Relational-Regularized Discriminative Sparse Learning for Alzheimer's Disease Diagnosis

Baiying Lei, *Member, IEEE*, Peng Yang, Tianfu Wang, Siping Chen, and Dong Ni, *Member, IEEE*

***Abstract*—Accurate identification and understanding informative feature is important for early Alzheimer's disease (AD) prognosis and diagnosis. In this paper, we propose a novel discriminative sparse learning method with relational regularization to jointly predict the clinical score and classify AD disease stages using multimodal features. Specifically, we apply a discriminative learning technique to expand the class-specific difference and include geometric information for effective feature selection. In addition, two kind of relational information are incorporated to explore the intrinsic relationships among features and training subjects in terms of similarity learning. We map the original feature into the target space to identify the informative and predictive features by sparse learning technique. A unique loss function is designed to include both discriminative learning and relational regularization methods. Experimental results based on a total of 805 subjects [including 226 AD patients, 393 mild cognitive impairment (MCI) subjects, and 186 normal controls (NCs)] from AD neuroimaging initiative database show that the proposed method can obtain a classification accuracy of 94.68% for AD versus NC, 80.32% for MCI versus NC, and 74.58% for progressive MCI versus stable MCI, respectively. In addition, we achieve remarkable performance for the clinical scores prediction and classification label identification, which has efficacy for AD disease diagnosis and prognosis. The algorithm comparison demonstrates the effectiveness of the introduced learning techniques and superiority over the state-of-the-arts methods.**

***Index Terms*—Alzheimer's disease (AD) diagnosis, discriminative sparse learning, feature selection, relational regularization.**

I. Introduction

ALZHEIMER'S disease neuroimaging initiative (ADNI), funded by NIH in 2003, has received ever increasing attention for onset prediction and progression modeling of Alzheimer's disease (AD) and its early stage, e.g., mild cognitive impairment (MCI) [1]. The initiative was designed to expedite the scientific neuroimaging data evaluation [e.g., magnetic resonance imaging (MRI), positron emission tomography (PET), and cerebrospinal fluid (CSF)] [2]. Making a definite diagnosis of AD patients requires an invasive biopsy, which is quite expensive and inconvenient. In this regard, it is of vital importance to identify sensitive and specific biomarkers for early AD progression assessment and monitoring of new treatments. Also, it is critical to develop an automatic diagnosis tool for possible early treatment due to the financial and psychological burden of AD. As a result, many machine learning methods and pattern analysis of AD-related pathologies have been proposed to address this issue. Since various neuroimaging modalities (e.g., MRI and PET) can provide complementary information, they have been widely applied in AD study [2]–[8].

Feature selection, i.e., finding effective biomarkers, is important for AD diagnosis and prediction [2]–[5], [9]–[15]. Recent studies have demonstrated feature selection has the capability of overcoming dimensional curse issues after removing the indistinctive features [16]–[25]. To this date, numerous attempts have witnessed to identify the disease-related and informative features for class label identification and clinical score prediction [2]–[7], [26]. For instance, Zhang and Shen [2] proposed a multitask sparse learning method for feature selection method to predict clinical scores and identify disease status, and showed that such a joint learning could obtain better performance than performing them separately. Zhu *et al.* [5] also showed that the consideration of information inherent in observations was helpful to improve final AD diagnosis results. However, most methods are with simple vector stacking, and they degrade the performance due to ignorance of the underlying relational information included in the samples and imaging data. By contrast, manifold learning-based method, including the complementarity of the heterogeneous features and samples have shown boosted performance in both classification and regression tasks [2]–[5], [27]–[29]. Inspired from it, we develop a relational regularized sparse learning method, which take their feature and subject inherent information into consideration. Specifically, feature and subject relational information are integrated in a least squares regression (LSR) framework using $l_{2,1}$-norm to investigate the underlying relationship. Using the relational characteristics and $l_{2,1}$ norm on the weight coefficients, a novel objective function is devised

