Medical Image Analysis 000 (2015) 1–10

Contents lists available at ScienceDirect

Medical Image Analysis

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

A novel relational regularization feature selection method for joint regression and classification in AD diagnosis

Xiaofeng Zhu[a], Heung-Il Suk[b], Li Wang[a], Seong-Whan Lee[b,*], Dinggang Shen[a,b], Alzheimer's Disease Neuroimaging Initiative

[a] Department of Radiology and BRIC, The University of North Carolina at Chapel Hill, USA
[b] Department of Brain and Cognitive Engineering, Korea University, Republic of Korea

ARTICLE INFO

Article history:
Received 9 November 2014
Revised 10 June 2015
Accepted 21 October 2015
Available online xxx

Keywords:
Alzheimer's disease
Feature selection
Sparse coding
Manifold learning
MCI conversion

ABSTRACT

In this paper, we focus on joint regression and classification for Alzheimer's disease diagnosis and propose a new feature selection method by embedding the relational information inherent in the observations into a sparse multi-task learning framework. Specifically, the relational information includes three kinds of relationships (such as feature-feature relation, response–response relation, and sample-sample relation), for preserving three kinds of the similarity, such as for the features, the response variables, and the samples, respectively. To conduct feature selection, we first formulate the objective function by imposing these three relational characteristics along with an $\ell_{2,1}$-norm regularization term, and further propose a computationally efficient algorithm to optimize the proposed objective function. With the dimension-reduced data, we train two support vector regression models to predict the clinical scores of ADAS-Cog and MMSE, respectively, and also a support vector classification model to determine the clinical label. We conducted extensive experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset to validate the effectiveness of the proposed method. Our experimental results showed the efficacy of the proposed method in enhancing the performances of both clinical scores prediction and disease status identification, compared to the state-of-the-art methods.

1. Introduction

Alzheimer's Disease (AD) is characterized as a genetically complex and irreversible neurodegenerative disorder and often found in persons aged over 65. Recent studies have shown that there are about 26.6 million AD patients worldwide, and 1 out of 85 people will be affected by AD by 2050 (Brookmeyer et al., 2007; Zhang et al., 2012; Zhou et al., 2011; Zhu et al., 2014a; 2014b). Thus, there have been great interests for early diagnosis of AD and its prodromal stage, Mild Cognitive Impairment (MCI).

It has been shown that the neuroimaging tools, including Magnetic Resonance Imaging (MRI) (Fjell et al., 2010), Positron Emission Tomography (PET) (Wee et al., 2013; Morris et al., 2001), and functional MRI (Suk et al., 2013), help understand the neurodegenerative process in the progression of AD. Furthermore, machine learning methods can effectively handle complex patterns in the observed subjects for either identifying clinical labels, such as AD, MCI, and Normal Control (NC) (Cheng et al., 2013; Franke et al., 2010; Walhovd et al., 2010), or regressing the clinical scores, such as Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) and Mini-Mental State Examination (MMSE) (McEvoy et al., 2009; Wee et al., 2012).

In computer-aided AD diagnosis, the available sample size is usually small, but the feature dimensionality is high. For example, the sample size used in (Jie et al., 2013) was as small as 99, while the feature dimensionality (including both MRI and PET features) was hundreds or even thousands. The small sample size makes it difficult to build an effective model, and the high-dimensional data could lead to an overfitting problem although the number of intrinsic features may be very low (Weinberger et al., 2004; Suk et al., 2014; Zhu et al., 2015c; 2015b). To this end, researchers predefined the disease-related features and used the low-dimensional feature vector for disease identification. For example, Wang et al. (2011) considered the brain areas of medial temporal lobe structures, medial and lateral parietal, as well as prefrontal cortical areas, and showed that these areas were useful to predict most memory scores and classify AD from NC subjects.

However, to further enhance diagnostic accuracy and better understand the disease-related brain atrophies, it's necessary to select

* Corresponding author at: Department of Radiology and BRIC, The University of North Carolina at Chapel Hill, USA; and and Department of Brain and Cognitive Engineering, Korea University, Republic of Korea.
E-mail addresses: sw.lee@korea.ac.kr (S.-W. Lee), dgshen@med.unc.edu (D. Shen).

http://dx.doi.org/10.1016/j.media.2015.10.008
1361-8415/

features in a data-driven manner. It has been shown that the feature selection helps overcome both problems of high dimensionality and small sample size by removing uninformative features. Among various feature selection techniques, manifold learning methods has been successfully used in either regression or classification (Cho et al., 2012; Cuingnet et al., 2011; Liu et al., 2014; Zhang and Shen, 2012; Zhang et al., 2011; Suk et al., 2015). For example, Cho et al. (2012) adopted a manifold harmonic transformation method on the cortical thickness data. Meanwhile, while most of the previous studies focused on separately identifying brain disease and estimating clinical scores (Jie et al., 2013; Liu et al., 2014; Suk and Shen, 2013), there also have been some efforts to tackle both tasks simultaneously in a unified framework. For example, Zhang and Shen (2012) proposed a feature selection method for simultaneous disease diagnosis and clinical scores prediction, and achieved promising results. However, to our best knowledge, the previous manifold-based feature selection methods considered only the manifold of the samples, not manifold of either the features or the response variables.

For better understanding of the underlying mechanism of AD, our interest in this paper is to predict both clinical scores and disease status jointly, which we call as Joint Regression and Classification (JRC) problem. In particular, we devise new regularization terms to reflect the relational information inherent in the observations and then combine them with an $\ell_{2,1}$-norm regularization term within a multi-task learning framework for joint sparse feature selection in the JRC problem. The rationale for the proposed regularization method is as follows: (1) If some features are related to each other, then the same or similar relation is expected to be preserved between the respective weight coefficients. (2) Due to the algebraic operation in the least square regression, *i.e.,* matrix multiplication, the weight coefficients are linked to the response variables via regressors, *i.e.,* feature vectors in our work. Therefore, it is meaningful to impose the relation between a pair of weight coefficients to be similar to the relation between the respective pair of target response variables. (3) As considered in many manifold learning methods (Belkin et al., 2006; Fan et al., 2008; Zhu et al., 2011; 2013b; 2013c), if a pair of samples are similar to each other, then their respective response values should be also similar to each other. By imposing these three relational characteristics along with the $\ell_{2,1}$-norm regularization term on the weight coefficients, we formulate a new objective function to conduct feature selection and further solve it with a new computationally efficient optimization algorithm. Then, we can select effective features to build a classifier for clinical label identification and two regression models for ADAS-Cog and MMSE scores prediction, respectively.

2. Image preprocessing

In this work, we used the publicly available ADNI dataset for performance evaluation.

2.1. Subjects

We selected the subjects satisfying the following general inclusion/exclusion criteria[1]: (1) The MMSE score of each NC is between 24 and 30. Their Clinical Dementia Rating (CDR) is of 0. Moreover, the NC is non-depressed, non MCI, and non-demented. (2) The MMSE score of each 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. (3) The MMSE score of each Mild AD subject is between 20 and 26, with the CDR of 0.5 or 1.0.

In this paper, we use baseline MRI and PET obtained from 202 subjects including 51 AD subjects, 52 NC subjects, and 99 MCI subjects. Moreover, 99 MCI subjects include 43 MCI-C and 56 MCI-NC[2]. The detailed demographic information is summarized in Table 1. For reference, we presented sample slices of MRI and PET for one typical subject belonging each class (AD, MCI, and NC) in Fig. 1.

Table 1
Demographic information of the subjects. (MCI-C: MCI Converters; MCI-NC: MCI Non-Converters).

	AD	NC	MCI-C	MCI-NC
Female/male	18/33	18/34	15/28	17/39
Age	75.2 ± 7.4	75.3 ± 5.2	75.8 ± 6.8	74.8 ± 7.1
Education	14.7 ± 3.6	15.8 ± 3.2	16.1 ± 2.6	15.8 ± 3.2
MMSE	23.8 ± 2.0	29.0 ± 1.2	26.6 ± 1.7	28.4 ± 1.7
ADAS-Cog	18.3 ± 6.0	12.1 ± 3.8	12.9 ± 3.9	8.03 ± 3.8

2.2. Image processing

We downloaded raw Digital Imaging and COmmunications in Medicine (DICOM) MRI scans from the ADNI website[3]. All structural MR images used in this work were acquired from 1.5T scanners. Data were collected across a variety of scanners with protocols individualized for each scanner. Moreover, these MR images were already reviewed for quality, and automatically corrected for spatial distortion caused by gradient nonlinearity and B1 field inhomogeneity. Moreover, PET images were acquired 30–60 min post Fluoro-DeoxyGlucose (FDG) injection. They were then averaged, spatially aligned, interpolated to a standard voxel size, intensity normalized, and smoothed to a common resolution of 8*mm* full width at half maximum.

