involve more available modalities and collect more useful information from them, thus enhancing the classification performance. Third, in this work, features are generated from a 93 manual labels template, which may not be sufficient to represent the underlying information from original data. Therefore, it is an interesting future direction to investigate how to make a better choice of template. Finally, we will also investigate the effect of non-handcrafted features using current deep learning techniques and the longitudinal image data on the AD classification task in the future work.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgments

This work is supported by National Natural Science Foundation of China (NSFC 62071314) and Sichuan Science and Technology Program (2021YFG0326, 2020YFG0079).

References

[1] L. Mucke, Alzheimer's disease, Nature 461 (7266) (2009) 895–897, doi:10.1038/461895a.
[2] A. Association, Alzheimer's disease facts and figures, Alzheimer's Dement. 11 (3) (2015) 332–384, doi:10.1016/j.jalz.2015.02.003.
[3] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, H.M. Arrighi, Forecasting the global burden of Alzheimer's disease, Alzheimer's Dement. 3 (3) (2007) 186–191, doi:10.1016/j.jalz.2007.04.381.
[4] L.K. McEvoy, C. Fennema-Notestine, J.C. Roddey, D.J. Hagler, D. Holland, D.S. Karow, C.J. Pung, J.B. Brewer, A.M. Dale, Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment, Radiology 251 (1) (2009) 195–205, doi:10.1148/radiol.2511080924.

[5] F. Bouwman, W. van der Flier, N. Schoonenboom, E. van Elk, A. Kok, F. Rijmen, M. Blankenstein, P. Scheltens, Longitudinal changes of CSF biomarkers in memory clinic patients, Neurology 69 (10) (2007) 1006–1011, doi:10.1212/01.wnl.0000271375.37131.04.

[6] J.C. Morris, M. Storandt, J.P. Miller, D.W. McKeel, J.L. Price, E.H. Rubin, L. Berg, Mild cognitive impairment represents early-stage Alzheimer disease, Arch. Neurol. 58 (3) (2001) 397–405, doi:10.1001/archneur.58.3.397.

[7] A.M. Fjell, K.B. Walhovd, C. Fennema-Notestine, L.K. McEvoy, D.J. Hagler, D. Holland, J.B. Brewer, A.M. Dale, A.D.N. Initiative, et al., CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease, J. Neurosci. 30 (6) (2010) 2088–2101, doi:10.1523/jneurosci.3785-09.2010.

[8] B. Lei, M. Yang, P. Yang, F. Zhou, W. Hou, W. Zou, X. Li, T. Wang, X. Xiao, S. Wang, Deep and joint learning of longitudinal data for Alzheimer's disease prediction, Pattern Recognit. 102 (2020) 107247, doi:10.1016/j.patcog.2020.107247.

[9] M. Liu, D. Zhang, E. Adeli, D. Shen, Inherent structure-based multiview learning with multitemplate feature representation for Alzheimer's disease diagnosis, IEEE Trans. Biomed. Eng. 63 (7) (2016) 1473–1482, doi:10.1109/tbme.2015.2496233.

[10] R. Higdon, N.L. Foster, R.A. Koeppe, C.S. DeCarli, W.J. Jagust, C.M. Clark, N.R. Barbas, S.E. Arnold, R.S. Turner, J.L. Heidebrink, S. Minoshima, A comparison of classification methods for differentiating frontotemporal dementia from Alzheimer's disease using FDG-PET imaging, Stat. Med. 23 (2) (2004) 315–326, doi:10.1002/sim.1719.

[11] K. Walhovd, A. Fjell, A. Dale, L. McEvoy, J. Brewer, D. Karow, D. Salmon, C. Fennema-Notestine, Multi-modal imaging predicts memory performance in normal aging and cognitive decline, Neurobiol. Aging 31 (7) (2010) 1107–1121, doi:10.1016/j.neurobiolaging.2008.08.013.

[12] T. Tong, K. Gray, Q. Gao, L. Chen, D. Rueckert, Multi-modal classification of Alzheimer's disease using nonlinear graph fusion, Pattern Recognit. 63 (2017) 171–181, doi:10.1016/j.patcog.2016.10.009.