Manuscript received September 29, 2016; accepted December 8, 2016. Date of publication January 16, 2017; date of current version March 15, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61402296, Grant 61571304, and Grant 61427806, in part by the National Key Research and Develop Program under Grant 2016YFC0104703, in part by the Key Project of Department of Education of Guangdong Province under Grant 2014GKXM052, in part by the Shenzhen Key Basic Research Project under Grant JCYJ20150525092940986 and Grant JCYJ20150525092940982, in part by the Guangdong Medical under Grant B2016094, and in part by the National Natural Science Foundation of Shenzhen University under Grant 2016077. This paper was recommended by Associate Editor Y. Xiang. *(Corresponding author: Dong Ni.)*

The authors are with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: nidong@szu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2016.2644718

2168-2267 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

to identify the informative feature for joint regression and classification.

In spite of these attempts, it is known that discriminative learning [30]–[36] via distance expanding among various classes fails to be considered in AD diagnosis. Intuitively, this technique can achieve a better classification performance than the conventional sparse learning techniques. For example, feature selection based on a discriminative LSR (DLSR) model proposed by Xiang *et al.* [31] introduced an ε-dragging strategy via slack variable ε to expand the distances and integrate geometrical information between data points. In [33], a marginal scalable DLSR (MSDLSR) learning for homogeneous feature selection is proposed to improve the DLSR method. Joint structured sparsity norms, row sparsity, and column sparsity are integrated in the MSDLSR framework. In [35], the DLSR method is further extended to incorporate the structured sparsity and multimodal information. This method was shown to outperform the traditional DLSR method since it incorporated distance learning. Inspired by [31] and [33]–[35], a new DLSR framework for feature selection is developed to take the internal relational information in the observations into account. Different from previous work [31], [33]–[35], we incorporate relational information for regularization. Specifically, we first expanded class distance via ε-dragging method (i.e., the DLSR framework) to discriminatively learn the label characteristics, and then embedded the relation information into the DLSR framework. The motivation behind this is that the DLSR enables to effectively select the informative feature in a discriminative way, while the regularization terms enable to impose relation information for performance boosting.

Apart from the above-mentioned approaches, there are numerous attempts to develop feature selection techniques for joint regression and classification in AD diagnosis and prognosis [2], [4], [5]. For example, the relationship learning-based methods utilizing Laplacian score and Fisher's score as the selection criteria have attracted numerous interests [4], [5], [29]. Recently, great success has been witnessed to find the informative feature jointly via simultaneous multitask learning. To our best knowledge, previous methods usually conducted feature selection first, and then built regression or classification models. Different from the previous study, our joint AD/MCI regression and classification framework is developed in a discriminative way using relational information.

In this paper, we develop a feature selection method for joint regression and classification via discriminative sparse learning and relational regularization. We propose a novel loss function, which not only expands the distance to get the geometrical information, but also makes use of the inherent information in the observations. The relation information contained in the loss function is imposed to preserve the similarity. The joint similarity and discriminative learning in the novel designed loss function can enhance the diagnosis performance. The experimental results on ADNI baseline dataset with 805 subjects show the efficacy of the proposed brain disease diagnosis and prognosis method. The achieved encouraging classification and regression performance demonstrates the superiority and advantages of our proposed simple yet effective method.