The image processing for all MR and PET images was conducted by following the same procedures in Zhang and Shen (2012). Specifically, we first performed anterior commissure-posterior commissure correction using MIPAV software[4] for all images, and used the N3 algorithm (Sled et al., 1998) to correct the intensity inhomogeneity. Second, we extracted a brain on all structural MR images using a robust skull-stripping method (Wang et al., 2013), followed by manual edition and intensity inhomogeneity correction. After removal of cerebellum based on registration (Tang et al., 2009; Wu et al., 2011; Xue et al., 2006) and also intensity inhomogeneity correction by repeating N3 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 CSF. Next, we used HAMMER (Shen and Davatzikos, 2002) to register the template into subject specific space for preserving local image volume of each subjects. We then obtained the Region-Of-Interest (ROI) labeled images using the Jacob template, which dissects a brain into 93 ROIs (Kabani, 1998). For each of all 93 ROIs in the labeled image of a subject, we computed the GM tissue volumes in ROIs by integrating the GM segmentation result of the subject. For each subject, we aligned the PET images to their respective MR T1 images using affine registration and then computed the average intensity of each ROI. Therefore, for each subject, we obtained 93 features for MRI and 93 features for PET.

3. Method

3.1. 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

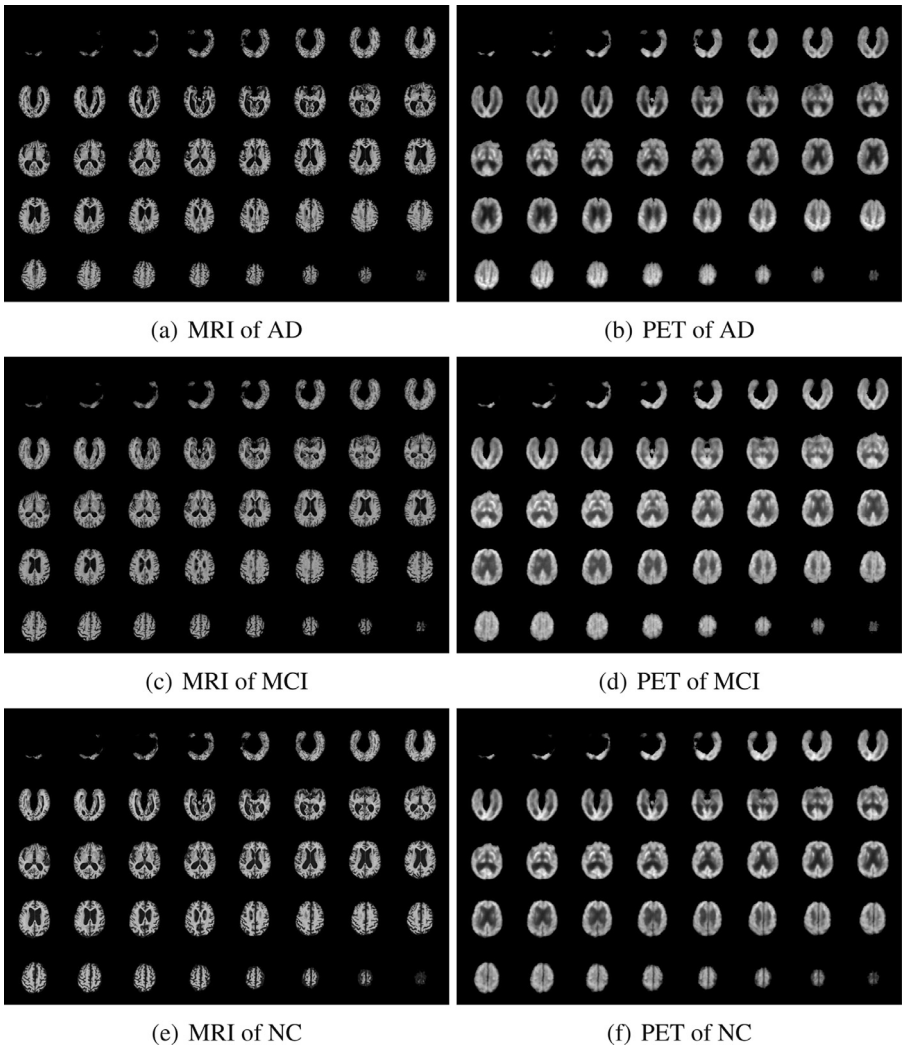
[1] Please refer to 'http://adni.loni.usc.edu/' for up-to-date information.

[2] Here, MCI-C and MCI-NC denote, respectively, those who progressed to AD in 18 months and those who didn't.

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

[4] http://mipav.cit.nih.gov/clickwrap.php.

Please cite this article as: X. Zhu et al., A novel relational regularization feature selection method for joint regression and classification in AD diagnosis, Medical Image Analysis (2015), http://dx.doi.org/10.1016/j.media.2015.10.008



(a) MRI of AD (b) PET of AD

(c) MRI of MCI (d) PET of MCI

(e) MRI of NC (f) PET of NC

Fig. 1. Example slices of MRI (left column) and PET (right column) for subjects belonging to different classes.

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.

3.2. Relational regularization

Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{Y} \in \mathbb{R}^{n\times c}$ denote, respectively, the d neuroimaging features and c clinical response values of n subjects or samples[5]. In this work, we assume that the response values of clinical scores and clinical label[6] can be represented by a linear combination of the features. Then, the problems of regressing clinical scores and determining class label can be formulated by a least square regression model as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{W}) &= \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 \\ &= \|\mathbf{Y} - \hat{\mathbf{Y}}\|_F^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{c}(y_{ij} - \hat{y}_{ij})^2 \end{aligned} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d\times c}$ is a weight coefficient matrix and $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{W}$. While the least square regression model has been successfully used in many applications, it is shown that the solution is often overfitted to the dataset with small samples and high-dimensional features in its original form, especially, in the field of neuroimaging analysis. To this end, a variety of its variants using different types of regularization terms have been suggested to circumvent the overfitting problem and find a more generalized solution (Suk et al., 2013; Yuan and Lin, 2006; Zhang and Shen, 2012), which can be mathematically simplified as follows:

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) + \mathcal{R}(\mathbf{W}) \tag{2}$$

where $\mathcal{R}(\mathbf{W})$ denotes a set of regularization terms.

From a machine learning point of view, a well-defined regularization term can produce a generalized solution to the objective function, and thus results in a better performance for the final goal. In this paper, we devise novel regularization terms that effectively utilize various pieces of information inherent in the observations.

Note that since, in this work, we extract features from ROIs, which are structurally or functionally related to each other, it is natural to expect that there exist relations among features. Meanwhile, if two features are highly related to each other, then it is reasonable to have the respective weight coefficients also related. However, to the best of our knowledge, none of the previous representation (or regression) methods in the literature considered and guaranteed this

[5] In this work, we have one sample per subject.
[6] In this paper, we represented the class label with 0–1 encoding.

Please cite this article as: X. Zhu et al., A novel relational regularization feature selection method for joint regression and classification in AD diagnosis, Medical Image Analysis (2015), http://dx.doi.org/10.1016/j.media.2015.10.008

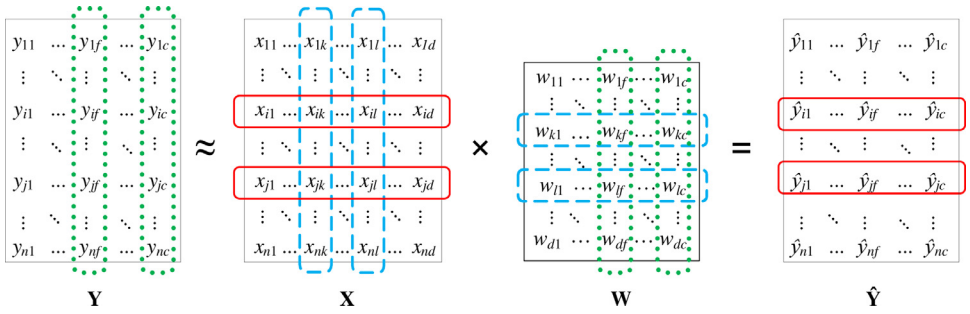


Fig. 2. An illustration of the relational information that can be obtained from the observations. The red solid rectangles, the blue dash rectangles, and the green dotted rectangles denote, respectively, the 'sample-sample' relation, 'feature-feature' relation and 'response-response' relation.

characteristic in their solutions. To this end, we devise a regularization term with the assumption that, if some features, *e.g.,* $\mathbf{x}_i$ and $\mathbf{x}_j$ in the blue dash rectangles of Fig. 2, are involved in regressing the response variables and are also related to each other, their corresponding weight coefficients (*i.e.,* $\mathbf{w}_i$ and $\mathbf{w}_j$) should have the same or similar relation since the *i*-feature $\mathbf{x}_i$ in $\mathbf{X}$ corresponds to the *i*th row $\mathbf{w}_i$ in $\mathbf{W}$ in our regression framework. We call this relation as the '*feature-feature*' relation in this work. To utilize the '*feature–feature*' relation, we penalize the loss function with the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ (*i.e.,* m_{ij}) on $\|\mathbf{w}_i - \mathbf{w}_i\|_2^2$. Specifically, we impose the relation between columns in $\mathbf{X}$to be reflected in the relation between the respective rows in $\mathbf{W}$ by defining the following embedding function:

$$\mathcal{R}_1(\mathbf{W}) = \frac{1}{2}\sum_{i,j}^{d} m_{ij}\|\mathbf{w}^i - \mathbf{w}^j\|_2^2 \tag{3}$$

where m_{ij} denotes an element in the feature similarity matrix $\mathbf{M} = [m_{ij}] \in \mathbb{R}^{d\times d}$which encodes the relation between features in the samples. With respect to the similarity measure between vectors of $\mathbf{a}$ and $\mathbf{b}$, throughout this paper, we first use a radial basis function kernel as defined as follows:

$$f(\mathbf{a}, \mathbf{b}) = exp\left(-\frac{\|\mathbf{a}-\mathbf{b}\|_2^2}{2\sigma^2}\right) \tag{4}$$

where σ denotes a kernel width. As for the similarity matrix $\mathbf{M}$, we first construct a data adjacency graph by regarding each sample as a node and using *k* nearest neighbors along with a heat kernel function defined in Eq. (4) to compute the edge weights, *i.e.,* similarities. For example, if a sample $\mathbf{x}_j$ is selected as one of the *k* nearest neighbors of a sample $\mathbf{x}_i$, then the similarity m_{ij} between these two samples or nodes is set to the value of $f(\mathbf{x}_i, \mathbf{x}_j)$; otherwise, their similarity is set to zero, *i.e.,* $m_{ij} = 0$.

