Brain Imaging and Behavior (2019) 13:27–40
DOI 10.1007/s11682-017-9731-x

ORIGINAL RESEARCH

Discriminative self-representation sparse regression for neuroimaging-based alzheimer's disease diagnosis

Xiaofeng Zhu[1] · Heung-Il Suk[2] · Seong-Whan Lee[2] · Dinggang Shen[1,2]

Published online: 17 June 2017

Abstract In this paper, we propose a novel feature selection method by jointly considering (1) '*task-specific*' relations between response variables (e.g., clinical labels in this work) and neuroimaging features and (2) '*self-representation*' relations among neuroimaging features in a sparse regression framework. Specifically, the task-specific relation is devised to learn the relative importance of features for representation of response variables by a linear combination of the input features in a supervised manner, while the self-representation relation is used to take into account the inherent information among neuroimaging features such that any feature can be represented by a weighted sum of the other features, regardless of the label information, in an unsupervised manner. By integrating these two different relations along with a group sparsity constraint, we formulate a new sparse linear regression model for class-discriminative feature selection. The selected features are used to train a support vector machine for classification. To validate the effectiveness of the proposed method, we conducted experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset; experimental results showed superiority of the proposed method over the state-of-the-art methods considered in this work.

Keywords Alzheimer's disease (AD) · Mild cognitive impairment (MCI) · Feature selection · Joint sparse learning · Self-representation

✉ Dinggang Shen
dgshen@med.unc.edu

Xiaofeng Zhu
xiaofeng@med.unc.edu

Seong-Whan Lee
sw.lee@korea.ac.kr

Heung-Il Suk
hisuk@korea.ac.kr

[1] Department of Radiology and BRIC, The University of North Carolina at Chapel Hill, Chapel Hill, USA

[2] Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea

Introduction

Alzheimer's Disease (AD), the most common form of dementia around the world, is the sixth leading cause of death in the United States. According to the report of the Alzheimer's Association[1] and the Centers for Disease Control[2], in 2015, nearly 44 million people had AD or related dementia and only 1-in-4 AD patients had been diagnosed worldwide. In 2015, the population of AD, in the United States, was around 5.3 million, with the chance of the number increasing to more than 16 million by 2050. Moreover, in 2015, the global cost of AD was estimated to be 605 billion, equivalently 1% of the entire world's Gross Domestic Product (GDP), while the cost in the United States was around 226 billion. Thus, the treatment of AD is placing a huge financial burden on society.

Studies have demonstrated that early diagnosis of AD and its early stage (i.e., Mild Cognitive Impairment (MCI)) are of high importance clinically, as effective treatments on early-stage patients would have more influence for slowing down disease progression. However, current clinical assessments, e.g., Mini-Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale-Cognitive subscale

[1] http://www.alz.org/

[2] http://www.alzheimers.net/resources/alzheimers-statistics/

(ADAS-Cog), still present low sensitivity and specificity in early diagnosis of AD. Recently, a number of studies (Mwangi et al. 2014; Suk et al. 2015b; Thung et al. 2015; Zhu et al. 2014d) utilized neuroimaging techniques to show different soft tissues in the brain with good contrast and thus present important information about brain atrophy possibly caused by neurodegeneration.

Neuroimaging tools, such as Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and functional MRI, have become powerful tools for characterizing neurodegenerative progress of AD by helping overcome the limitations of the conventional cognitive assessments, such as imprecise diagnosis and low diagnosis confidence (Suk et al. 2015a, b; Thung et al. 2014; Zhu et al. 2014c, 2015b). For example, (Greicius et al. 2004) showed that the disrupted functional connectivity between posterior cingulate and hippocampus led to the posterior cingulate hypometabolism. (Guo et al. 2010) indicated that AD patients exhibited significant decreases in gray matter volume in the hippocampus, parahippocampal gyrus, insula, and superior temporal gyrus.

With a large amount of the neuroimaging dataset publicly available on the web, e.g., the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, machine learning techniques have been playing a core role in investigation and analysis of high-dimensional neuroimaging data, providing unprecedented opportunities to investigate AD-related problems with high confidence and precision. For example, (Duchesne et al. 2009) employed linear regression models to estimate one-year MMSE changes from structural MRI; (Wang et al. 2010) designed high-dimensional kernel-based regression methods to estimate ADAS-Cog and MMSE.

Among machine learning techniques for the analysis of neuroimaging data, sparse linear regression has attracted researchers' interests by modeling the relation between representation of neuroimaging data and clinical response variables, such as clinical scores and clinical status. For example, (Zhu et al. 2014c) designed a regression model on multi-modality neuroimaging data for joint prediction of clinical scores (including ADAS-Cog and MMSE) and clinical status. Although the linear regression model is helpful for finding the relationship between neuroimaging features and response variables, it naturally suffers from a high-dimensional problem, which is very common in neuroimaging data analysis. The straightforward approach to circumvent the so-called 'curse-of-dimensionality' is to collect as many training samples as possible. It is, however, not feasible in reality due to high cost and time consumption.

Recently, subspace learning (Spedding et al. 2015; Lee et al. 2014) and feature selection (López-de-Ipiña et al. 2015) were commonly used to reduce the number of features from high-dimensional neuroimaging data to deal with the issue of 'curse-of-dimensionality'. For example, (Zhu et al. 2015a) proposed to conduct feature selection by transferring original features from different modalities to a common space with Canonical Correlation Analysis (CCA) for jointly predicting clinical scores (e.g., ADAS-Cog and MMSE) and identifying multi-class disease status. Usually, feature selection methods (Wang et al. 2011; Zhang and Shen 2012; Zhu et al. 2014a) find informative feature subsets from an original feature set, while subspace learning methods, such as Fisher's Linear Discriminant Analysis (LDA) (Spedding et al. 2015; Zhu et al. 2015b), CCA (Kim et al. 2014; Lorenzi et al. 2016), and Locality Preserving Projection (LPP) (Zhu et al. 2015b), transform original features into a low-dimensional space (Duda et al. 2012). In regards to the interpretability of the results, feature selection methods are more preferable over subspace learning methods, particularly in neuroimaging studies, as selected features can directly link anatomical structures for providing intuitive understanding of disease.

In this work, we focus on a feature selection method and propose to consider different kinds of relations inherent in data with the goal of selecting brain regions related to AD for clinical diagnosis and improving the performance of representative features for identifying AD status. Specifically, we propose a self-representation feature selection regression model to select a representative feature subset by simultaneously considering a '*sample-level*' relation between the features and response variables as well as a '*feature-level*' relation among the features. The goal of our method is to use the '*sample-level*' relation to conduct a task-oriented supervised step and use the '*feature-level*' relation to conduct a self-representation-oriented unsupervised step. Specifically, we first define an objective function with an element-wise similarity loss function (i.e., least square loss function) and the self-representation property of features, to characterize the sample-level relation and the feature-level relation, respectively. Then, we penalize our objective function with an $\ell_{2,1}$-norm regularizer to output representative features. Furthermore, we iteratively optimize these two steps such that each of them may adjust the other in order to achieve an optimal process of feature selection. Finally, the selected features are fed into a Support Vector Machine (SVM) classifier for clinical diagnosis.

Compared to the previous state-of-the-art feature selection methods for AD diagnosis, the main contributions of our work are three-fold. First, we propose a novel sparse feature selection method by exploiting the inherent structures of the features along with the relations between response variables and neuroimaging features. The rationale of our method is that, the features are dependent in real applications, thus each feature can be (sparsely) represented by other features. Moreover, if a feature is important for the classification task, then it is reasonable to assume that the feature can be also informative to represent other features.

Meanwhile, the designed model should also achieve the minimum regression error measured by the residual between the response matrix and its prediction.

Second, unlike the existing methods in sparse learning (Elhamifar and Vidal 2009; Yuan and Lin 2006) and low-rank representation (Liu et al. 2010), which considered the self-representation property of data from the view of relations among samples, this work considers such properties from aspects of relations among features for self-representation as well as relations among samples for target tasks, i.e., clinical status identification or clinical scores prediction. Our motivation for this is that, the task-oriented feature selection strategy pursues the minimum regression error under a supervised learning concept, while the self-representation-oriented strategy selects features that are involved in representing other features under a unsupervised learning concept. These two strategies are integrated in a unified framework of sparse linear regression.