[13] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M.K. Chung, S.C. Johnson, Spatially augmented LP boosting for AD classification with evaluations on the ADNI dataset, Neuroimage 48 (1) (2009) 138–149, doi:10.1016/j. neuroimage.2009.05.056.

[14] 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, doi:10.1016/j.neuroimage.2011.01.008.

[15] K.R. Gray, P. Aljabar, R.A. Heckemann, A. Hammers, D. Rueckert, Random forest-based similarity measures for multi-modal classification of Alzheimer's disease, Neuroimage 65 (2013) 167–175, doi:10.1016/j. neuroimage.2012.09.065.

[16] Y. Wang, L. Zhou, B. Yu, L. Wang, C. Zu, D. Lalush, W. Lin, X. Wu, J. Zhou, D. Shen, 3D auto-context-based locality adaptive multi-modality GANs for PET synthesis, IEEE transactions on medical imaging 38 (6) (2018) 1328–1339, doi:10.1109/TMI.2018.2884053.

[17] B. Zhan, J. Xiao, C. Cao, X. Peng, C. Zu, J. Zhou, Y. Wang, Multi-constraint generative adversarial network for dose prediction in radiotherapy, Medical Image Analysis 77 (2022).

[18] Y. Luo, L. Zhou, B. Zhan, Y. Fei, J. Zhou, Y. Wang, D. Shen, Adaptive rectification based adversarial network with spectrum constraint for high-quality PET image synthesis, Medical Image Analysis 77 (2022).

[19] L. Hu, J. Li, X. Peng, J. Xiao, B. Zhan, C. Zu, X. Wu, J. Zhou, Y. Wang, Semi-supervised NPC segmentation with uncertainty and attention guided consistency, Knowledge-based Systems 239 (2022).

[20] P. Tang, P. Yang, D. Nie, X. Wu, J. Zhou, Y. Wang, Unified medical image segmentation by learning from uncertainty in an end-to-end manner, Knowledge-based Systems 241 (2022).

[21] H. Li, X. Peng, J. Zeng, J. Xiao, D. Nie, C. Zu, X. Wu, J. Zhou, Y. Wang, Explainable attention guided adversarial deep network for 3D radiotherapy dose distribution prediction, Knowledge-Based Systems (2022) In press.

[22] Y. Wang, B. Yu, L. Wang, C. Zu, D. Lalush, W. Lin, X. Wu, J. Zhou, D. Shen, L. Zhou, 3D conditional generative adversarial networks for high-quality PET image estimation at low dose, Neuroimage 174 (2018) 550–562.

[23] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, et al., Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment, Front. Neurosci. 12 (2018) 777, doi:10.3389/fnins.2018.00777.

[24] Y. Huang, J. Xu, Y. Zhou, T. Tong, X. Zhuang, ADNI, et al., Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network, Front. Neurosci. 13 (2019) 509, doi:10.3389/fnins.2019. 00509.

[25] M. Liu, D. Cheng, K. Wang, Y. Wang, Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis, Neuroinform 16 (3–4) (2018) 295–308, doi:10.1007/s12021-018-9370-4.

[26] W. Lin, W. Lin, G. Chen, H. Zhang, Q. Gao, Y. Huang, T. Tong, M. Du, A.D.N. Initiative, et al., Bidirectional mapping of brain MRI and PET with 3D reversible GAN for the diagnosis of Alzheimer's disease, Front. Neurosci. 15 (2021) 357, doi:10.3389/fnins.2021.646013.

[27] L. Zhao, Q. Hu, W. Wang, Heterogeneous feature selection with multi-modal deep neural networks and sparse group LASSO, IEEE Trans. Multimed. 17 (11) (2015) 1936–1948, doi:10.1109/tmm.2015.2477058.

[28] F. Liu, C.Y. Wee, H. Chen, D. Shen, Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification, Neuroimage 84 (2014) 466–475, doi:10.1016/j.neuroimage.2013.09.015.