In the meantime, given a feature vector $\mathbf{x}^i$, in our joint regression and classification framework, we use a different set of weight coefficients to regress the elements in the response vector $\mathbf{y}^i$. In other words, the elements of each column in $\mathbf{W}$ are linked to the elements of each column in $\mathbf{Y}$ via feature vectors. By taking this mathematical property into account, we further impose the relation between column vectors in $\mathbf{W}$ to be similar to the relation between the respective target response variables (*i.e.,* respective column vectors) in $\mathbf{Y}$, which is called as '*response-response*' relation as defined below:

$$\mathcal{R}_2(\mathbf{W}) = \frac{1}{2}\sum_{i,j}^{c} g_{ij}\|\mathbf{w}_i - \mathbf{w}_j\|_2^2 \tag{5}$$

where g_{ij} denotes an element in the matrix $\mathbf{G} = [g_{ij}] \in \mathbb{R}^{c\times c}$which represents the similarity between every pair of target response variables (*i.e.,* every pair of column vectors).

We also utilize the relational information between samples, called as '*sample-sample*' relation. That is, if samples are similar to each other, then their respective response values should be also similar to each other. To this end, we define a regularization term as follows:

$$\mathcal{R}_3(\mathbf{W}) = \frac{1}{2}\sum_{i,j}^{n} s_{ij}\|\hat{\mathbf{y}}^i - \hat{\mathbf{y}}^j\|_2^2 \tag{6}$$

where s_{ij} is an element in the matrix $\mathbf{S} = [s_{ij}] \in \mathbb{R}^{n\times n}$which measures the similarity between every pair of samples. We should note that this kind of sample–sample relation has been successfully used in many manifold learning methods (Belkin et al., 2006; Zhu et al., 2013b; 2013c). The elements of the matrices $\mathbf{G}$ and $\mathbf{S}$ can be computed similarly as in the computation of $\mathbf{M}$ as described above.

We argue that the simultaneous consideration of these newly devised regularization terms, *i.e.,* feature–feature relation, sample–sample relation, and response–response relation, can effectively reflect the relational information inherent in observations in finding an optimal solution. Fig. 2 illustrates these relational regularizations in a matrix form. Regarding feature selection, we believe that due to the underlying brain mechanisms that influence both the clinical scores and a clinical label, *i.e.,* response variables, if one feature plays a role in predicting one response variable, then it also devotes to the prediction of the other response variables. To this end, we further impose to use the same features across the tasks of clinical scores and clinical label prediction. Mathematically, this can be implemented by an $\ell_{2,1}$-norm regularization term on $\mathbf{W}$, *i.e.,* $\|\mathbf{W}\|_{2,1} = \sum_i \|\mathbf{w}^i\|_2$. Concretely, $\|\mathbf{w}^i\|_2$, the ℓ_2-norm of the *i*th row vector in $\mathbf{W}$, is equally imposed on the *i*th feature across different tasks, which thus forces the coefficients that weight the *i*-th feature for different tasks to be grouped together. Earlier, Zhang and Shen (2012)considered the same regularization term in their multi-task learning and validated its efficacy in AD/MCI diagnosis.

Finally, our objective function is formulated as follows:

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) + \alpha_1\mathcal{R}_1(\mathbf{W}) + \alpha_2\mathcal{R}_2(\mathbf{W}) + \alpha_3\mathcal{R}_3(\mathbf{W}) + \lambda\|\mathbf{W}\|_{2,1} \tag{7}$$

where α_1, α_2, α_3, and λ denote control parameters of the respective regularization terms, respectively. It is noteworthy that unlike the previous regularization methods such as local linear embedding (Roweis and Saul, 2000), locality preserving projection (He et al., 2005; Zhu et al., 2013a; 2014c), and high-order graph matching (Liu et al., 2013) that focused on the sample similarities by imposing nearby samples to be still nearby in the transformed space, the proposed method utilizes richer information obtained from the observations for finding the optimal weight coefficients $\mathbf{W}$. The matrices $\mathbf{X}$ and $\mathbf{Y}$ are used to obtain the similarities, where $\mathbf{X}$ and $\mathbf{Y}$ are composed of MRI/PET features and target values, respectively. According to the previous work in Zhu et al. (2014a), theoretically the loss function in Eq. (1) can be designed to expect that the predictions of the model should be correlated for the similar subjects. But, in practice, it is not guaranteed due to unexpected noises in features. In this regard, we explicitly impose such correlational characteristic (*e.g.,* the proposed three kinds of relations) in the final objective function. Thus, it is

Please cite this article as: X. Zhu et al., A novel relational regularization feature selection method for joint regression and classification in AD diagnosis, Medical Image Analysis (2015), http://dx.doi.org/10.1016/j.media.2015.10.008

expected that the proposed method can find a generalizable solution robust to noise or outlier.

3.3. Optimization

With respect to the optimization of parameters $\mathbf{W}$, due to the use of the similarity weights of m_{ij} in Eq. (3), g_{ij} in Eq. (5), and s_{ij} in Eq. (6), it is beneficiary to transform the respective regularization terms to the trace forms using Laplacian matrices (Belkin et al., 2006; Zhu et al., 2012; 2015a). Let $\mathbf{H^M}$, $\mathbf{H^G}$, and $\mathbf{H^S}$, respectively, be diagonal matrices with their diagonal elements being the column-wise or row-wise sum of the similarity weight matrices of **M, G**, and **S**, *i.e.*, $h^{\mathbf{M}}_{ii} = \sum_{j=1}^{d} m_{ij}, h^{\mathbf{G}}_{ii} = \sum_{j=1}^{c} g_{ij}$, and $h^{\mathbf{S}}_{ii} = \sum_{j=1}^{n} s_{ij}$. The regularization terms can be rewritten as follows:

$$\mathcal{R}_1(\mathbf{W}) = tr(\mathbf{W}^T\mathbf{L_M}\mathbf{W}) \tag{8}$$

$$\mathcal{R}_2(\mathbf{W}) = tr(\mathbf{W}\mathbf{L_G}\mathbf{W}^T) \tag{9}$$

$$\mathcal{R}_3(\mathbf{W}) = tr((\mathbf{X}\mathbf{W})^T\mathbf{L_S}(\mathbf{X}\mathbf{W})) \tag{10}$$

where $\mathbf{L_M} = \mathbf{H^M} - \mathbf{M}, \mathbf{L_G} = \mathbf{H^G} - \mathbf{G}$, and $\mathbf{L_S} = \mathbf{H^S} - \mathbf{S}$, which are called *Laplacian* matrices. Then our objective function in Eq. (7) can be rewritten as follows:

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) + \alpha_1 tr(\mathbf{W}^T\mathbf{L_M}\mathbf{W}) + \alpha_2 tr(\mathbf{W}\mathbf{L_G}\mathbf{W}^T) + \alpha_3 tr((\mathbf{X}\mathbf{W})^T\mathbf{L_S}(\mathbf{X}\mathbf{W})) + \lambda\|\mathbf{W}\|_{2,1}. \tag{11}$$

Note that Eq. (11) is a convex but non-smooth function. By setting the derivative of the objective function in Eq. (11) with respect to $\mathbf{W}$ to zero, we can obtain the form of

$$\mathbf{A}\mathbf{W} + \mathbf{W}\mathbf{B} = \mathbf{Z} \tag{12}$$

where $\mathbf{A} = (\mathbf{X}^T\mathbf{X} + \alpha_1\mathbf{L_M} + \alpha_3\mathbf{X}^T\mathbf{L_S}\mathbf{X} + \lambda\mathbf{Q}), \mathbf{B} = \alpha_2\mathbf{L_G}, \mathbf{Z} = \mathbf{X}^T\mathbf{Y}$, and $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with the i-th diagonal element set to

$$q_{ii} = \frac{1}{2\|\mathbf{w}^i\|_2}. \tag{13}$$

Here, we should note that due to the possibility of being zero for $\mathbf{w}^i$ in Eq. (13), we add a small constant to the denominator in implementation, by following Nie *et al.*'s work (Nie et al., 2010).