Last but not least, this work simultaneously considers binary classification and multi-class classification, instead of only conducting binary classification, as most of the state-of-the-art methods did for AD diagnosis, such as (Suk et al. 2015a; Wang et al. 2011; Zhang and Shen 2012). In real clinical applications, given neuroimaging data of an subject, he or she can be categorized into one of the following status, such as AD, Normal Control (NC), progressive MCI (pMCI) and stable MCI (sMCI). This obviously belongs to a multi-class classification problem.

Materials and image preprocessing

For performance evaluation, we use the ADNI-1 dataset.[3] The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations. The main goal of ADNI is to test if MRI, PET, other biological markers, and clinical and neuropsychological assessments can be combined to predict the progression of MCI and early AD. To this end, ADNI has recruited over 800 adults (age range $\in [55, 90]$) to participate in research. Specifically, around 200 cognitively normal older individuals were followed for 3 years, 400 people with MCI were followed for 3 years, and 200 people with early AD were followed for 2 years.[4] The research protocol was approved by each local institutional review board and written informed consent was obtained from each participant for the study.

Table 1 Demographic and clinical information of the subjects

	NC (N = 226)	AD (N = 186)	MCI (N = 393)
Male/Female	118/108	99/87	253/140
Age (Mean ± SD)	75.8±5.0	75.3±7.5	74.8±7.3
	[59.9, 89.6]	[55.1, 90.9]	[54.4, 89.3]
Education (Mean ± SD)	16.0±2.9	14.7±3.1	15.6±3.0
	[6, 20]	[4, 20]	[4, 20]
MMSE (Mean ± SD)	29.1±1.0	23.3±2.2	27.0±1.8
	[25, 30]	[18, 27]	[23, 30]
CDR (Mean ± SD)	0±0	0.75±0.3	0.50±0
	[0, 0]	[0.5, 1]	[0.5, 0.5]

(MMSE: Mini-Mental State Examination, CDR: Clinical Dementia Rating, N: number of subjects, SD: Standard Deviation, [min, max])

Subjects

General inclusion/exclusion criteria with respect to the general eligibility criteria in ADNI are as follows[5]: The MMSE score of an NC subject is between 24 and 30. Their Clinical Dementia Rating (CDR) is of 0. Moreover, the NC subject is non-depressed, non MCI, and non-demented. The MMSE score of an MCI subject is between 24 and 30. Their CDR is of 0.5. Moreover, each MCI subject is an absence of significant level of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia. The MMSE score of a mild AD subject is between 20 and 26, with the CDR of 0.5 or 1.0, and meets the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS/ADRDA) criteria for probable AD.

In our experiments, we only use the baseline MRI data acquired from 226 NC, 186 AD, and 393 MCI subjects. MCI subjects were clinically further subdivided into pMCI, sMCI, ncMCI, and uMCI. 118 pMCI subjects indicate that the subjects converted from MCI to AD in 24 months, while 124 sMCI subjects didn't not convert to AD in both 24 months and 36 months. Besides, 49 ncMCI subjects did not convert in 24 months but converted in 36 months, while 102 uMCI subjects were MCI at base line but were never converted to AD at any available time points among 0 – 96 months. We summarize the demographics of the subjects in Table 1.

Image preprocessing

We downloaded raw Digital Imaging and Communications in Medicine (DICOM) MRI scans from the public ADNI

[3] Available at 'http://www.loni.usc.edu/ADNI'

[4] Please refer to 'www.adni-info.org' for up-to-date information.

[5] Refer to 'http://www.adni-info.org' for more details.

website. These MRI scans were already reviewed for quality, and automatically corrected for spatial distortion caused by gradient nonlinearity and B1 field inhomogeneity.

The image processing for all MR images was conducted following the procedures in (Zhu et al. 2014b). Specifically, we first performed anterior commissure-posterior commissure correction using MIPAV software[6] on all images, and applied the algorithm in (Sled et al. 1998) to correct intensity inhomogeneity. Second, we extracted a brain on all structural MR images using a robust skull-stripping method, followed by manual edition and intensity inhomogeneity correction. After removal of cerebellum, based on registration and intensity inhomogeneity correction by repeating N3 (Sled et al. 1998) for three times, we used FAST algorithm in the FSL package (Zhang et al. 2001) to segment the structural MR images into three different tissues: Gray Matter (GM), White Matter (WM), and CerebroSpinal Fluid (CSF). Next, we used HAMMER (Shen and Davatzikos 2002) for registration and obtained the Region-Of-Interest (ROI)-labeled image based on the Jacob template (Kabani 1998), which dissects a cerebrum into 93 ROIs. For each of all 93 ROIs in the labeled image of a subject, we computed the GM tissue volumes of the ROIs and used them as structural features. With this, we acquired 93 features from an MRI image.

Method

In this section, we describe our framework for AD classification by proposing a novel feature selection method. Figure 1 presents a schematic diagram of our method for predicting a class label. We first extract features from MRI data and then construct a feature matrix **X**, with the feature vectors of MRI data, and a corresponding response matrix **Y** representing a class label at each column. With our new feature selection method, we select representative features and then use them to train a Support Vector Machine (SVM) for clinical label identification.

Notations

In this paper, we denote matrices as boldface uppercase letters, vectors as boldface lowercase letters, and scalars as normal italic letters, respectively. For a matrix $\mathbf{X} = [x_{ij}]$, its i-th row and j-th column are denoted as $\mathbf{x}^i$ and $\mathbf{x}_j$, respectively. Also, we denote the Frobenius norm and $\ell_{2,1}$-norm of a matrix $\mathbf{X}$ as $\|\mathbf{X}\|_F = \sqrt{\sum_i \|\mathbf{x}^i\|_2^2} = \sqrt{\sum_j \|\mathbf{x}_j\|_2^2}$ and $\|\mathbf{X}\|_{2,1} = \sum_i \|\mathbf{x}^i\|_2 = \sum_i \sqrt{\sum_j x_{ij}^2}$, respectively. We further denote the transpose operator, the trace operator, and the inverse of a matrix $\mathbf{X}$ as $\mathbf{X}^T$, $tr(\mathbf{X})$, and $\mathbf{X}^{-1}$, respectively.

Feature selection with a sparse linear regression model

Let $\mathbf{X} = [\mathbf{x}^1; ...; \mathbf{x}^n] = [\mathbf{x}_1, ..., \mathbf{x}_d] \in \mathbb{R}^{n\times d}$ (where $\mathbf{x}^i \in \mathbb{R}^{1\times d}$ and $\mathbf{x}_j \in \mathbb{R}^{n\times 1}$, $i = 1, ..., n$; $j = 1, ..., d$) and $\mathbf{Y} = [\mathbf{y}^1; ...; \mathbf{y}^n] \in \{0, 1\}^{n\times c}$, respectively, be the feature matrix and the clinical status (*a.k.a.,* response matrix) of all MRI data, where n, d, and c denote the numbers of samples (or subjects), feature variables, and response variables, respectively. We use a class indicator vector for response variables by denoting the class label of the i-th sample $\mathbf{x}^i$ as $\mathbf{y}^i = [y_{i1}, ..., y_{ij}, ..., y_{ic}] \in \{0, 1\}^c$, where y_{ij} denotes the label information of the j-th class of the i-th sample $\mathbf{x}^i$ and $y_{ij} = 1$ if $\mathbf{x}^i$ belongs to the j-th class, while $y_{ij} = 0$ otherwise.

A linear regression model formulates a linear relation between feature variables and the response variables as follows:

$$\mathbf{Y} \approx \mathbf{XW} + \mathbf{eb} = \hat{\mathbf{Y}} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d\times c}$ is a regression coefficient matrix or weight matrix, $\hat{\mathbf{Y}}$ is the prediction of $\mathbf{X}$ on the space spanned by $\mathbf{XW}$, $\mathbf{b} \in \mathbb{R}^{1\times c}$ is a bias term, and $\mathbf{e} \in \mathbb{R}^{n\times 1}$ denotes a column vector with all ones. To find optimal coefficients of the weight matrix $\mathbf{W}$, with which the response variables in $\mathbf{Y}$ are represented by a linear combination of the features $\mathbf{X}$, we use the least squares error for a loss function $l(\mathbf{W})$ as follows:

$$\begin{aligned} l(\mathbf{W}) &= \|\mathbf{Y} - (\mathbf{XW} + \mathbf{eb})\|_F^2 \\ &= \|\mathbf{Y} - \hat{\mathbf{Y}}\|_F^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{c}(y_{ij} - \hat{y}_{ij})^2. \end{aligned} \tag{2}$$

From a matrix similarity point of view, Eq. 2 computes the similarity between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ with the sum of elementwise differences between them. The lower the loss, the more similar they are.