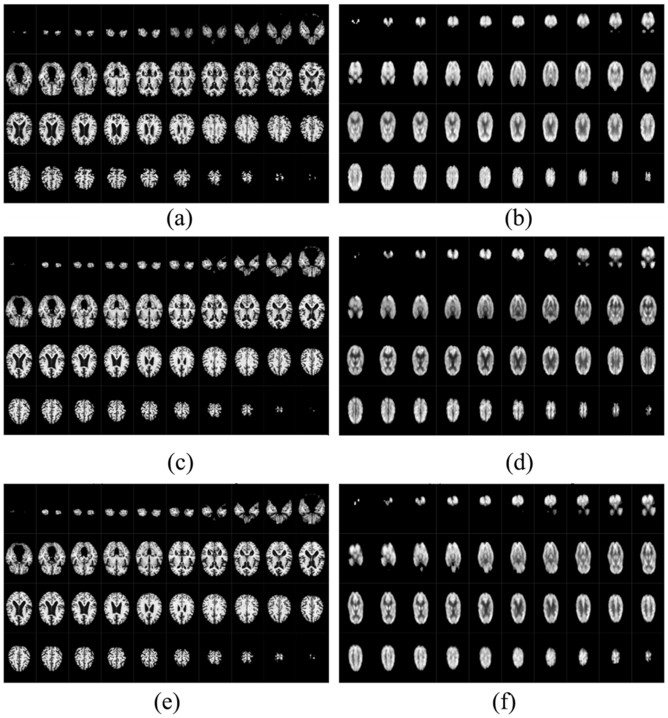


Fig. 1. Typical sample slices after preprocessing. (a) AD of MRI modality. (b) AD of PET modality. (c) MCI of MRI modality. (d) MCI of PET modality. (e) NC of MRI modality. (f) NC of PET modality.

II. Methodology

Fig. 2 illustrates our proposed method for joint AD/MCI classification and regression. We can see from Fig. 1 that our method includes three main contributions: 1) multimodal feature extraction for joint prediction and regression; 2) discriminative learning by ε-dragging strategy; and 3) relational regularization by feature and subject similarity learning. We will explain all the steps in detail in the following sections.

A. Subjects and Image Processing

We select the public available ADNI dataset to illustrate the proposed method. We adopt the ADNI general eligibility criteria detailed in the following. ADNI subjects are aged between 55 years and 90 years. We adopt the same general inclusion/exclusion criteria, namely: 1) the range of MMSE scores of healthy subjects is 24–30, nondemented, nondepressed, and non-MCI; and 2) MMSE of MCI subjects is also ranged from 24 to 30, which meets the National Institute of Neurological and Communicative Disorders and Stroke, and the AD and Related Disorders Association criteria for probable AD. The written informed consent are given from all study subjects at the time of enrollment to collect imaging and genetic sample by completing questionnaires approved by each participating site institutional review board.

We first perform preprocessing for all the studied subjects. Specifically, the T1-weighted MRI brain images was