In solving Eq. (12), it is not trivial to find the optimum solution due to the inter-dependence in computing matrices of $\mathbf{W}$ and $\mathbf{Q}$. To this end, in this work, we apply an iterative approach by alternatively computing $\mathbf{Q}$ and $\mathbf{W}$. That is, at the t-th iteration, we first update the matrix $\mathbf{W}(t)$ with the matrix $\mathbf{Q}(t-1)$, and then update the matrix $\mathbf{Q}(t)$ with the updated matrix $\mathbf{W}(t)$. Refer to Algorithm 1 and Appendix A, respectively, for implementation details and the proof of convergence of our algorithm.

Algorithm 1: Pseudo code of solving Eq. (11).

Input: $\mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{Y} \in \mathbb{R}^{n\times c}, \alpha_1, \alpha_2, \alpha_3, \lambda$;
Output: $\mathbf{W}$;
Initialize $t = 0$ and set $\mathbf{Q}(t)$ a random diagonal matrix;
repeat
  Compute $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{Z}$ in Eq. (12);
  Factorize matrices $\mathbf{A} = \mathbf{P}^T \times \mathbf{P}$ and $\mathbf{B} = \mathbf{R} \times \mathbf{R}^T$;
  Perform singular value decomposition on $\mathbf{P}$ and $\mathbf{R}$;
  Update $\tilde{\mathbf{W}}(t+1)$ by Eq. (16) and Eq. (17);
  Compute $\mathbf{W}(t+1)$ by Eq. (18);
  Update $\mathbf{Q}(t+1)$ by Eq. (13);
  $t = t+1$;
until *Eq. (11)* converges;

Although there exists a general solver with this iterative approach[7], its computational complexity is known to be cubic. In this paper, we propose a simple but computationally more efficient algorithm. In Eq. (12), since both $\mathbf{A}$ and $\mathbf{B}$ are positive semi-definite, we can decompose them into two triangular matrices by Cholesky factorization (Golub and Van Loan, 1996):

$$\mathbf{A} = \mathbf{P}^T \times \mathbf{P}$$
$$\mathbf{B} = \mathbf{R} \times \mathbf{R}^T.$$

By applying a Singular Value Decomposition (SVD) on each of the triangular matrices, $\mathbf{P}$ and $\mathbf{R}$, we can further decompose them as follows:

$$\mathbf{P} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T$$
$$\mathbf{R} = \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^T$$

where $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are diagonal matrices whose elements correspond to eigenvalues, and $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{V}_1$, and $\mathbf{V}_2$ are unitary matrices, *i.e.*, $\mathbf{U}_1 \times \mathbf{U}_1^T = \mathbf{U}_1^T \times \mathbf{U}_1 = \mathbf{I}, \mathbf{U}_2 \times \mathbf{U}_2^T = \mathbf{U}_2^T \times \mathbf{U}_2 = \mathbf{I}, \mathbf{V}_1 \times \mathbf{V}_1^T = \mathbf{V}_1^T \times \mathbf{V}_1 = \mathbf{I}$, and $\mathbf{V}_2 \times \mathbf{V}_2^T = \mathbf{V}_2^T \times \mathbf{V}_2 = \mathbf{I}$. Then, we can rewrite Eq. (12) as follows:

$$\mathbf{V}_1\mathbf{\Sigma}_1^T\mathbf{\Sigma}_1\mathbf{V}_1^T\mathbf{W} + \mathbf{W}\mathbf{U}_2\mathbf{\Sigma}_2\mathbf{\Sigma}_2^T\mathbf{U}_2^T = \mathbf{Z}. \tag{14}$$

By multiplying $\mathbf{V}_1^T$ and $\mathbf{U}_2$ to both sides of Eq. (14), we can obtain

$$\mathbf{\Sigma}_1^T\mathbf{\Sigma}_1\mathbf{V}_1^T\mathbf{W}\mathbf{U}_2 + \mathbf{V}_1^T\mathbf{W}\mathbf{U}_2\mathbf{\Sigma}_2\mathbf{\Sigma}_2^T = \mathbf{V}_1^T\mathbf{Z}\mathbf{U}_2. \tag{15}$$

Let $\tilde{\mathbf{\Sigma}}_1 = \mathbf{\Sigma}_1^T\mathbf{\Sigma}_1, \tilde{\mathbf{\Sigma}}_2 = \mathbf{\Sigma}_2\mathbf{\Sigma}_2^T, \tilde{\mathbf{W}} = \mathbf{V}_1^T\mathbf{W}\mathbf{U}_2$, and $\mathbf{E} = \mathbf{V}_1^T\mathbf{Z}\mathbf{U}_2$, then we obtain the form of

$$\tilde{\mathbf{\Sigma}}_1\tilde{\mathbf{W}} + \tilde{\mathbf{W}}\tilde{\mathbf{\Sigma}}_2 = \mathbf{E}. \tag{16}$$

Note that both $\tilde{\mathbf{\Sigma}}_1 = \left[\tilde{\sigma}_{ii}^1\right] \in \mathbb{R}^{d\times d}$ and $\tilde{\mathbf{\Sigma}}_2 = [\tilde{\sigma}_{jj}^2] \in \mathbb{R}^{c\times c}$ are diagonal matrices. Therefore, it is straightforward to obtain $\tilde{\mathbf{W}} = \left[\tilde{w}_{ij}\right] \in \mathbb{R}^{d\times c}$ as follows:

$$\tilde{w}_{ij} = \frac{e_{ij}}{\tilde{\sigma}_{ii}^1 + \tilde{\sigma}_{jj}^2} \tag{17}$$

where e_{ij} denotes the (i,j)-th element in $\mathbf{E}$. From the matrix $\tilde{\mathbf{W}}$, we can obtain $\mathbf{W}$ by

$$\mathbf{W} = \mathbf{V}_1\tilde{\mathbf{W}}\mathbf{U}_2^T. \tag{18}$$

It is noteworthy that, thanks to the decomposed diagonal matrices obtained by Cholesky factorization and SVD, we can greatly reduce the computational cost in solving the optimization problem.

3.4. Feature selection and model training

Because of using the $\ell_{2,1}$-norm regularization term in our objective function, after finding the optimal solution with Algorithm 1, we have some zero or close to zero row vectors in $\mathbf{W}$. In terms of least square regression, the corresponding features are not necessary in regressing the response variables. Meanwhile, from the prediction perspective, the lower the ℓ_2-norm value of a row vector, the less informative the respective feature in our observation. To this end, we first sort rows in $\mathbf{W}$ in a descending order based on the ℓ_2-norm value of each row, *i.e.*, $\|\mathbf{w}^j\|_2, j \in \{1, \ldots, d\}$, and then select the features that correspond to the K top-ranked rows[8].

With the selected features, we then train support vector machines, which have been successfully used in many fields (Suk and Lee, 2013; Zhang and Shen, 2012). Note that the selected features are jointly used to predict two clinical scores and one clinical label. Specifically, we build two Support Vector Regression (SVR) (Smola and Schölkopf, 2004) models to predict ADAS-Cog and MMSE scores,

[7] For example, a built-in function 'lyap' in MATLAB.

[8] In this work, the proposed optimization method (*i.e.*, Algorithm 1) outputs many zero-rows, which determine the value of K.

respectively, and one Support Vector Classification (Burges, 1998) model to identify a clinical label, via the public LIBSVM toolbox [9].

4. Experimental results

4.1. Experimental setting

We considered three binary classification problems: AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC. For MCI vs. NC, both MCI-C and MCI-NC were labeled as MCI. For each set of experiments, we used 93 MRI features or 93 PET features as regressors, and 2 clinical scores along with 1 class label for responses in the least square regression model.

Due to the limited small number of samples, we used a 10-fold cross-validation technique to measure the performances. Specifically, we partitioned the data of each class into 10 disjoints sets, *i.e.,* 10 folds. Then we selected two sets, one from each class, for testing while using the remaining 18 sets (*e.g.,* 9 sets from AD and 9 sets from NC in the case of AD vs. NC classification) for training in the binary classification task. We repeated the process 10 times to avoid the possible bias occurring in dataset partitioning. The final results were computed by averaging the repeated experiments. For model selection, *i.e.,* tuning parameters in Eq. (11) and SVR/SVC parameters[10], we further split the training samples into 5 subsets for 5-fold inner cross-validation. In our experiments, we conducted exhaustive grid search on the parameters with the spaces of $\alpha_i \in \{10^{-6}, \ldots, 10^{2}\}, i \in \{1, 2, 3\}$,and $\lambda \in \{10^{2}, \ldots, 10^{8}\}$. We empirically set $k = 3$and $\sigma = 1$to calculate three kinds of similarity, such as m_{ij} in Eq. (3), g_{ij} in Eq. (5), and s_{ij} in Eq. (6). The parameters that resulted in the best performance in the inner cross-validation were finally used in testing.

To evaluate the performance of all competing methods, we employed the metrics of Correlation Coefficient (CC) and Root Mean Squared Error (RMSE) between the target clinical scores and the predicted ones in regression, and also the metrics of classification ACCuracy (ACC), SENsitivity (SEN), SPEcificity (SPE), Area Under Curve (AUC), and Receiver Operating Characteristic (ROC) curves in classification.