Task-oriented feature selection

As for feature selection, we assume that a small number of features is useful to represent the response variables. Furthermore, since the clinical status, i.e., the response variables in $\mathbf{Y}$, has high neurophysiological relations among them, it is commonplace to assume that the same neuroimaging features are used to represent the response

[6] Available at ‘http://mipav.cit.nih.gov/clickwrap.php’

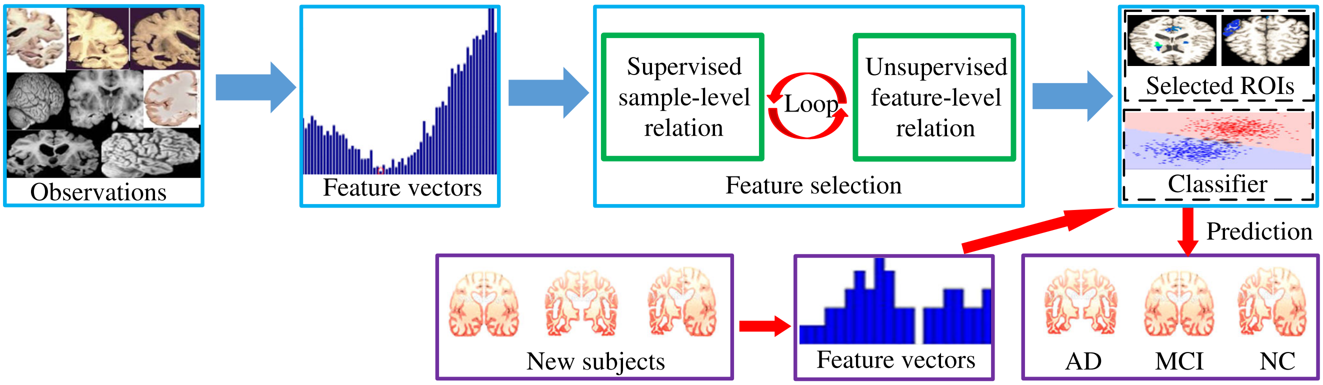


Fig. 1 The framework of the proposed method

variables. To this end, we add an $\ell_{2,1}$-norm regularizer into the loss function in Eq. 2 as follows:

$$\min_{\mathbf{W},\mathbf{b}} \|\mathbf{Y} - \mathbf{XW} - \mathbf{eb}\|_F^2 + \lambda\|\mathbf{W}\|_{2,1}, \tag{3}$$

where λ is a sparsity control parameter. The $\ell_{2,1}$-norm regularizer $\|\mathbf{W}\|_{2,1}$ penalizes all coefficients in the same row of $\mathbf{W}$ together for jointly selecting or un-selecting features for predicting the response variables (Suk et al. 2015a; Yuan and Lin 2006; Zhang and Shen 2012). Specifically, the $\ell_{2,1}$-norm regularizer results in zero-rows in $\mathbf{W}$ after optimizing (1). The features that are multiplied with zero-rows do not play any role to represent the response variables in $\mathbf{Y}$, and thus can be removed from the feature set. From a clinical standpoint, the selected features for representation of the response variables can be interpreted as potential biomarkers[7] to identify a clinical status. Since the solution of $\mathbf{W}$ is oriented to response variables of the clinical values, i.e., clinical status, the prediction of which is the main task of this work, we call it a '*task-oriented*' method.

Self-representation-oriented feature selection

The main assumption in Eq. 1 is that the features that are important to represent a response variable are also informative of other response variables, while features uninformative to represent a response variable are also uninformative of other response variables. Thus, such features should be jointly selected or un-selected in representing the response variables, i.e., clinical status.

In this section, we propose to further include an important characteristic among features by maximally utilizing the information inherent in observations $\mathbf{X}$ into the objective function as a regularizer. Note that AD may affect multiple brain regions simultaneously, rather than just a single region (Zhang et al. 2011; Zhang and Shen 2012). Justified by neurophysiological characteristics observed in the previous AD studies, we assume that there are dependencies among ROIs (i.e., features). We devise a new regularizer to utilize this relational characteristic among ROIs in feature selection. Specifically, we define a linear regression model such that each feature $\mathbf{x}_i$ in $\mathbf{X}$ can be represented as a linear combination of other features (including itself):

$$\mathbf{x}_i \approx \sum_{j=1}^{d} (\mathbf{x}_j s_{ji} + \mathbf{e}p_i), i, j = 1, ..., d. \tag{4}$$

where the element s_{ji} of matrix $\mathbf{S}$ is a weight coefficient between the i-th feature vector $\mathbf{x}_i$ and the j-th feature vector $\mathbf{x}_j$ and $\mathbf{p} = [p_1, ..., p_j, ..., p_d] \in \mathbb{R}^{1\times d}$ is a bias term.

By regarding the prediction of each feature as a task and constraining the sparsity across tasks with an $\ell_{2,1}$-norm regularizer, we define a new objective function as follows:

$$\min_{\mathbf{S},\mathbf{p}} \|\mathbf{X} - \mathbf{XS} - \mathbf{ep}\|_F^2 + \alpha\|\mathbf{S}\|_{2,1}, \tag{5}$$

where α is a sparsity control parameter. The $\ell_{2,1}$-norm regularizer $\|\mathbf{S}\|_{2,1}$ penalizes all coefficients in the same row of $\mathbf{S}$ together for joint selection or un-selection in predicting the feature matrix $\mathbf{X}$. Note that since a vector $\mathbf{x}_i$ in the observation $\mathbf{X}$ can be used to represent itself[8], there always exists a feasible (trivial) solution. That is, its corresponding coefficient in $\mathbf{S}$ equals to one and all the other coefficients equal to zero. However, due to our assumption of dependencies among ROIs, i.e., $rank(\mathbf{X}) < \min(n, d)$, there also exist

[7] Note that we use a set of ROI volumes as features.

[8] The term 'self-similarity', widely used in machine learning and computer vision, such as the literature (Liu et al. 2010; Zhu et al. 2015), indicates that each sample/feature can be represented by both other samples/features and itself. In this work, we assume that features are dependent, so it is reasonable to indicate that each ROI (or feature) can be sparsely represented by all ROIs (or features).

non-trial solutions in the space of $\mathbf{I} - null(\mathbf{X})$ (Liu et al. 2010). To differentiate from the 'task-oriented' method, we call Eq. 5 a '*self-representation-oriented*' method.

The first term in Eq. 5 measures the distance of all column vectors in $\mathbf{X}$ to the subspace spanned by $\mathbf{XS}$, thus the optimal matrix $\mathbf{S}^*$ clearly makes the regression error between $\mathbf{X}$ and its predictions $\mathbf{XS}^*$ as small as possible. On the other hand, the optimal matrix $\mathbf{S}^*$ also reflects the importance of different features. If a feature is important for the model, then it should participate in the representation of other features and help lead to a significant representation coefficient in the row, and vice versa. Specifically, we approximately extend Eq. 4 (by ignoring the bias term) to have:

$$\begin{cases} \mathbf{x}_1 \approx s_{11}\mathbf{x}_1 + \cdots + s_{i1}\mathbf{x}_i + \cdots + s_{d1}\mathbf{x}_d \\ \vdots \\ \mathbf{x}_i \approx s_{1i}\mathbf{x}_1 + \cdots + s_{ii}\mathbf{x}_i + \cdots + s_{di}\mathbf{x}_d \\ \vdots \\ \mathbf{x}_d \approx s_{1d}\mathbf{x}_1 + \cdots + s_{id}\mathbf{x}_i + \cdots + s_{dd}\mathbf{x}_d . \end{cases} \tag{6}$$

Equation 6 indicates that each feature vector e.g., $\mathbf{x}_i$ in left-hand side of Eq. 6 is represented by a linear combination of the feature vectors (including itself) in right-hand side of Eq. 6 and the corresponding weight vector is the i-th column $\mathbf{s}_i$ of $\mathbf{S}$. Obviously, the larger the values in the $\mathbf{s}_i$, the more the corresponding feature vectors involved in the representation of the feature vector $\mathbf{x}_i$. If there is a zero-row in the optimal matrix $\mathbf{S}^*$, e.g., a zero-vector $\mathbf{s}^j = [s_{j1}, ..., s_{j,j}, ..., s_{jd}]$, then the corresponding feature i.e., $\mathbf{x}_j$ in right-hand side of Eq. 6 will not participate in the representation of features. The features participating in the representation of all features should be important, while those not participating in the representation process should be discarded by means of feature selection.