[29] B. Jie, D. Zhang, B. Cheng, D. Shen, A.D.N. Initiative, Manifold regularized multitask feature learning for multimodality disease classification, Hum. Brain Mapp. 36 (2) (2014) 489–507, doi:10.1002/hbm.22642.

[30] X. Hao, Y. Bao, Y. Guo, M. Yu, D. Zhang, S.L. Risacher, A.J. Saykin, X. Yao, L. Shen, Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer's disease, Med. Image Anal. 60 (2020) 101625, doi:10.1016/j.media.2019.101625.

[31] X. Zhu, H.I. Suk, D. Shen, Multi-modality canonical feature selection for Alzheimer's disease diagnosis, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2014, pp. 162–169.

[32] W. Shao, Y. Peng, C. Zu, M. Wang, D. Zhang, Hypergraph based multi-task feature selection for multimodal classification of Alzheimer's disease, Comput. Med. Imaging Gr. 80 (2020) 101663, doi:10.1016/j. compmedimag.2019.101663.

[33] M. Khatami, T. Schmidt-Wilcke, P.C. Sundgren, A. Abbasloo, B. Sch ölkopf, T. Schultz, BundleMAP: anatomically localized classification, regression, and hypothesis testing in diffusion MRI, Pattern Recognit. 63 (2017) 593–600, doi:10.1016/j.patcog.2016.09.020.

[34] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326, doi:10.1126/science.290.5500.2323.

[35] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004, doi:10.1017/cbo9780511804441.

[36] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (11) (2021).

[37] L.K. Saul, S.T. Roweis, Think globally, fit locally: unsupervised learning of low dimensional manifolds, Departmental Papers (CIS) (2003) 12.

[38] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint $\lambda_{2,1}$-norms minimization, Adv. Neural Inf. Process. Syst. 23 (2021).

[39] F.R. Bach, G.R. Lanckriet, M.I. Jordan, Multiple kernel learning, conic duality, and the SMO algorithm, in: Proceedings of the Twenty-first International Conference on Machine Learning - ICML '04, ACM Press, 2004, p. 6, doi:10.1145/1015330.1015424.

[40] S. Kl öppel, C.M. Stonnington, C. Chu, B. Draganski, R.I. Scahill, J.D. Rohrer, N.C. Fox, C.R. Jack Jr, J. Ashburner, R.S. Frackowiak, Automatic classification of MR scans in Alzheimer's disease, Brain 131 (3) (2008) 681–689.

[41] B. Jie, M. Liu, J. Liu, D. Zhang, D. Shen, Temporally constrained group sparse learning for longitudinal data analysis in Alzheimer's disease, IEEE Trans. Biomed. Eng. 64 (1) (2017) 238–249, doi:10.1109/tbme.2016. 2553663.

[42] D. Zhang, D. Shen, A.D.N. Initiative, Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers, PLoS ONE 7 (3) (2012) e33182, doi:10.1371/journal.pone.0033182.

[43] J. Sled, A. Zijdenbos, A. Evans, A nonparametric method for automatic correction of intensity nonuniformity in MRI data, IEEE Trans. Med. Imaging 17 (1) (1998) 87–97, doi:10.1109/42.668698.

[44] Y. Zhang, J.M. Brady, S. Smith, Hidden Markov random field model for segmentation of brain MR image, in: Medical Imaging 2000: Image Processing, 3979, International Society for Optics and Photonics, SPIE, 2000, pp. 1126–1137, doi:10.1117/12.387617.

[45] R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective, J. R. Stat. Soc. B 73 (3) (2011) 273–282, doi:10.1111/j.1467-9868. 2011.00771.x.

[46] 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, doi:10.1016/j.neuroimage.2011. 09.069.

[47] N.C. Fox, J.M. Schott, Imaging cerebral atrophy: normal ageing to Alzheimer's disease, Lancet 363 (9406) (2004) 392–394, doi:10.1016/s0140-6736(04)15441-x.

[48] C. Misra, Y. Fan, C. Davatzikos, Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI, Neuroimage 44 (4) (2009) 1415–1422.