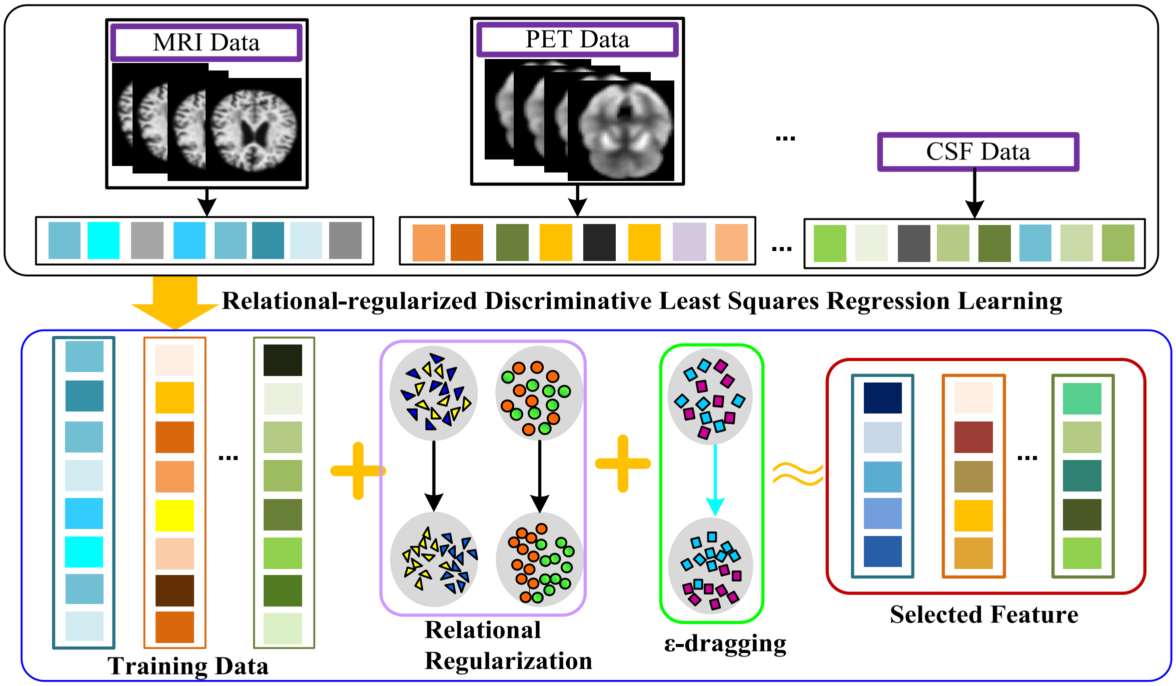

MRI Data
PET Data
CSF Data
...
Relational-regularized Discriminative Least Squares Regression Learning
Training Data
Relational Regularization
ε-dragging
Selected Feature
+
+
≈

Fig. 2. Illustration of flowchart of the proposed method (Note that large square in different colors denotes different features in various data, triangle means subject–subject relationship in two different classes, circle represents feature–feature relationship in two different classes, and rectangle means the label space in two different classes).

first registered by HAMMER [37], and then applied intensity normalization, skull-stripping, and cerebellum removing. The skull stripping is performed to clean skull. The cerebellum is removed by mapping a labeled atlas to the skull-stripped image. The segmentation is then applied to segment the brain images into three tissues [i.e., white matter, gray matter (GM), and CSF] by the FAST method [38]. After segmentation, the brain image is nonlinearly registered with a HAMMER tool [37]. In fact, we perform a two-step registration. Namely, the first step is the linear alignment, and the second step is the nonlinear registration from HAMMER, which is a standard pipeline in various publications and widely applied [2]–[5], [13], [29]. The template used in our scheme is Jacob template [2]–[5], [13], [29]. The total region of interest (ROI) number in our template is 93. For the preprocessing PET data, we further apply linear registration with 9 degree of freedom from T1 to PET, so that the ROIs in T1 space could be propagated to PET data. Mean intensity of the aligned ROIs in PET images is used as features. Each subject is divided into multiple 93 ROIs by atlas warping, and volume of GM tissue of each ROI is extracted as a feature. To visualize the processing, Fig. 1 provides the typical MRI and PET slices belonging to different classes (AD, MCI, and NC). In fact, our preprocessing and feature extraction from ROI regions have been widely applied in the literature. Similar to the study in [2], we normalized the features.

B. Notation

To be consistent in this paper, matrices are denoted by the capital bold letters, vectors are represented by small bold letters, and the regular variables are denoted by nonbold letters. For F-dimensional feature vector in the baseline data, the data of S subjects is represented by $\mathbf{X} \in \mathbb{R}^{S\times F}$. In this paper, we concatenate all the modalities together, as a result, the total dimension of feature vector is 189 (note that the feature dimension for CSF modality is 3 in this paper). The uth row vector and vth column vector of $\mathbf{X}$ are represented by $\mathbf{x}_u$ and $\mathbf{x}^v$, respectively. C categories of labels for S subjects are denoted as $\mathbf{Y} = \mathbb{R}^{S\times C}$. $\mathbf{W} \in \mathbb{R}^{F\times C}$ is the set of weight matrices to transform from the original features to label space. Our objective is to develop the classification and regression model to identify the labels and clinical scores using multimodal data. To establish the learning model for the data analysis task, the response variables $\mathbf{Y}$ are assigned in each class label for classification. Namely, in matrix $\mathbf{X}$, each subject's features are put as a row to get the weight coefficients in each $\mathbf{W}$. Accordingly, we are able to obtain the corresponding clinical scores in each $\mathbf{Y}$. For classification task, it is expected that the distance between data points in the same class is as small as possible, while the distance between different classes is as large as possible after mapping.