Proposed objective function

By simultaneously considering aforementioned constraints, we combine Eq. 3 with Eq. 5 as follows:

$$\min_{\mathbf{W},\mathbf{b},\mathbf{S},\mathbf{p}} \|\mathbf{Y} - \mathbf{XW} - \mathbf{eb}\|_F^2 + \lambda\|\mathbf{W}\|_{2,1}$$
$$+\beta\|\mathbf{X} - \mathbf{XS} - \mathbf{ep}\|_F^2 + \alpha\|\mathbf{S}\|_{2,1}$$

where $\beta > 0$ is a tuning parameter. Equation 7 conducts two feature selection models (i.e., $\mathbf{W}$ and $\mathbf{S}$), the one under supervised learning (i.e., the first two terms) and the others under unsupervised learning (i.e., the last two terms). However, in this work, we simultaneously consider two models (the task-oriented relations (i.e., the difference between $\mathbf{Y}$ and $\mathbf{XW}$) and the self-representation-oriented relations (i.e., the difference between $\mathbf{X}$ and $\mathbf{XS}$)) for feature selection. Ultimately, the optimal solutions of $\mathbf{W}^*$ and $\mathbf{S}^*$ are aimed to select informative features for clinical response representation. To this end, we integrate the matrices of $\mathbf{W}$ and $\mathbf{S}$ into a single matrix by letting $\lambda = \alpha$ as follows:

$$\min_{\mathbf{W},\mathbf{b},\mathbf{S},\mathbf{p}} \|\mathbf{Y}-\mathbf{XW}-\mathbf{eb}\|_F^2+\beta\|\mathbf{X}-\mathbf{XS}-\mathbf{ep}\|_F^2+\lambda\|[\mathbf{W},\mathbf{S}]\|_{2,1} \tag{7}$$

where $[\mathbf{W}, \mathbf{S}] = \begin{pmatrix} w_{11} & \dots & w_{1c} & s_{11} & \dots & s_{1d} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ w_{d1} & \dots & w_{dc} & s_{d1} & \dots & s_{dd} \end{pmatrix} \in \mathbb{R}^{d\times(c+d)}$ is defined as a joint analyzer (Wang et al. 2011), i.e., a horizontal concatenation of $\mathbf{W}$ and $\mathbf{S}$. Specifically, $\tilde{\mathbf{W}} = [\mathbf{W}, \mathbf{S}]$ reflects the importance of features for joint representation of the response variables and the features.

An intuitive interpretation of the propose method is as follows: It alternately optimizes a supervised step i.e., $\mathbf{W}$ and $\mathbf{b}$ in Eq. 7 and a unsupervised step i.e., $\mathbf{S}$ and $\mathbf{p}$ in Eq. 7, where the supervised step i.e., the first term in Eq. 7 learns a task-specific feature selection model and the unsupervised step i.e., the second term in Eq. 7 learns a self-representation feature selection model. Concretely, the proposed method uses the element-wise similarity measure under the supervision of label information to learn a feature selection model $\mathbf{W}$ and $\mathbf{b}$, and also uses the self-representation property of features to learn another feature selection model $\mathbf{S}$ and $\mathbf{p}$. As these two models jointly conduct feature selection on the same feature matrix $\mathbf{X}$, they should be designed to output the same sparsity. In this way, after optimizing $\tilde{\mathbf{W}}$, the $\ell_{2,1}$-norm regularizer pushes Eq. 7 to output zero values through the whole rows of $\tilde{\mathbf{W}}$, i.e., the same sparse rows on both $\mathbf{W}$ and $\mathbf{S}$. This process that simultaneously satisfies two constraints (i.e., the sample-level relation constraint and the feature-level relation constraint) makes the selection of informative features more confident.

Finally, after optimizing Eq. 7, we can discard the shared irrelevant or noisy components (i.e., the features whose regression coefficient vectors are zero in the rows on both $\mathbf{W}$ and $\mathbf{S}$). Given the selected representative features, we use them to train an SVM as a classifier.

The following remarks show two aspects of the importance of the self-representation characteristics inherent in observed features and the rationale of the combination on the 'task-oriented' feature selection and the 'self-representation-oriented' feature selection.

Remark 1 In this section, we reveal the feature-level relation among self-representation, which characterizes the property that each feature can be linearly approximated by a subset of other features (called as representative features in this work) in feature selection. Self-similarity has been widely used in computer vision and machine learning (Zhu et al. 2015). In computer vision, non-local self-similarity means that patches at different locations in an image may be similar to each other. In machine learning, self-representation can also be modeled as a sparse representation model or a low-rank representation model depending on tasks. However, previous models belong to sample-level representation, while our self-representation regularizer is feature-level representation. In other words, the goal of previous literature is to sparsely represent each sample by other samples (e.g., $\|\mathbf{X}^T - \mathbf{X}^T\mathbf{M}\|_F^2 \Rightarrow \mathbf{x}^i \approx \sum_{j=1}^{n} \mathbf{x}^j m_{ij}$, where m_{ij} is the similarity between the i-th sample $\mathbf{x}^i$ and the j-th sample $\mathbf{x}^j$, $i, j = 1, ..., n$.), while our goal is designed to sparsely represent each feature by other features, i.e., $\mathbf{x}_i \approx \sum_{j=1}^{d} \mathbf{x}_j s_{ji}, i, j = 1, ..., d$.

Remark 2 Self-representation is designed on the assumption that there is redundancy in features as features are linearly correlated Such assumption has been used in a lot of literature on either feature selection or subspace learning for the analysis of high-dimensional neuroimaging data (Suk et al. 2015a; Wang et al. 2011; Zhang and Shen 2012; Zhu et al. 2014a). On the other hand, The proposed self-representation regularizer is designed for unsupervised learning, as label information can be used to enhance the performance of feature selection models. Hence, it is reasonable to add the element-wise similarity loss function to conduct supervised learning by making use of the label information. Mathematically, it is very similar to conduct unsupervised feature selection in supervised learning, with another assumption that the original feature gets best reconstruction by the self-representation property of features in unsupervised learning. Moreover, both the element-wise similarity loss function and the self-representation regularizer enable the proposed framework (with an intuitive and easy way) to select representative features.

Optimization

The objective function in Eq. 7 is convex but non-smooth. In this work, we utilize the framework of iteratively reweighted least square (Wipf and Nagarajan 2010) to optimize Eq. 7.

With simple algebraic operations, Eq. 7 becomes

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{b},\mathbf{S},\mathbf{p}} \; tr((\mathbf{Y}-\mathbf{XW}-\mathbf{eb})^T(\mathbf{Y}-\mathbf{XW}-\mathbf{eb})) \\ &+\beta tr((\mathbf{X}-\mathbf{XS}-\mathbf{ep})^T(\mathbf{X}-\mathbf{XS}-\mathbf{ep})) + \lambda tr(\tilde{\mathbf{W}}^T\mathbf{D}\tilde{\mathbf{W}})\end{aligned} \tag{8}$$

where $\tilde{\mathbf{W}} = [\mathbf{W}, \mathbf{S}] \in \mathbb{R}^{d\times(c+d)}$ and $\mathbf{D} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with $\mathbf{D}_{ii} = \frac{1}{2\|\tilde{\mathbf{W}}^i\|_2^2}$, $i = 1, ..., d$. By setting the derivative of Eq. 8 *w.r.t.* **b** to zero, we have:

$$2\mathbf{e}^T\mathbf{XW} + 2\mathbf{e}^T\mathbf{eb} - 2\mathbf{e}^T\mathbf{Y} = 0 \tag{9}$$

After a simple mathematical transformation, we have:

$$\mathbf{b} = \frac{1}{n}\mathbf{e}^T\mathbf{Y} - \frac{1}{n}\mathbf{e}^T\mathbf{XW}. \tag{10}$$