4.2. Competing methods

To validate the effectiveness of the proposed method, we performed extensive experiments comparing with the state-of-the-art methods. Specifically, we considered rigorous experimental conditions: (1) In order to show the validity of the feature selection strategy, we performed the tasks of regression and classification without precedent feature selection, and considered them as a baseline method. Hereafter, we use the suffix "N" to indicate that no feature selection was involved in. For example, by MRI-N, we mean that either the classification or regression was performed using the full MRI features. (2) One of the main arguments in our work is to select features that can be jointly used for both regression and classification. To this end, we compare the multi-task based method with a single-task based method, in which the feature selection was carried out for regression and classification independently. In the following, the suffix "S" manifests a single-task based method. For example, MRI-S represents single-task based feature selection on MRI features. (3) We compare with two state-of-the-art methods: High-Order Graph Matching (HOGM) (Liu et al., 2013) and Multi-Modal Multi-Task (M3T) (Zhang and Shen, 2012). The former used a sample-sample relation along with an ℓ_1-norm regularization term in an optimization of single-task learning. The latter used multi-task learning with an $\ell_{2,1}$-norm regularization term only to select a common set of features for all tasks of regression and classification. Note that M3T is a special case of the proposed method by setting $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

4.3. Classification results

Table 2 shows the classification performances of the competing methods. We also compare the ROC curves of the competing methods on three classification problems in Fig. 3. From these results, we can draw three conclusions. First, it is important to conduct feature selection on the high-dimensional features before training a classifier. The baseline methods with no feature selection, *i.e.,* MRI-N, and PET-N, reported the worst performances. The simple feature selection method, *i.e.,* MRI-S and PET-S, still helped increase the classification accuracy by 1.7% (AD vs. NC), 8.4% (MCI vs. NC), and 4.2% (MCI-C vs. MCI-NC) compared to MRI-N, and by 1.7% (AD vs. NC), 4.8% (MCI vs. NC), and 3.9% (MCI-C vs. MCI-NC) compared to PET-N, respectively. The other more sophisticated methods further improved the accuracies. Note that the proposed method maximally enhanced the classification accuracies by 4.8% (AD vs. NC), 11.4% (MCI vs. NC), and 11.5% (MCI-C vs. MCI-NC) with MRI, and by 5.6% (AD vs. NC), 10.2% (MCI vs. NC), and 9.0% (MCI-C vs. MCI-NC) with PET, respectively, compared to the baseline method.

Second, it is beneficial to use joint regression and classification framework, *i.e.,* multi-task learning, for feature selection. As shown in Table 2, M3T and our method, which utilized the multi-task learning, achieved better classification performances than the single-task based method. Specifically, the proposed method showed the superiority to the single-task based method, *i.e.,* MRI-S and PET-S, improving the accuracies by 2.5% (AD vs. NC), 3.0% (MCI vs. NC), and 7.3% (MCI-C vs. MCI-NC) with MRI, and by 3.9% (AD vs. NC), 10.2% (MCI vs. NC), and 9.0% (MCI-C vs. MCI-NC) with PET, respectively.

Lastly, based on the fact that the best performances over the three binary classifications were all obtained by our method, we can say that the proposed regularization terms were effective to find class-discriminative features. It is worth noting that compared to the state-of-the-art methods, the accuracy enhancements by our method were 5% (vs. HOGM) and 4.7% (vs. M3T) with MRI, and 4.6% (vs. HOGM) and 4.2% (vs. M3T) with PET for MCI-C vs. MCI-NC classification, which is the most important for early diagnosis and treatment.

4.4. Regression results

Regarding the prediction of two clinical scores of MMSE and ADAS-Cog, we summarized the results in Table 3 and presented scatter plots of the predicted ADAS-Cog scores with MRI against the target ones in Fig. 4. In Table 3, we can see that, similar to the classification results, the regression performance of the methods without feature selection (MRI-N and PET-N) was worse than any of the other methods with feature selection. Moreover, our method consistently outperformed the competing methods for the cases of different pairs of clinical labels.

In the regression with MRI for AD vs. NC, our method showed the best CCs of 0.669 for ADAS-Cog and 0.679 for MMSE, and the best RMSEs of 4.43 for ADAS-Cog and 1.79 for MMSE. The next best performances in terms of CCs were obtained by M3T, *i.e.,* 0.649 for ADAS-Cog and 0.638 for MMSE, and those in terms of RMSEs were obtained by HOGM, *i.e.,* 4.53 for ADAS-Cog and 1.91 for MMSE. In the regression with MRI for MCI vs. NC, our method also achieved the best CCs of 0.472 for ADAS-Cog and 0.50 for MMSE, and the best RMSEs of 4.23 for ADAS-Cog and 1.63 for MMSE. For the case of MCI-C vs. MCI-NC with MRI, the proposed method improved the CCs by 0.092 for ADAS-Cog and 0.053 for MMSE compared to the next best CCs of

[9] Available at 'http://www.csie.ntu.edu.tw/~cjlin/libsvm/'.
[10] $C \in \{2^{-5}, \ldots, 2^{5}\}$in our experiments.

Table 2
Comparison of classification performances (%) of the competing methods. (ACCuracy (ACC), SENsitivity (SEN), SPEcificity (SPE), and Area Under Curve (AUC)).

Feature	Method	AD vs. NC					MCI vs. NC					MCI-C vs. MCI-NC				
		ACC	SEN	SPE	AUC	p-value	ACC	SEN	SPE	AUC	p-value	ACC	SEN	SPE	AUC	p-value
MRI	MRI-N	89.5	85.7	89.3	93.3	<0.001	68.3	92.6	43.9	78.2	<0.001	60.3	15.5	92.3	68.7	<0.001
	MRI-S	91.2	87.1	92.2	94.7	<0.001	76.7	93.3	47.6	81.5	<0.001	64.5	24.9	**95.8**	70.6	<0.001
	HOGM	93.4	**89.5**	92.5	97.1	0.002	77.7	**95.6**	51.4	84.4	<0.001	66.8	36.7	95.0	72.2	<0.001
	M3T	92.6	87.2	95.9	97.5	<0.001	78.1	94.5	54.0	83.1	<0.001	67.1	37.7	92.0	72.5	<0.001
	Proposed	**93.7**	88.6	**97.8**	**97.6**	–	**79.7**	94.8	**56.9**	**84.7**	–	**71.8**	**48.0**	92.8	**81.4**	–
PET	PET-N	86.2	88.5	87.8	90.2	<0.001	69.0	95.0	37.8	76.2	<0.001	62.2	21.6	93.1	71.3	<0.001
	PET-S	87.9	89.7	91.9	93.1	<0.001	73.8	96.5	39.2	77.6	<0.001	65.1	31.0	**95.5**	73.5	<0.001
	HOGM	91.7	91.1	92.8	95.6	0.003	74.7	96.5	43.2	79.3	<0.001	66.6	35.5	**95.5**	72.4	<0.001
	M3T	90.9	90.5	93.1	96.4	<0.001	77.2	94.5	44.3	80.5	<0.001	67.0	39.1	93.2	73.1	<0.001
	Proposed	**91.8**	**91.5**	**93.8**	**96.9**	–	**79.2**	**97.1**	**45.3**	**80.8**	–	**71.2**	**47.4**	93.0	**77.6**	–

Table 3
Comparison of regression performances of the competing methods in terms of Correlation Coefficient (CC) and Root Mean Square Error (RMSE).

Feature	Method	AD vs. NC					MCI vs. NC					MCI-C vs. MCI-NC				
		ADAS-Cog		MMSE			ADAS-Cog		MMSE			ADAS-Cog		MMSE		
		CC	RMSE	CC	RMSE	p-value	CC	RMSE	CC	RMSE	p-value	CC	RMSE	CC	RMSE	p-value
MRI	MRI-N	0.587	4.96	0.520	2.02	<0.001	0.329	4.48	0.309	1.90	<0.001	0.420	4.10	0.441	1.51	<0.001
	MRI-S	0.591	4.85	0.566	1.95	<0.001	0.347	4.27	0.367	1.64	<0.001	0.426	4.01	0.482	1.44	<0.001
	HOGM	0.625	4.53	0.598	1.91	<0.001	0.352	4.26	0.371	1.63	<0.001	0.435	3.94	0.521	1.41	<0.001
	M3T	0.649	4.60	0.638	1.91	<0.001	0.445	4.27	0.420	1.66	<0.001	0.497	4.01	0.550	1.41	<0.001
	Proposed	**0.669**	**4.43**	**0.679**	**1.79**	–	**0.472**	**4.23**	**0.500**	**1.62**	–	**0.589**	**3.83**	**0.603**	**1.40**	–
PET	PET-N	0.597	4.86	0.514	2.04	<0.001	0.333	4.34	0.331	1.70	<0.001	0.382	4.08	0.452	1.50	<0.001
	PET-S	0.620	4.83	0.593	2.00	<0.001	0.356	4.26	0.359	1.69	<0.001	0.437	4.00	0.478	1.48	<0.001
	HOGM	0.600	4.69	0.515	1.99	<0.001	0.360	4.21	0.368	1.67	<0.001	0.430	4.03	0.523	1.41	<0.001
	M3T	0.647	4.67	0.593	1.92	<0.001	0.447	4.24	0.432	1.68	<0.001	0.520	3.91	0.569	1.45	0.003
	Proposed	**0.671**	**4.41**	**0.620**	**1.90**	–	**0.513**	**4.13**	**0.485**	**1.66**	–	**0.526**	**3.87**	**0.570**	**1.37**	–