[49] R. Kiros, R. Salakhutdinov, R. Zemel, Multimodal neural language models, in: Proceedings of the International Conference on Machine Learning, PMLR, 2014, pp. 595–603.

[50] Y. Mroueh, E. Marcheret, V. Goel, Deep multimodal learning for audio-visual speech recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, IEEE, 2015, pp. 2130–2134, doi:10.1109/icassp.2015.7178347.

[51] D. Jimenez Rezende, S. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, N. Heess, Unsupervised learning of 3d structure from images, Adv. Neural Inf. Process. Syst. 29 (2016) 4996–5004.

[52] S. Makrogiannis, J. Wellen, Y. Wu, L. Bloy, S.K. Sarkar, A multimodal image registration and fusion methodology applied to drug discovery research, in: Proceedings of the IEEE 9th Workshop on Multimedia Signal Processing, IEEE, IEEE, 2007, pp. 324–327, doi:10.1109/mmsp.2007.4412883.

[53] C. Zu, B. Jie, M. Liu, S. Chen, D. Shen, D. Zhang, Label-aligned multi-task feature learning for multimodal classification of Alzheimer's disease and mild cognitive impairment, Brain Imaging Behav. 10 (4) (2015) 1148–1159, doi:10.1007/s11682-015-9480-7.

[54] C. Hinrichs, V. Singh, G. Xu, S.C. Johnson, Predictive markers for AD in a multimodality framework: an analysis of MCI progression in the ADNI population, Neuroimage 55 (2) (2011) 574–589, doi:10.1016/j. neuroimage.2010.10.081.

[55] S. Huang, J. Li, J. Ye, T. Wu, K. Chen, A. Fleisher, E. Reiman, Identifying Alzheimer's disease-related brain regions from multi-modality neuroimaging data using sparse composite linear discrimination analysis, Adv. Neural Inf. Process. Syst. 24 (2021).

[56] F. Nie, W. Zhu, X. Li, Unsupervised feature selection with structured graph optimization, in: Proceedings of the AAAI Conference on Artificial Intelligence, 30, 2016.
[57] S.P. Poulin, R. Dautoff, J.C. Morris, L.F. Barrett, B.C. Dickerson, Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity, Psychiatry Res. Neuroimaging 194 (1) (2011) 7–13, doi:10.1016/j.pscychresns.2011.06.014.

Yuang Shi is an undergraduate in the School of Computer Science, Sichuan University. His-research interests include machine learning and medical image analysis.

Chen Zu received the Ph.D. degree from Nanjing University of Aeronautics and Astronautics (NUAA) in 2018. He is currently a senior engineer in JD.COM. His-research interests include data mining, data processing and machine learning.

Mei Hong is a professor and Associate Dean of College of Computer Science & College of Software Engineering, Sichuan University. Her research interests include computer vision, air traffic control data processing, real-time software analysis, and testing and automated software testing.

Luping Zhou is a senior lecturer and ARC DECRA Fellow in the School of Electrical and Information Engineering, University of Sydney. She currently focuses on medical image analysis with statistical graphical models and deep learning.

Lei Wang is an associate professor of Faculty of Engineering and Information Sciences, University of Wollongong. His-research interest lies in machine learning, pattern recognition, and computer vision.

Xi Wu is a Professor and deputy dean of Department of Computer Science, Chengdu University of Information Technology. Dr. Wu's main research area is the development of novel methods for analysis of imaging data.

Jiliu Zhou is associated with Sichuan University and Chengdu University of Information Technology as full professor. His-research interests include image analysis and signal processing, mobile computing, medical image analysis, and intellectual computing.

Daoqiang Zhang is currently a professor in the Department of Computer Science and Engineering, NUAA. His-research interests include machine learning, pattern recognition, data mining, and medical image analysis.

Yan Wang received the Ph.D. degree from Sichuan University in 2015. She is currently an Associate Professor at School of Computer Science, Sichuan University. She studied at University of North Carolina at Chapel Hill, USA and University of Wollongong, Australia as joint training pH. D. student and Post doctorate in 2014–2015 and 2017–2018, respectively. Her research interests include computer vision, machine learning, deep learning, medical image analysis.