Similarly, by setting the derivative of Eq. 8 *w.r.t.* **p** to zero, we obtain optimal **p** as:

$$\mathbf{p} = \frac{1}{n}\mathbf{e}^T\mathbf{X} - \frac{1}{n}\mathbf{e}^T\mathbf{XS}. \tag{11}$$

By substituting Eqs. 10 and 11 into Eq. 8, we have:

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{S}} tr((\mathbf{Y}-\mathbf{XW}-\mathbf{e}(\tfrac{1}{n}\mathbf{e}^T\mathbf{Y}-\tfrac{1}{n}\mathbf{e}^T\mathbf{XW}))^T(\mathbf{Y}-\mathbf{XW}\\ &-\mathbf{e}(\tfrac{1}{n}\mathbf{e}^T\mathbf{Y}-\tfrac{1}{n}\mathbf{e}^T\mathbf{XW}))\\ &+\beta tr((\mathbf{X}-\mathbf{XS}-\mathbf{e}(\tfrac{1}{n}\mathbf{e}^T\mathbf{X}-\tfrac{1}{n}\mathbf{e}^T\mathbf{XS}))^T(\mathbf{X}-\mathbf{XS}\\ &-\mathbf{e}(\tfrac{1}{n}\mathbf{e}^T\mathbf{X}-\tfrac{1}{n}\mathbf{e}^T\mathbf{XS})))\\ &+\lambda tr(\tilde{\mathbf{W}}^T\mathbf{D}\tilde{\mathbf{W}}).\end{aligned} \tag{12}$$

Let $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{ee}^T \in \mathbb{R}^{n\times n}$, where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is an identity matrix, Eq. 12 can be rewritten as

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{S}} tr((\mathbf{HY}-\mathbf{HXW})^T(\mathbf{HY}-\mathbf{HXW})\\ &+\beta tr((\mathbf{HX}-\mathbf{HXS})^T(\mathbf{HX}-\mathbf{HXS}))\\ &+\lambda tr(\tilde{\mathbf{W}}^T\mathbf{D}\tilde{\mathbf{W}}).\end{aligned} \tag{13}$$

By taking the derivative of Eq. 13 *w.r.t.* **W** and setting it equal to zero, we obtain:

$$\mathbf{W} = (\mathbf{X}^T\mathbf{HX} + \lambda\mathbf{D})^{-1}\mathbf{X}^T\mathbf{HY}. \tag{14}$$

Similarly, by setting the derivative of Eq. 13 *w.r.t.* **S** and setting it equal to zero, we have:

$$\mathbf{S} = \left(\mathbf{X}^T\mathbf{HX} + \frac{\lambda}{\beta}\mathbf{D}\right)^{-1}\mathbf{X}^T\mathbf{HX}. \tag{15}$$

Next, we applied Algorithm 1 to solve the objective function in Eq. 7. Based on the theory of iteratively reweighted method in (Wipf and Nagarajan 2010), it can be proved that the value of our objective function of Eq. 7 monotonically decreases in each iteration until convergence using Algorithm 1. According to the iterative characteristics of Algorithm 1, the current **W** can be enhanced by the last optimal **S**, whose next optimization can be further improved by the updated optimal **W**. In this way, the proposed optimization method in Algorithm 1 helps obtain the optimal **W** and **S**, which finally ensure the output of class-discriminative features.

Algorithm 1 Pseudo code of solving Eq. 7

```
Input: $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{Y} \in \mathbb{R}^{n\times c}$, $\beta$, $\lambda$;
Output: $\mathbf{W}$, $\mathbf{S}$;
Initialize $t = 1$;
Initialize $\mathbf{W}(1)$ and $\mathbf{S}(1)$ as two random diagonal matrices /* $\mathbf{W}(t)$ (or $\mathbf{S}(t)$): the $t$-th iteration result of $\mathbf{W}$ (or $\mathbf{S}$) */;
$\tilde{\mathbf{W}}(1) = [\mathbf{W}(1), \mathbf{S}(1)]$ /* $\tilde{\mathbf{W}}(t)$: the $t$-th iteration result of $\tilde{\mathbf{W}}$ */;
repeat
    Update $\mathbf{W}(t+1)$ via Eq. 14;
    Update $\mathbf{b}(t+1)$ via Eq. 10;
    Update $\mathbf{S}(t+1)$ via Eq. 15;
    Update $\mathbf{p}(t+1)$ via Eq. 11;
    Update $\tilde{\mathbf{W}}(t+1)$ via $\tilde{\mathbf{W}} = [\mathbf{W}, \mathbf{S}]$;
    Update $\mathbf{D}(t+1)$ via $\mathbf{D}_{ii} = \frac{1}{2\|\tilde{\mathbf{W}}^i\|_2^2}$, $i = 1, ..., d$;
    $t = t+1$;
until Eq. 7 converges;
```

Experimental results

Experimental settings

In this work, we considered three binary classification tasks (e.g., AD vs. NC, MCI vs. NC, and pMCI vs. sMCI) and two multi-class classification tasks (e.g., AD vs. NC vs. MCI and AD vs. NC vs. pMCI vs. sMCI). To evaluate the performance of all competing methods, we referred to the metrics of classification accuracy, sensitivity, specificity, and Area Under the receiver operating characteristic Curve (AUC) for binary classification and referred only to accuracy for multi-class classification.

We used a 10-fold cross-validation technique for all methods. Specifically, we first randomly partitioned the whole dataset into 10 subsets. We then selected one subset for testing and used the remaining 9 subsets for training. We repeated the whole process 10 times to avoid possible bias during dataset partitioning for cross-validation. The final result was computed by averaging results from all experiments. For the model selection of our method, we applied a 5-fold inner cross-validation on the parameter spaces of $\lambda \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$ and $\beta \in \{10^{-5}, 10^{-4}, \ldots, 10^{1}\}$ in Eq. 7 and $C \in \{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$ in SVM. As for SVM, we used a LIBSVM.[9] The parameters that resulted in the best performance in the inner cross-validation were used in testing.

Competing methods

We selected the following methods for comparison.

- Original: This method doesn't involve a feature selection step, but uses all the features for classification.
- Fisher Score (FS) (Duda et al. 2012): A Fisher's criterion is used for feature selection. Specifically, we compute a Fisher's score for each feature individually, based on the way we sort the features in an ascending order.
- Laplacian Score (LS) (He et al. 2005): This conducts supervised feature selection with an assumption that data of the same class tends to distribute to each other in the feature space, while those of different classes are apart from each other. The importance of a feature is evaluated by its power of a Laplacian score.
- SELF-representation (SELF) (Zhu et al. 2015): as an unsupervised approach, this method finds weighting coefficients with which each feature can be represented by a linear combination of the features. It is worth noting that SELF is a special case of our method, with the objective function in Eq. 5.
- Multi-Modal Multi-Task (M3T) (Zhang and Shen 2012): This method includes two key steps: (1) using multi-task feature selection to determine a common subset of relevant features for multiple response variables (or multiple tasks) from each modality, and (2) a multi-kernel decision fusion to integrate the selected features from all modalities for prediction. It is worth noting that M3T is a special case of our method with the objective function, i.e., Eq. 1 on single-modality data.
- Sparse Joint Classification and Regression (SJCR) (Wang et al. 2011): This method uses a logistic loss function for label identification and a least square loss function for clinical scores estimation. It simultaneously learns these two functions with an $\ell_{2,1}$-norm regularizer for multi-task feature selection.

In our experiments, we used all feature selection methods separately (except Original) in order to conduct feature selection, and then used an SVM for three binary classification tasks and two multi-class classification tasks on the data with the selected features. Moreover, for fair comparison, we also conducted 5-fold inner cross-validation to conduct model selection for each competing method. Specifically, for eigen-value based methods, such as FS and LS, we determined their optimal features based on their respective eigne-values computed by the generalized eigen-decomposition method, according to (Duda et al. 2012; He et al. 2005). For achieving the best performance for SELF, SJCR and M3T, we optimized their sparsity parameter by cross-validating the values in the ranges of $\{10^{-3}, 10^{-2}..., 10^{3}\}$, $\{10^{-5}, 10^{-3}..., 10^{5}\}$ (as in (Wang et al. 2011)), and $\{10^{-5}, ..., 10^{2}\}$, respectively.

[9] Available at 'http://www.csie.ntu.edu.tw/cjlin/libsvm/'.