C. Feature Learning and Modeling

For feature selection, sparse learning is quite effective to address this regression problem. It is known that linear regression was one of the simplest and widely used regression analysis methods. The popularly applied linear regression via regularization can be addressed as an LSR optimization problem as

$$\min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_2^2 \qquad (1)$$

where $\|\blacksquare\|_F^2$ denotes the Frobenius norm of the matrix, $\|\blacksquare\|_2^2$ is l_2 norm, and λ is a regularization parameter.

In a continuous observation, using a class label is desired within two-class problem. It is noted that a label vector is manually assigned with "+1/−1" for two-class problems. For AD diagnosis, it is always desirable that the geometrical distance in different classes should be as large as possible. For this classification task, integrating geometrical information via distance learning is quite effective. However, the previous LSR does not include any geometrical information. To address it, we investigate expanding the distance for feature selection among different classes. Since the LSR framework lacks of the characteristics of discriminative learning, it is desirable to incorporate this geometrical information for learning in both classification and regression tasks. For this purpose, the conventional least squares model is extended to DLSR for feature selection. To this end, the DLSR framework is designed to consider the geometrical distance among different classes [31]. DLSR without bias parameter is defined as

$$\min_{\substack{\mathbf{W},\mathbf{P}\\ \text{s.t.}\mathbf{P}\geq 0}} \|\mathbf{Y}+\mathbf{B}\odot\mathbf{P}-\mathbf{X}\mathbf{W}\|_F^2+\lambda\|\mathbf{W}\|_2^2 \tag{2}$$

where $\odot$ is a Hadamard product operator of matrices, $\mathbf{P}\in\mathbb{R}^{S\times C}$ is a non-negative matrix and its element $p_{i,j}$ is a positive slack value $\varepsilon_{i,j}$ on the ith subject and jth class obtained by learning, $\mathbf{B}\in\mathbb{R}^{S\times C}$ is a constant matrix defined as

$$B_{ij}=\begin{cases}+1, & \text{if } \ y_i=j\\ -1, & \text{otherwise.}\end{cases} \tag{3}$$

Each element in $\mathbf{B}$ is a constant corresponding to the geometric direction. Namely, the dragging elements are recorded. Since our learning technique includes multimodal data, we further extend the DLSR framework to incorporate it as

$$\min_{\substack{\mathbf{W},\mathbf{P}\\ \text{s.t.}\mathbf{P}\geq 0}} \sum_{m=1}^{M}\|\mathbf{Y}+\mathbf{B}\odot\mathbf{P}-\boldsymbol{x_m}\mathbf{W}\|_F^2+\lambda\|\mathbf{W}\|_{2,1} \tag{4}$$

where $\|\blacksquare\|_{2,1}$ is a $l_{2,1}$-norm, which is defined as $\|\mathbf{W}\|_{2,1}=\sum_{i=1}^{F}\|\boldsymbol{w}_i\|_2$, $\boldsymbol{w}_i$ is the ith row vector of $\mathbf{W}$. DLSR uses a $l_{2,1}$ norm loss function to avoid the impact of the outliers. DLSR can uncover the correlation among different features and jointly select features for multiple tasks. Moreover, the DLSR was designed to achieve the minimal difference between $\mathbf{Y}$ and $\mathbf{B}\odot\mathbf{P}$. It can also enlarge the between-class distances among subjects in different classes. That is, we expect to obtain the optimal $\mathbf{P}$ so that the value of $p_{i,j}$ should be $1+\varepsilon_{i,j}$ for the sample grouped into the same class and $-\varepsilon_{i,j}$ for the sample grouped into different class. Accordingly, the between-class distance (i.e., the distance between two classes) will be enlarged, which enables to achieve discriminative learning of different classes based on multimodal data.