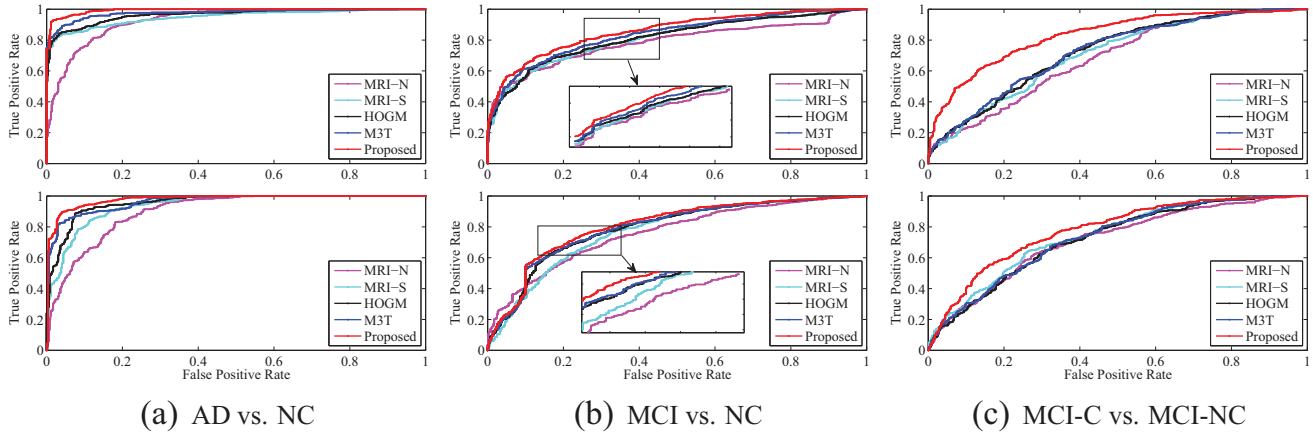


Fig. 3. Comparison of Receiver Operating Characteristic (ROC) curves for the competing methods on three binary classifications. The plots in the upper and the lower rows were, respectively, obtained with MRI and PET.

0.497 for ADAS-Cog and 0.550 for MMSE by M3T. Note that the proposed method with PET also reported the best CCs and RMSEs for both ADAS-Cog and MMSE over the three regression problems, *i.e.,* AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC.

4.5. Effects of the proposed regularization trms

In order to see the effects of each of the proposed regularization terms, such as sample-sample relation, feature-feature relation, and response-response relation[11], we further compared the performances of the proposed method with those of its counterparts that consider one of the terms or a pair of them. We present the performances of the counterpart methods and the proposed method in Fig. 5. For better understanding, we also presented the performances of M3T as baseline that doesn't consider any of three regularization terms. From the figure, we can observe the following that: (1) A method that utilizes any one of the three regularization terms is still better than M3T; (2) The inclusion of more than two regularization terms into the objective function resulted in better performances than a single regularization, and ultimately the full utilization of the three relational characteristics achieved the best performances.

4.6. Multiple modalities fusion

With respect to multi-modal fusion, it is known that different modalities can provide complementary information, and thus can enhance the diagnostic accuracy (Cui et al., 2011; Hinrichs et al., 2011; Kohannim et al., 2010; Perrin et al., 2009; Suk and Shen, 2013; Walhovd et al., 2010; Westman et al., 2012). For this reason, we also per-

[11] For example, we considered the feature-feature relation by setting $\alpha_1 = 0$and $\alpha_2 = 0$in Eq. (11).

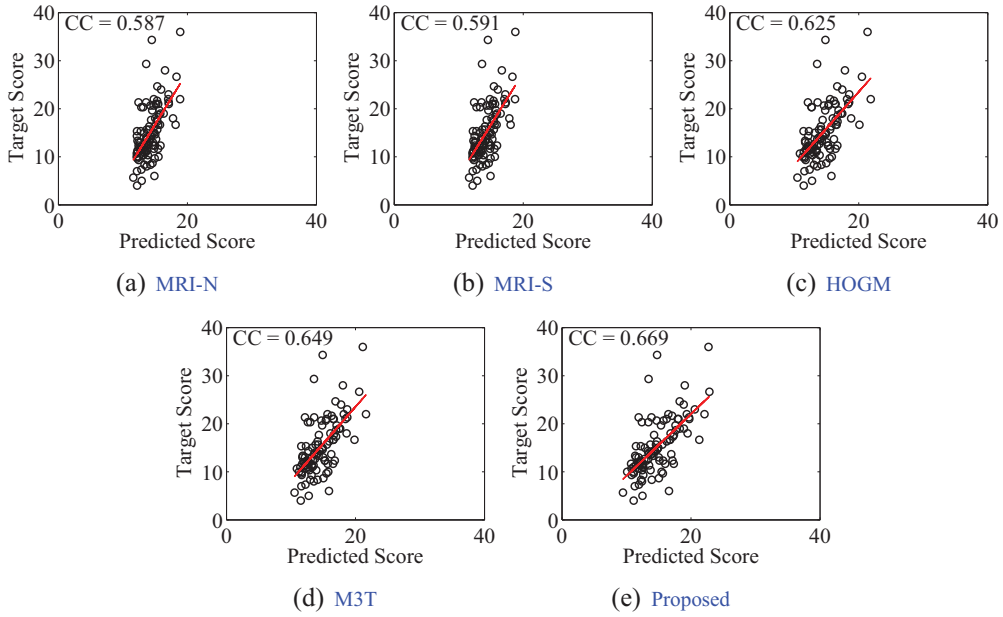


Fig. 4. Scatter plots of the target ADAS-Cog scores against the predicted ones, which were obtained with MRI for AD vs. NC.

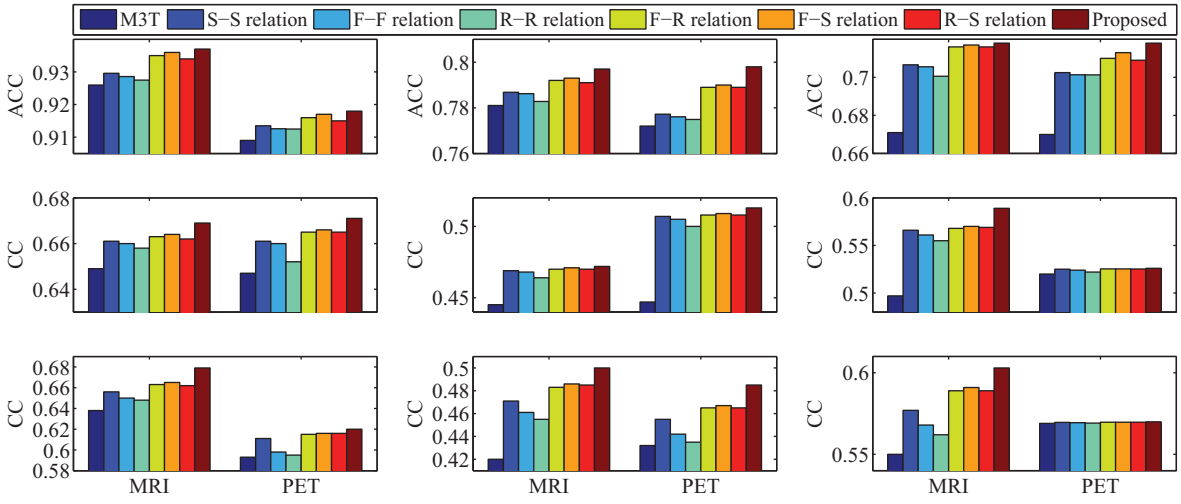


Fig. 5. Comparison of ACCuracy (ACC) (top row), Correlation Coefficient (CC) of ADAS-Cog (middle row), and CC of MMSE (bottom row) among the competing methods for three binary classifications: AD vs. NC (left column), MCI vs. NC (middle column), and MCI-C vs. MCI-NC (right column). 'S', 'F', and 'R' stand for 'Sample', 'Feature', and 'Response', respectively.

formed experiments using both MRI and PET (MP for short). We constructed a new feature matrix **X** with a concatenation of MRI and PET features at each row, but used the same response matrix **Y** as the above-described experiments.

Tables 4 and 5 summarize the results of clinical label identification and clinical scores estimation, respectively. In line with the previous researches, the modality fusion helped improve performances in both classification and regression. Moreover, all the methods with the modality fusion selected the aforementioned brain regions with higher 'Frequency' than the corresponding methods with a single modality, such as on average 99.2%, 93.1%, and 92.7%, respectively, for our method, HOGM and M3T, on the data with the modality fusion.

Finally, to check statistical significance, we conducted the paired *t*-tests (Dietterich, 1998) (at 95% significance level) on the classification and regression performances of our method and the competing methods (including the experiments in Sections 4.3–4.6). Tables 2 and 4 show the *p*-values obtained from the values of ACC, while Tables 3 and 5 show the *p*-values computed from the values of CC. All these resulting *p*-values indicate that our method is statistically superior to the competing methods on the tasks of *either* predicting clinical scores (*i.e.,* ADAS-Cog and MMSE) *or* identifying class label.