Table 2 Classification accuracy for AD vs. NC

	Accuracy	**Sensitivity**	**Specificity**	**AUC**
Original	0.781±0.52*	0.856±0.05*	0.725±0.14*	0.799±0.07*
FS	0.854±0.88*	0.862±0.16*	0.775±0.13*	0.852±0.06*
LS	0.852±0.22*	0.873±0.44*	0.758±0.19*	0.860±0.07*
SELF	0.844±0.22*	0.872±0.07*	0.782±0.19⋆	0.818±0.05*
M3T	0.875±0.92*	0.871±0.08*	0.788±0.37*	0.897±0.01⋆
SJCR	0.861±0.34*	0.853±0.09*	0.762±0.28⋆	0.883±0.05⋆
Proposed	**0.903**±**0.05**	**0.915**±**0.06**	**0.819**±**0.09**	**0.912**±**0.06**

Bold number in each column represents the best result

Binary classification results

We summarized the performances of the competing methods with three binary classification tasks in Table 2 (AD vs. NC), Table 3 (MCI vs. NC), and Table 4 (pMCI vs. sMCI).[10]

The proposed method outperformed all competing methods by improving the classification accuracies on average over three binary classification tasks by 8.70% (vs. Original), 4.10% (vs. FS), 4.87% (vs. LS), 4.40% (vs. SELF), 3.80% (vs. SJCR), and 3.73% (vs. M3T). Specifically, compared to the worst performed method of Original and the best performed method of M3T among the competing methods, our method enhanced the classification accuracy by 12.2% (AD vs. NC) and 4.6% (MCI vs. NC). Based on these results, we conclude that the proposed feature selection method helped enhance classification performances by selecting more class-discriminative features. It is noteworthy that all methods achieved the highest classification performance on AD vs. NC and the lowest classification performance on pMCI vs. sMCI.

All feature selection methods outperformed the method using full features for classification, i.e., Original, which implies the effectiveness of feature selection with respect to the high-dimension problem in the classification of neuroimaging data. For example, the classification accuracy of Original is lower than FS (as it achieved the lowest performance among the other competing methods) and M3T (as it achieved the best performance among the other competing methods) on average of as much as 3.83% and 4.97%, respectively, over three binary classification tasks.

In comparison between LS and SELF, which adopt, respectively, a task-oriented and a self-representation-oriented strategies for feature selection, there was no significant difference in classification accuracy. Meanwhile, the supervised feature selection methods (i.e., FS, M3T and SJCR) obtained slightly higher performances than LS and SELF, yet there are still no significant differences in accuracy. However, the proposed method, which adopts both the task-oriented and self-representation-oriented by integrating the supervised and unsupervised learning concepts in a unified framework, clearly outperformed all these methods with significance, less than 0.05 of p-value in a statistical test. In this regard, we argue the effectiveness of joint task-oriented and self-representation-oriented regularization.

Multi-class classification results

In clinical applications, there exist multiple stages in the spectrum of AD and NC, such as pMCI and sMCI, but previous literature mostly focused on binary classification tasks. In this work, we consider the practical applications to conduct two multi-class classification tasks (i.e., AD vs. NC vs. MCI and AD vs. NC vs. pMCI vs. sMCI). We summarized the performance of all methods in Table 5.

Similar to the binary classification results, we observed that the proposed method outperformed the competing methods for both 3-class and 4-class classification tasks. More specifically, in the 3-class classification, our method achieved a classification accuracy of 63.9% by improving 14.5% (vs. Original), 6.6% (vs. FS), 6.0% (vs. LS), 4.6% (vs. SELF), 3.2% (vs. M3T), and 3.6% (vs. SJCR), respectively. In the 4-class classification, our method produced the highest classification accuracy of 59.3% by improving by 11.1% (vs. Original), 8.0% (vs. FS), 6.6% (vs. LS), 5.7% (vs. SELF), 4.0% (vs. M3T), and 5.3% (vs. SJCR).

Compared to binary classification tasks, the classification accuracy in multi-class classification tasks is decreased by around 14.0% and 18.6% on average, respectively, for the 3-class classification and the 4-class classification. One possible reason is that the subtle structure difference between pMCI subjects and sMCI subjects (or between AD subjects and MCI subjects) makes the multi-class classification much more difficult. Another possible reason may be the

[10] In Tables 2, 3, and 4, the boldface denotes the maximum performance in each column. (⋆: Statistically significant from the proposed method with $p < 0.05$ and ∗: Statistically significant different from the proposed method with $p < 0.001$ on the paired-sample t-tests at 95% significance level between results of our method and all other competing methods).

Table 3 Classification accuracy of MCI vs. NC

	Accuracy	**Sensitivity**	**Specificity**	**AUC**
Original	0.651±0.28*	0.623±0.51*	0.694±0.04*	0.652±0.09*
FS	0.675±0.26*	0.634±0.02*	0.782±0.23*	0.720±0.01*
LS	0.667±0.38*	0.650±0.52*	0.785±0.12*	0.741±0.05*
SELF	0.689±0.05*	0.648±0.14*	0.821±0.12*	0.752±0.04*
M3T	0.676±0.23*	0.659±0.17⋆	0.822±0.25⋆	0.777±0.02⋆
SJCR	0.683±0.06*	0.651±0.02⋆	**0.854±0.23**	0.775±0.04⋆
Proposed	**0.722±0.08**	**0.688±0.06**	0.850±0.09	**0.790±0.02**

Bold number in each column represents the best result

Table 4 Classification accuracy of pMCI vs. sMCI

	Accuracy	**Sensitivity**	**Specificity**	**AUC**
Original	0.645±0.15*	0.623±0.52*	0.601±0.24*	0.700±0.10*
FS	0.686±0.29⋆	0.651±0.35*	0.647±0.34*	0.732±0.10*
LS	0.673±0.80*	0.648±0.75*	0.638±0.57*	0.738±0.10*
SELF	0.673±0.07*	0.670±0.25⋆	0.655±0.08*	0.734±0.09*
M3T	0.675±0.27*	0.678±0.14⋆	0.697±0.45⋆	0.775±0.02⋆
SJCR	0.680±0.12*	0.664±0.34⋆	0.668±0.62*	0.766±0.07⋆
Proposed	**0.713±0.11**	**0.681±0.04**	**0.755±0.15**	**0.781±0.10**

Bold number in each column represents the best result

Table 5 Classification accuracy of multi-class classification tasks

Method	AD vs. NC vs. MCI	AD vs. NC vs. pMCI vs. sMCI
Original	0.494±4.26*	0.482±4.81*
FS	0.573±3.19*	0.513±1.54*
LS	0.579±3.07*	0.527±2.61*
SELF	0.593±2.35*	0.536±1.29*
M3T	0.607±1.08⋆	0.553±2.47*
SJCR	0.603±2.32*	0.540±2.43*
Proposed	**0.639±1.38**	**0.593±1.54**

Bold number in each column represents the best result

Table 6 Top 10 selected ROIs and the Average Number (AN) of selected features by our method on five classification tasks

Method	**Top 10 Regions**	**AN**
AD vs. NC	30, 46, 48, 61, 69, 76, 80, 83, 84, 22	58.0
MCI vs. NC	30, 46, 48, 69, 76, 83, 22, 78, 64, 70	43.2
pMCI vs. sMCI	30, 46, 48, 67, 69, 80, 83, 84, 90, 58	53.0
AD vs. NC vs. MCI	22, 30, 46, 48, 61, 69, 76, 80, 83, 84	50.6
AD vs. NC vs. pMCI vs. sMCI	30, 46, 48, 69, 76, 80, 83, 84, 62, 61	35.3