However, the DLSR did not consider any relationships among the observations, which had been shown to boost the AD diagnosis performance [4], [5], [29]. In this paper, predefined brain areas of ROIs are functionally or structurally related to each other, so it is natural to expect that there exist relations among features and subjects. Different from the previous study, novel regularization terms are devised for effective utilization of inherent information. Based on the conclusion that a well-defined regularization term may produce a generalized solution to boost the classification performance [4], [5], [29], in this paper, we first formulate regularizations to include two kinds of relation information inherent in observations under the assumption that, if some features (or subjects) are related to each other, the same (or similar) relation should be preserved in the corresponding weight coefficients. Since the least squares solution still has the over-fitting issue, we use the regularization method to find a more generalized solution. In the literature, numerous regularization terms have developed to find a generalized solution from a machine learning point of view. For our task, we devise two novel regularizers via Laplacian matrices and graphs to obtain the similarity in the local structures. In the rest of this section, we will explain it in details and discuss all its characteristics. We then integrate these regularization terms into the discriminative learning framework (i.e., DLSR). Aiming at boosting AD diagnosis performance by identifying the most useful features based on the optimal regression matrix $\mathbf{W}$, we devise our loss function as

$$\min_{\substack{\mathbf{W},\mathbf{P}\\ \text{s.t.}\mathbf{P}\geq 0}} \sum_{m=1}^{M}\|\mathbf{Y}+\mathbf{B}\odot\mathbf{P}-\boldsymbol{x_m}\mathbf{W}\|_F^2+\lambda\|\mathbf{W}\|_{2,1}+\lambda_1 R_1(\mathbf{W}) + \lambda_2 R_2(\mathbf{W}) \tag{5}$$

where R_1 is a feature–feature relation-based regularization term and λ_1 is its regularization parameter, R_2 is a subject–subject relation-based regularization term and λ_2 is its regularization parameter.

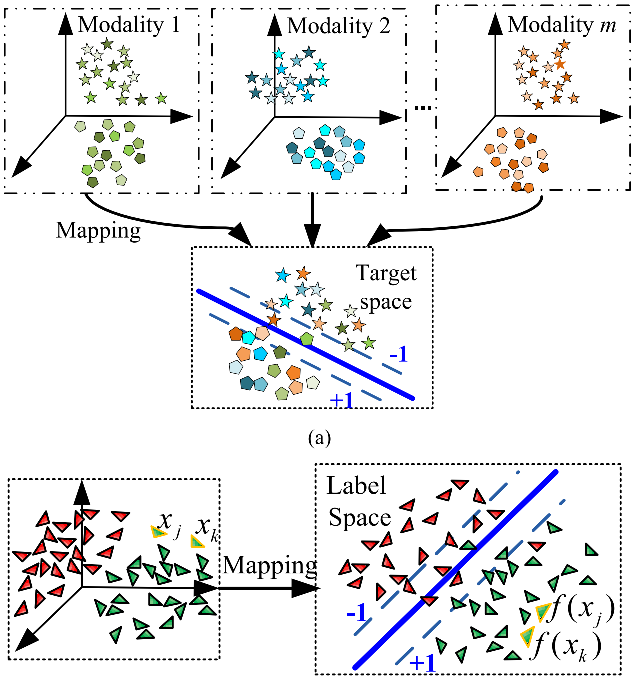
Fig. 3 shows the relationship learning to map the original representation to the target space. The motivation for the mapping is that ROIs extracted from a predefined brain area are functionally or structurally related with each other since they are from the same subject. Also, the subjects from the same classes preserve the similarity information. If two features are highly related to each other, the similar relation is preserved using the respective weight coefficients, and hence feature vectors of the weight coefficients are explored for regression. Due to the correlation of ROIs, incorporating the existing multirelation information is highly desirable. Specifically, each modality's relation is imposed as the relation between columns (i.e., the features) of $\mathbf{X}$ to be reflected in the relation between the corresponding rows in $\mathbf{W}$. When representing the same subject, the feature in the target space is similar to each other after mapping. To incorporate the relation, we use the widely used graph Laplacian. Specifically, let $\mathbf{F}=\{f_{uv}\}\in\mathbb{R}^{F\times F}$ measure the similarity between the uth feature and vth feature of $\mathbf{X}$ in the original feature space, we use a heat kernel defined as