5. Conclusions

In this work, we proposed a novel feature selection method by devising new regularization terms that consider relational information inherent in the observations for joint regression and classification in the computer-aided AD diagnosis. In our extensive experiments on the ADNI dataset, we validated the effectiveness of the proposed method by comparing with the state-of-the-art methods for both the clinical scores (ADAS-Cog and MMSE) prediction and the clinical label identification. The utilization of the devised three regularization terms that consider relational information in observation, *i.e.,* sample–sample relation, feature–feature relation, and response–response relation, were helpful to improve the perfor-

Table 4
Performance comparison among competing methods with multi-modal fusion. (ACCuracy (ACC), SENsitivity (SEN), SPEcificity (SPE), Area Under Curve (AUC), fusion of MRI and PET (MP)).

Method	AD vs. NC					MCI vs. NC					MCI-C vs. MCI-NC				
	ACC	SEN	SPE	AUC	p-value	ACC	SEN	SPE	AUC	p-value	ACC	SEN	SPE	AUC	p-value
MP-N	89.7	92.2	89.5	94.1	<0.001	71.6	96.1	43.9	82.7	<0.001	62.7	22.6	93.5	73.2	<0.001
MP-S	90.8	92.6	93.8	96.7	<0.001	76.3	**97.0**	49.9	83.4	<0.001	66.9	33.9	96.0	75.7	<0.001
HOGM	95.2	92.8	95.4	97.8	0.001	79.5	96.6	58.6	84.6	0.003	67.6	45.5	**96.8**	75.1	<0.001
M3T	94.0	92.0	96.3	98.0	<0.001	78.4	95.0	57.7	83.9	<0.001	67.9	47.0	93.3	75.7	<0.001
Proposed	**95.7**	**96.6**	**98.2**	**98.1**	–	**79.9**	**97.0**	**59.2**	**84.9**	–	**72.4**	**49.1**	94.6	**82.9**	–

Table 5
Comparison of regression performances of the competing methods in terms of Correlation Coefficient (CC) and Root Mean Square Error (RMSE) by fusing MRI and PET (MP).

Method	AD vs. NC					MCI vs. NC					MCI-C vs. MCI-NC				
	ADAS-Cog		MMSE			ADAS-Cog		MMSE			ADAS-Cog		MMSE		
	CC	RMSE	CC	RMSE	p-value	CC	RMSE	CC	RMSE	p-value	CC	RMSE	CC	RMSE	p-value
MP-N	0.626	4.80	0.587	1.99	<0.001	0.365	4.29	0.335	1.69	<0.001	0.431	4.09	0.455	1.47	<0.001
MP-S	0.634	4.83	0.585	1.92	<0.001	0.359	4.25	0.371	1.67	<0.001	0.449	4.00	0.496	1.41	<0.001
HOGM	0.633	4.64	0.602	1.83	<0.001	0.364	4.20	0.365	1.65	<0.001	0.450	3.93	0.531	1.40	<0.001
M3T	0.653	4.61	0.639	1.91	<0.001	0.450	4.23	0.433	1.64	<0.001	0.522	3.81	0.567	1.36	<0.001
Proposed	**0.680**	**4.40**	**0.682**	**1.78**	–	**0.520**	**4.02**	**0.508**	**1.61**	–	**0.591**	**3.78**	**0.622**	**1.35**	–

mances in the JRC problem, and outperformed the state-of-the-art methods.

It should be noted that while the proposed method was successful to enhance the performances for AD/MCI diagnosis, the current method considered only the linear relationships inherent in the observations. Therefore, it will be our forthcoming research issue to extend the current work to the nonlinear formulation via the kernel methods.

Acknowledgments

This work was supported in part by NIH grants (EB006733, EB008374, EB009634, MH100217, AG041721, AG042599), the ICT R&D program of MSIP/IITP [B0101-15-0307, Basic Software Research in Human-level Lifelong Machine Learning (Machine Learning Centre)], and the National Research Foundation of Korea (NRF) grant funded by the Korea government (NRF-2015R1A2A1A05001867). Xiaofeng Zhu was supported in part by the National Natural Science Foundation of China under grants (61263035 and 61573270), the Guangxi Natural Science Foundation under grant (2015GXNSFCB139011), and the funding of Guangxi 100 Plan.

Appendix A

We prove that the proposed Algorithm 1 makes the value of the objective function in Eq. (11) monotonically decrease. We first give a Lemma from (Nie et al., 2010) as follows, which will be used in our proof.

Lemma 1. *For any nonzero row vectors* $(\mathbf{w}(t))^i \in \mathbb{R}^c$ *and* $(\mathbf{w}(t+1))^i \in \mathbb{R}^c$, *where* $i \in \{1, \cdots, d\}$ *and* t *denotes an index of iteration, the following inequality holds:*

$$\sum_{i=1}^{d}\left(\left(\frac{\|(\mathbf{w}(t+1))^i\|_2^2}{2\|(\mathbf{w}(t))^i\|_2}-\|(\mathbf{w}(t+1))^i\|_2\right)-\left(\frac{\|(\mathbf{w}(t))^i\|_2^2}{2\|(\mathbf{w}(t))^i\|_2}-\|(\mathbf{w}(t))^i\|_2\right)\right)\geq 0. \tag{A.1}$$

Theorem 1. *In each iteration, Algorithm 1 monotonically decreases the objective function value in Eq. (11).*

Proof. In Algorithm 1, we denote part of Eq. (11), *i.e.*, without the last term $\lambda\|\mathbf{W}\|_{2,1}$, in the t-th iteration as $\mathcal{L}(t) = \|\mathbf{Y}-\mathbf{X}\mathbf{W}(t)\|_F^2+\alpha_1 tr((\mathbf{W}(t))^T\mathbf{L_M}\mathbf{W}(t)) + \alpha_2 tr(\mathbf{W}(t)\mathbf{L_G}(\mathbf{W}(t))^T)+\alpha_3 tr((\mathbf{X}\mathbf{W}(t))^T\mathbf{L_S}\mathbf{X}\mathbf{W}(t))$. We also denote $\mathbf{Q}(t)$ as the optimal value in the t-th iteration for $\mathbf{Q}$. According to (Nie et al., 2010), optimizing the non-smooth convex form $\|\mathbf{W}\|_{2,1}$ can be transferred to iteratively optimize $\mathbf{Q}$ and $\mathbf{W}$ in $tr(\mathbf{W}^T\mathbf{Q}\mathbf{W})$. Therefore, according to the steps of line 6 and 7 in Algorithm 1, we have

$$\begin{aligned}&\mathcal{L}(t+1)+\lambda tr((\mathbf{W}(t+1))^T\mathbf{Q}(t)\mathbf{W}(t+1))\\&\leq \mathcal{L}(t)+\lambda tr((\mathbf{W}(t))^T\mathbf{Q}(t)\mathbf{W}(t)).\end{aligned} \tag{A.2}$$

By changing the trace form into the form of summation, we have

$$\mathcal{L}(t+1)+\lambda\sum_{i=1}^{d}\frac{\|(\mathbf{w}(t+1))^i\|_2^2}{2\|(\mathbf{w}(t))^i\|_2}\leq\mathcal{L}(t)+\lambda\sum_{i=1}^{d}\frac{\|(\mathbf{w}(t))^i\|_2^2}{2\|(\mathbf{w}(t))^i\|_2}. \tag{A.3}$$

With a simple modification, we can have

$$\begin{aligned}&\mathcal{L}(t+1)+\lambda\sum_{i=1}^{d}\left(\frac{\|(\mathbf{w}(t+1))^i\|_2^2}{2\|(\mathbf{w}(t))^i\|_2}-\|(\mathbf{w}(t+1))^i\|_2+\|(\mathbf{w}(t+1))^i\|_2\right)\\&\leq\mathcal{L}(t)+\lambda\sum_{i=1}^{d}\left(\frac{\|(\mathbf{w}(t))^i\|_2^2}{2\|(\mathbf{w}(t))^i\|_2}-\|(\mathbf{w}(t))^i\|_2+\|(\mathbf{w}(t))^i\|_2\right).\end{aligned} \tag{A.4}$$

After reorganizing terms, we finally have

$$\begin{aligned}&\mathcal{L}(t+1)+\lambda\sum_{i=1}^{d}\|(\mathbf{w}(t+1))^i\|_2+\lambda\sum_{i=1}^{d}\left(\left(\frac{\|(\mathbf{w}(t+1))^i\|_2^2}{2\|(\mathbf{w}(t))^i\|_2}-\|(\mathbf{w}(t+1))^i\|_2\right)-\left(\frac{\|(\mathbf{w}(t))^i\|_2^2}{2\|(\mathbf{w}(t))^i\|_2}-\|(\mathbf{w}(t))^i\|_2\right)\right)\\&\leq\mathcal{L}(t)+\lambda\sum_{i=1}^{d}\|(\mathbf{w}(t))^i\|_2.\end{aligned} \tag{A.5}$$

According to Lemma 1, the third term of the left side in Eq. (A.5) is non-negative. Therefore, the following inequality holds

$$\mathcal{L}(t+1)+\lambda\sum_{i=1}^{d}\|(\mathbf{w}(t+1))^i\|_2\leq\mathcal{L}(t)+\lambda\sum_{i=1}^{d}\|(\mathbf{w}(t))^i\|_2. \tag{A.6}$$