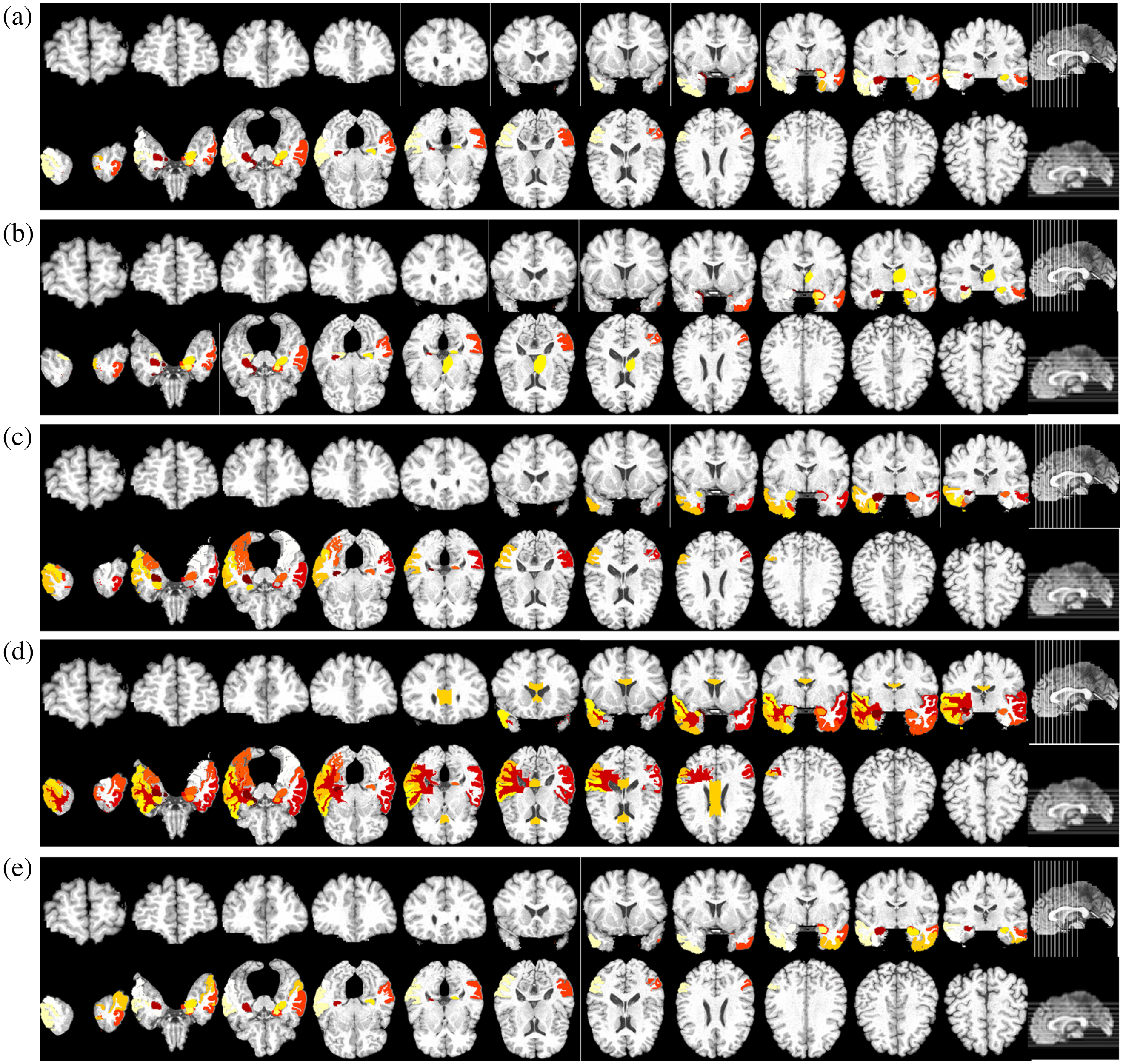


Fig. 2 Top 10 selected regions of the proposed method in five classification tasks: (**a**) AD vs. NC; (**b**) MCI vs. NC; (**c**) pMCI vs. sMCI; (**d**) AD vs. NC vs. MCI; (**e**) AD vs. NC vs. pMCI vs. sMCI

imbalance of training data among classes, i.e., 186 ADs vs. 226 NCs vs. 393 MCIs in 3-class classification and 186 ADs vs. 226 NCs vs. 118 pMCIs vs. 124 sMCIs in 4-class classification.

Most discriminative brain regions

We investigated the selected features to identify the potential biomarkers in AD diagnosis. We listed the most frequently selected ROIs in Table 6 and also visualized them in Fig. 2. As presented in Table 6, on average, our method selected 58.0 (AD vs. NC), 43.2 (MCI vs. NC), 53.0 (pMCI vs. sMCI), 50.6 (AD vs. MCI vs. NC), and 35.3 (AD vs. pMCI vs. sMCI vs. NC) numbers of features, out of 93 features (*a.k.a.,* ROIs), respectively.

From Table 6, we can see that the commonly selected regions in all five different classification tasks were uncus right (22),[11] hippocampal formation right (30), uncus left (46), middle temporal gyrus left (48), hippocampal formation

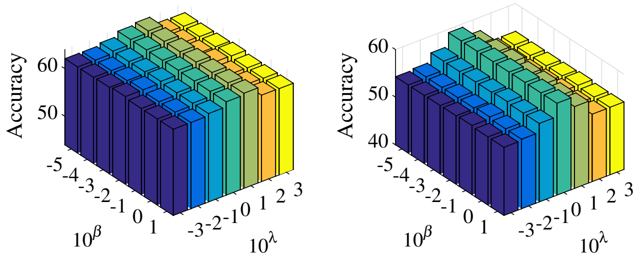
[11] The number in the parentheses represents an index of an ROI. Please refer to Table 7 for the full name of the respective ROIs.

Table 7 The names of the selected ROIs in this work

ID	ROI Name	ID	ROI Name
22	uncus right	30	hippocampal formation right
43	temporal lobe WM right	45	superior temporal gyrus left
46	uncus left	48	middle temporal gyrus left
58	perirhinal cortex right	61	perirhinal cortex left
62	inferior temporal gyrus left	64	entorhinal cortex left
67	lateral occipitotemporal gyrus right	69	hippocampal formation left
70	thalamus left	76	amygdala left
78	parahippocampal gyrus right	80	middle temporal gyrus right
82	corpus callosum	83	amygdala right
84	inferior temporal gyrus right	85	superior temporal gyrus right
90	lateral occipitotemporal gyrus left		

left (69), amygdala left (76), middle temporal gyrus right (80), and amygdala right (83). Moreover, these discriminative brain regions have been pointed out in previous literature on binary classification (Zhang and Shen 2012) and have been also shown to be highly related to AD or related dementia (e.g., MCI) in clinical diagnosis (Chételat et al. 2005; Convit et al. 2000; Fox and Schott 2004; Misra et al. 2009). In this regard, we can say that these regions could be potential biomarkers for AD or MCI diagnosis.

It is worth noting that: (1) most of the competing methods in our experiments selected the ROIs listed in Table 6 as class-discriminative features. (2) Even though most of the methods selected similar ROIs as the top brain regions, our method selected them with the highest frequency.



(a) RMSE (b) Top 10 selected SNPs

Fig. 4 Classification accuracy of the proposed objective function with varied values of β and γ in two multi-class classification tasks

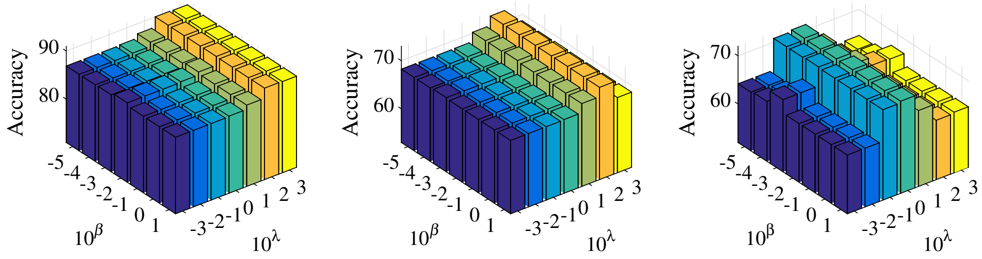
Sensitivity with Different Parameter Values

Figures 3 and 4 present the sensitivity in performance according to the changes in the values of β and γ in Eq. 7. The results show that our proposed method is sensitive to the parameters within only a small range. Specifically, the best parameter combination was always found since 1) the magnitude of $\|\mathbf{X}-\mathbf{XS}-\mathbf{ep}\|_F^2$ was almost approached to the magnitude of the data fitting term (i.e., $\|\mathbf{Y}-\mathbf{XW}-\mathbf{eb}\|_F^2$) (by tuning the values of β); and 2) the large values of γ caused the matrices of $\mathbf{W}$ and $\mathbf{S}$ to be sparse. This indicates the importance of the penalty term $\|\mathbf{X}-\mathbf{XS}-\mathbf{ep}\|_F^2$; and also the sparse constraint on the proposed objective function plays an important role for selecting informative features.