$$f_{uv}=\exp\Big(-\big\|\mathbf{x}^u-\mathbf{x}^v\big\|_2^2\Big) \tag{6}$$

where $\mathbf{x}^u$ is the uth column of the input data $\mathbf{X}$. Based on the similarity, we develop the first feature–feature relation-based regularization term as

$$R_1(\mathbf{W})=\sum_{u,v=1}^{F} f_{uv}\big\|\boldsymbol{w}^u-\boldsymbol{w}^v\big\|_2^2 \tag{7}$$

where $\mathbf{w}^u$ is the uth row of $\mathbf{W}$.



(b)

Fig. 3. Relationship learning in terms of feature and subject information to be mapped into the target domain. (a) Mapping via feature relationship in different modalities (Note that different green colors mean the features in modality 1, different blue colors denote the features in modality 2, and different brown colors represent the features in the modality *m*. The shape of pentagon denotes the positive training class and the shape of pentagram represents the negative training class). Each class has multiple modalities. After mapping (learning), it is easy to separate different classes in the target space since features of the same class become closer, and vice versa. (b) Mapping via subject relationship, where red is the positive training class and green is the negative training class, *j*th and *k*th subject is mapped into the label space.

In addition, we incorporate subject–subject relation graph as the second regularization term. We know that if the subjects are similar to each other, their corresponding labels and clinical scores should be also similar to each other. Subject and subject relation is reserved in the element-wise differences of weighting coefficients. Similar to the previous term, we use a heat kernel to exploit the subject–subject similarities and define the similarity between the *j*th and *k*th subject as

$$\phi_{jk} = \exp\left(-\left\|\mathbf{x}_j - \mathbf{x}_k\right\|_2^2\right) \tag{8}$$

where $\mathbf{x}_j$ is the *j*th row of input $\mathbf{X}$. Here, subject–subject relation regularizer is defined as

$$R_2(\mathbf{W}) = \sum_{j,k=1}^{S} \phi_{jk}\left\|\mathbf{x}_j\mathbf{W} - \mathbf{x}_k\mathbf{W}\right\|_2^2. \tag{9}$$

Finally, we propose a discriminative sparse learning model, along with a least-squares loss function in order to select the most relevant and discriminant features correlated with the actual clinical scores. The loss function would control the prediction error, while the sparsity assumption leads to the least number of contributing features. We formulate the proposed objective function in a general form as follows:

$$\sum_{m=1}^{M}\|\mathbf{Y}+\mathbf{B}\odot\mathbf{P}-\mathbf{x}_m\mathbf{W}\|_F^2+\lambda_1\sum_{u,v=1}^{F} f_{uv}\left\|\boldsymbol{w}^u-\boldsymbol{w}^v\right\|_2^2 + \lambda_2\sum_{j,k=1}^{S}\phi_{jk}\left\|\mathbf{x}_j\mathbf{W}-\mathbf{x}_k\mathbf{W}\right\|_2^2+\lambda\|\mathbf{W}\|_{2,1}. \tag{10}$$

Because of $l_{2,1}$ norm in (10), the optimal $\mathbf{W}$ contains some zero or close to zero row vectors. $l_{2,1}$ norm is convex, and thus it has a global minimum solution. The corresponding features with nonzero weights are important in both classification and regression tasks. Hence, we rank l_2 norm value of each row vector in $\mathbf{W}$, i.e., $\|w^k\|_2^2$, $k = 1, \ldots, K$, and then select the features corresponding to the top-ranked rows. Note that $l_{2,1}$-norm calculates the sum of the l_2-norm of each row of $\mathbf{W}$ to render many rows to be zeros and desirable for feature selection. Also, the nonzero rows in $\mathbf{W}$ are corresponding to the informative features in subsequent learning models. We incorporate two regularizers composed of feature and subject relationships into a general loss function, and we call this algorithm as relational-regularized DLSR learning (R2DLSR). We preserve the internal information in terms of feature and subject relations in the loss function. Both the similarity information and discriminative learning are integrated to expand the solution. Our method is the first study to integrate the feature and subject information as two relational regularizers in a discriminative learning framework, which is not easy to solve in the current sparse models. We can solve the optimization problem in the loss function in an alternative way [39]. We illustrate the optimization steps in detail in the next section. By taking advantage of both local structural similarity and relational information inherent in data, our method can identify the most informative features for joint regression and classification.