□

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.media.2015.10.008

References

Belkin, M., Niyogi, P., Sindhwani, V., 2006. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. 7, 2399–2434.
Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, M.H., 2007. Forecasting the global burden of Alzheimer's disease.. Alzheimer's & dementia :J. Alzheimer's Assoc. 3 (3), 186–191.
Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discover 2 (2), 121–167.
Cheng, B., Zhang, D., Chen, S., Kaufer, D., Shen, D., 2013. Semi-supervised multimodal relevance vector regression improves cognitive performance estimation from imaging and biological biomarkers. Neuroinformatics 11 (3), 339–353.
Cho, Y., Seong, J.-K., Jeong, Y., Shin, S.Y., 2012. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. NeuroImage 59 (3), 2217–2230.
Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T., Jin, J.S., the Alzheimer's Disease Neuroimaging Initiative, 2011. Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. PLoS One 6 (7), e21896.
Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage 56 (2), 766–781.
Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10 (7), 1895–1923.
Fan, Y., Gur, R.E., Gur, R.C., Wu, X., Shen, D., Calkins, M.E., Davatzikos, C., 2008. Unaffected family members and schizophrenia patients share brain structure patterns: a high-dimensional pattern classification study. Biol. Psychiatry 63 (1), 118–124.
Fjell, A.M., Walhovd, K.B., Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Holland, D., Brewer, J.B., Dale, A.M., the Alzheimer's Disease Neuroimaging Initiative, 2010. CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease. J. Neurosci. 30 (6), 2088–2101.
Franke, K., Ziegler, G., Klöppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. NeuroImage 50 (3), 883–892.
Golub, G.H., Van Loan, C.F., 1996. Matrix Computations (3rd Ed.). Johns Hopkins University Press.
He, X., Cai, D., Niyogi, P., 2005. Laplacian score for feature selection. In: Proceedings of the NIPS, pp. 1–8.
Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., 2011. Predictive markers for AD in a multimodality framework: An analysis of MCI progression in the ADNI population. NeuroImage 55 (2), 574–589.
Jie, B., Zhang, D., Cheng, B., Shen, D., 2013. Manifold regularized multi-task feature selection for multi-modality classification in Alzheimers disease. In: Proceedings of the MICCAI, pp. 9–16.
Kabani, N.J., 1998. 3D anatomical atlas of the human brain. NeuroImage 7, 0700–0717.
Kohannim, O., Hua, X., Hibar, D.P., Lee, S., Chou, Y.-Y., Toga, A.W., Jr., C.R.J., Weiner, M.W., Thompson, P.M., 2010. Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiol. Aging 31 (8), 1429–1442.
Liu, F., Suk, H.-I., Wee, C.-Y., Chen, H., Shen, D., 2013. High-order graph matching based feature selection for Alzheimer's disease identification. In: Proceedings of the MICCAI, pp. 311–318.
Liu, F., Wee, C.-Y., Chen, H., Shen, D., 2014. Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. NeuroImage 84, 466–475.
McEvoy, L.K., Fennema-Notestine, C., Roddey, J.C., Hagler, D.J., Holland, D., Karow, D.S., Pung, C.J., Brewer, J.B., Dale, A.M., 2009. Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment.. Radiology 251 (5), 195–205.
Morris, J., Storandt, M., Miller, J., et al, 2001. Mild cognitive impairment represents early-stage Alzheimer disease. Arch. Neurol. 58 (3), 397–405.
Nie, F., Huang, H., Cai, X., Ding, C.H.Q., 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In: Proceedings of the NIPS, pp. 1813–1821.
Perrin, R.J., Fagan, A.M., Holtzman, D.M., 2009. Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. Nature 461, 916–922.
Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326.
Shen, D., Davatzikos, C., 2002. HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. 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 Trans. Med. Imaging 17 (1), 87–97.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222.
Suk, H.-I., Lee, S.-W., 2013. A novel Bayesian framework for discriminative feature extraction in brain-computer interfaces. IEEE Trans. Pattern Anal. Mach. Intell. 35 (2), 286–299.
Suk, H.-I., Lee, S.-W., Shen, D., 2014. Subclass-based multi-task learning for Alzheimer's disease diagnosis. Front. Aging Neurosci. 6 (168).
Suk, H.-I., Lee, S.-W., Shen, D., 2015. Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. Brain Struct. Funct. 1–19.
Suk, H.-I., Shen, D., 2013. Deep learning-based feature representation for AD/MCI classification. In: Proceedings of the MICCAI, pp. 583–590.
Suk, H.-I., Wee, C.-Y., Shen, D., 2013. Discriminative group sparse representation for mild cognitive impairment classification. In: Proceedings of the MLMI, pp. 131–138.
Tang, S., Fan, Y., Wu, G., Kim, M., Shen, D., 2009. RABBIT: rapid alignment of brains by building intermediate templates. NeuroImage 47 (4), 1277–1287.
Walhovd, K., Fjell, A., Dale, A., McEvoy, L., Brewer, J., Karow, D., Salmon, D., Fennema-Notestine, C., 2010. Multi-modal imaging predicts memory performance in normal aging and cognitive decline. Neurobiol. Aging 31 (7), 1107–1121.
Wang, H., Nie, F., Huang, H., Risacher, S., Saykin, A.J., Shen, L., 2011. Identifying AD-sensitive and cognition-relevant imaging biomarkers via joint classification and regression. In: Proceedings of the MICCAI, pp. 115–123.
Wang, Y., Nie, J., Yap, P.-T., Li, G., Shi, F., Geng, X., Guo, L., Shen, D., 2014. Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. PLoS One 9 (1).
Wee, C.-Y., Yap, P.-T., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D., 2012. Resting-state multi-spectrum functional connectivity networks for identification of MCI patients. PloS One 7 (5), e37828.
Wee, C.-Y., Yap, P.-T., Shen, D., 2013. Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Hum. Brain Mapp. 34 (12), 3411–3425.
Weinberger, K.Q., Sha, F., Saul, L.K., 2004. Learning a kernel matrix for nonlinear dimensionality reduction. In: Proceedings of the ICML, pp. 17–24.
Westman, E., Muehlboeck, J.-S., Simmons, A., 2012. Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. NeuroImage 62 (1), 229–238.
Wu, G., Jia, H., Wang, Q., Shen, D., 2011. Sharpmean: groupwise registration guided by sharp mean image and tree-based registration. NeuroImage 56 (4), 1968–1981.
Xue, Z., Shen, D., Davatzikos, C., 2006. Statistical representation of high-dimensional deformation fields with application to statistically constrained 3D warping. Med. Image Anal. 10 (5), 740–751.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. 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., Shen, D., et al., 2012. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. PloS One 7 (3), e33182.
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 Trans. Med. Imaging 20 (1), 45–57.
Zhou, L., Wang, Y., Li, Y., Yap, P.-T., Shen, D., et al., 2011. Hierarchical anatomical brain networks for MCI prediction: revisiting volumetric measures. PloS One 6 (7), e21935.
Zhu, X., Huang, Z., Cheng, H., Cui, J., Shen, H.T., 2013a. Sparse hashing for fast multimedia search. ACM Trans. Inf. Syst. 31 (2), 9.
Zhu, X., Huang, Z., Cui, J., Shen, T., 2013b. Video-to-shot tag propagation by graph sparse group lasso. IEEE Trans. Multim. 13 (3), 633–646.
Zhu, X., Huang, Z., Shen, H.T., Cheng, J., Xu, C., 2012. Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recogn. 45 (8), 3003–3016.
Zhu, X., Huang, Z., Yang, Y., Tao Shen, H., Xu, C., Luo, J., 2013c. Self-taught dimensionality reduction on the high-dimensional small-sized data. Pattern Recogn. 46 (1), 215–229.
Zhu, X., Li, X., Zhang, S., 2015a. Block-row sparse multiview multilabel learning for image classification. IEEE Trans. Cybern. 0 (0), online.
Zhu, X., Suk, H., Shen, D., 2014a. A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. NeuroImage 100, 91–105.
Zhu, X., Suk, H., Shen, D., 2014b. A novel multi-relation regularization method for regression and classification in AD diagnosis. In: Proceedings of the MICCAI, pp. 401–408.
Zhu, X., Suk, H.-I., Lee, S.-W., Shen, D., 2015b. Canonical feature selection for joint regression and multi-class identification in alzheimers disease diagnosis. Brain Imaging Behav. 1–11.
Zhu, X., Suk, H.-I., Lee, S.-W., Shen, D., 2015c. Subspace regularized sparse multitask learning for multi-class neurodegenerative disease identification. IEEE Trans. Biomed. Eng. 0 (0), online.
Zhu, X., Zhang, L., Huang, Z., 2014c. A sparse embedding and least variance encoding approach to hashing. IEEE Trans. Image Process. 23 (9), 3737–3750.
Zhu, X., Zhang, S., Jin, Z., Zhang, Z., Xu, Z., 2011. Missing value estimation for mixed-attribute data sets. IEEE Trans. Knowl. Data Eng. 23 (1), 110–121.

Please cite this article as: X. Zhu et al., A novel relational regularization feature selection method for joint regression and classification in AD diagnosis, Medical Image Analysis (2015), http://dx.doi.org/10.1016/j.media.2015.10.008