Discussion and conclusion

In this paper, we focused on the *high feature-dimension* problem for both binary classification and multi-class classification in AD diagnosis. Specifically, we proposed a novel feature selection method by integrating a task-oriented regularization in supervised learning and a self-representation-oriented regularization in unsupervised learning in a linear regression framework. Our experimental results on the ADNI dataset with MRI imaging data validated the effectiveness of the proposed method by enhancing classification

Fig. 3 Classification accuracy of the proposed objective function with varied values of β and γ in three binary classification tasks



(a) RMSE (b) SNPs (c) ROIs

accuracies in both binary classification and multi-class classification problems.

In the study of high-dimensional neuroimaging data, multi-modality data have been demonstrated to improve performance of AD diagnosis due to the beneficiary of complementary information from different modalities (Zhang et al. 2011; Zhang and Shen 2012; Zhu et al. 2014c). However, the proposed model in Eq. 7 cannot be directly applied for multi-modality data, as literature showed that multi-modality neuroimaging data may not share the same sparsity across modalities (Zhu et al. 2014c), since the last term in Eq. 7 could push all modalities to select the same features. In our future work, we will focus on extending the proposed model to its multi-modality version for further improving the performance of AD diagnosis via mapping multiple subspaces spanned by multi-modality data into a common space.

Acknowledgments This work was supported in part by NIH grants (EB006733, EB008374, MH100217, MH108914, AG041721, AG049371, AG042599, AG053867, EB022880, MH110274), the Nation Natural Science Foundation of China (Grant No: 61573270), the Brain Research Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT & Future Planning (NRF-2014M3C7A1046050), the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning, and by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (No. 2017-0-00451).

Compliance with Ethical Standards

Conflict of interests Xiaofeng Zhu, Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen declare that they have no conflict of interest.

Ethical approval We confirm that this article does not contain any studies with either human participants or animals performed by any of the authors.

Informed Consent We confirm that informed consent was obtained from all individual participants included in the study.

References

Chételat, G., Eustache, F., Viader, F., Sayette, V.D.L., Pélerin, A., Mézenge, F., Hannequin, D., Dupuy, B., Baron, J.-C., & Desgranges, B. (2005). FDG-PET measurement is more accurate than neuropsychological assessments to predict global cognitive deterioration in patients with mild cognitive impairment. *Neurocase*, *11*(1), 14–25.

Convit, A., De Asis, J., De Leon, M., Tarshish, C., De Santi, S., & Rusinek, H. (2000). Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimers disease. *Neurobiology of Aging*, *21*(1), 19–26.

Duchesne, S., Caroli, A., Geroldi, C., Collins, D.L., & Frisoni, G.B. (2009). Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. *NeuroImage*, *47*(4), 1363–1370.

Duda, R.O., Hart, P.E., & Stork, D.G. (2012). Pattern classification. Wiley.

Elhamifar, E., & Vidal, R. (2009). Sparse subspace clustering. In *CVPR* (pp. 2790–2797).

Fox, N.C., & Schott, J.M. (2004). Imaging cerebral atrophy: Normal ageing to Alzheimer's disease. *The Lancet*, *363*(9406), 392–394.

Greicius, M.D., Srivastava, G., Reiss, A.L., & Menon, V. (2004). Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(13), 4637–4642.

Guo, X., Wang, Z., Li, K., Li, Z., Qi, Z., Jin, Z., Yao, L., & Chen, K. (2010). Voxel-based assessment of gray and white matter volumes in Alzheimer's disease. *Neuroscience Letters*, *468*(2), 146–150.

He, X., Cai, D., & Niyogi, P. (2005). Laplacian score for feature selection. In *NIPS* (pp. 507–514).

Kabani, N.J. (1998). 3D anatomical atlas of the human brain. *NeuroImage*, *7*, 0700–0717.

Kim, H.J., Adluru, N., Bendlin, B.B., Johnson, S.C., Vemuri, B.C., & Singh, V. (2014). Canonical correlation analysis on riemannian manifolds and its applications. In *European Conference on Computer Vision* (pp. 251–267).

Lee, G.-Y., Kim, J., Kim, J.H., Kim, K., & Seong, J.-K. (2014). Online learning for classification of Alzheimer disease based on cortical thickness and hippocampal shape analysis. *Healthcare Informatics Research*, *20*(1), 61–68.

Liu, G., Lin, Z., & Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In *ICML* (pp. 663–670).

López-de-Ipiña, K., Solé-Casals, J., Eguiraun, H., Alonso, J.B., Travieso, C.M., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P., & Beitia, B. (2015). Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach. *Computer Speech and Language*, *30*(1), 43–60.

Lorenzi, M., Simpson, I.J., Mendelson, A.F., Vos, S.B., Cardoso, M.J., Modat, M., Schott, J.M., & Ourselin, S. (2016). Multimodal image analysis in Alzheimers disease via statistical modelling of non-local intensity correlations. Scientific Reports, 6.

Misra, C., Fan, Y., & Davatzikos, C. (2009). 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), 1415–1422.

Mwangi, B., Tian, T.S., & Soares, J.C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, *12*(2), 229–244.

Shen, D., & Davatzikos, C. (2002). HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging*, *21*(11), 1421–1439.

Sled, J.G., Zijdenbos, A.P., & Evans, A.C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, *17*(1), 87–97.

Spedding, A.L., Di Fatta, G., & Saddy, J.D. (2015). An LDA and probability-based classifier for the diagnosis of Alzheimer's Disease from structural MRI. In *BIBM* (pp. 1404–1411).

Suk, H.-I., Lee, S.-W., & Shen, D. (2015a). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Structure and Function*, *220*(2), 841–859.

Suk, H.-I., Wee, C.-Y., Lee, S.-W., & Shen, D. (2015b). Supervised discriminative group sparse representation for mild cognitive impairment diagnosis. *Neuroinformatics*, *13*(3), 277–295.

Thung, K.-H., Wee, C.-Y., Yap, P.-T., & Shen D. (2014). Neurodegenerative disease diagnosis using incomplete multimodality data via matrix shrinkage and completion. *NeuroImage*, *91*, 386–400.

Thung, K.-H., Wee, C.-Y., Yap, P.-T., & Shen, D. (2015). Identification of progressive mild cognitive impairment patients using incomplete longitudinal MRI scans. *Brain Structure and Function* 1–17.

Wang, H., Nie, F., Huang, H., Risacher, S., Saykin, A.J., & Shen, L. (2011). Identifying AD-sensitive and cognitionrelevant imaging biomarkers via joint classification and regression. In *MICCAI* (pp. 115–123).

Wang, Y., Fan, Y., Bhatt, P., & Davatzikos, C. (2010). High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. *NeuroImage*, *50*(4), 1519–1535.

Wipf, D., & Nagarajan, S. (2010). Iterative reweighted and methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing*, *4*(2), 317–329.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, *68*(1), 49–67.

Zhang, D., & Shen, D. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, *59*(2), 895–907.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., & Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, *55*(3), 856–867.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57.

Zhu, P., Zuo, W., Zhang, L., Hu, Q., & Shiu, S.C.K. (2015). Unsupervised feature selection by regularized selfrepresentation. *Pattern Recognition*, *48*(2), 438–446.

Zhu, X., Suk, H.-I., & Shen, D. (2014a). Matrix-Similarity Based Loss Function and Feature Selection for Alzheimer's Disease Diagnosis. In *CVPR* (pp. 3089–3096).

Zhu, X., Suk, H.-I., & Shen, D. (2014b). Multi-modality canonical feature selection for Alzheimers disease diagnosis. In *MICCAI* (pp. 162–169).

Zhu, X., Suk, H.-I., & Shen, D. (2014c). A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. *NeuroImage*, *14*(0), 1–30.

Zhu, X., Suk, H.-I., & Shen, D. (2014d). A novel multi-relation regularization method for regression and classification in AD diagnosis. In *MICCAI* (pp. 401–408).

Zhu, X., Suk, H.-I., Lee, S.-W., & Shen, D. (2015a). Canonical feature selection for joint regression and multi-class identification in Alzheimers disease diagnosis. *Brain Imaging and Behavior*, pp. 1–11.

Zhu, X., Suk, H.-I., Wang, L., Lee, S.-W., Shen, D., Initiative, A.D.N., et al. (2015b). A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. *Medical Image Analysis, 38*, 205–214.