III. Optimization Algorithm

Although our objective function is convex, it is difficult to be solved because regularization terms are based on the nonsmooth sparsity-inducing norms in the objective function. $l_{2,1}$-norm minimization is more challenging to solve than l_1-norm minimization problem. There are some previous optimization algorithms [4], but they are too computationally intensive to solve our problem. For this purpose, an efficient iterative algorithm is developed.

In the similarity measurement, Laplacian graph at each time point is built based on a diagonal matrix and formulated as: $\boldsymbol{D}_f = f_{uv}, \boldsymbol{D}_s = \phi_{jk}$, let $\boldsymbol{S}_f$ and $\boldsymbol{S}_s$ be the summation of the diagonal entry of $\boldsymbol{D}_f$ and $\boldsymbol{D}_s$, respectively, then the graph Laplacian $\boldsymbol{L}_f$ for the feature and subject space are: $\boldsymbol{L}_f = \boldsymbol{D}_f - \boldsymbol{S}_f$ and $\boldsymbol{L}_s = \boldsymbol{D}_s - \boldsymbol{S}_s$, respectively. The regularization term $R_1(\mathbf{W})$ can be reformulated as $R_1(\mathbf{W}) = \mathrm{Tr}(\mathbf{W}^{\boldsymbol{T}}\boldsymbol{L}_f\mathbf{W})$, where Tr(■) is trace function. Similarly, we can have $R_2(\mathbf{W}) = \mathrm{Tr}((\mathbf{X}\mathbf{W})^{\boldsymbol{T}}\boldsymbol{L}_s\mathbf{X}\mathbf{W})$. To solve the optimization

Algorithm 1 Iterative Algorithm to Solve the Optimization Problem in the Relational-Regularized Discriminative Learning for Feature Selection

```
Input:  Baseline multimodal training data of $S$ subjects and $F$ dimensional features: $\mathbf{X} \in \mathbb{R}^{S \times F}$
        $C$ dimensional labels and clincial score vector of $S$ subjects $\mathbf{Y} \in \mathbb{R}^{S \times C}$
        Parameters: regularization paramters and iteration times
        Set iteration $r$=0 and initialize $\mathbf{W} \in \mathbb{R}^{F \times C}$ according to the linear model; Initilialize P;
        Repeat
            Calculate $\boldsymbol{L_f}$, $\boldsymbol{L_s}$, and $\boldsymbol{L_D}$, according to the above definitions;
            Let $\widehat{\mathbf{Y}} = \mathbf{Y} + \mathbf{B} \odot \mathbf{P}$
            Update $\mathbf{W}$ by solving the simple problem in Eq.(14): $\mathbf{W}_{r+1} = \mathbf{A}_{\boldsymbol{r}}^{\mathbf{-1}}\mathbf{Q}_r$;
            Let $\mathbf{T} = \mathbf{X}W - \mathbf{Y}$
            Update P by $P_{r+1} = max(\mathbf{B} \odot \mathbf{T}, \mathbf{0})$,
            $r$=$r$+1;
        until convergence (i.e., $r$<=50 or $\left\|\widehat{\mathbf{W}} - \mathbf{W}\right\|_{\mathbf{2}}$<$\mathbf{10^{-6}}$ )
Output: Selected top ranked features.
```

problem, the objective can be reformulated as

$$\min_{\substack{\mathbf{W},\mathbf{P}\\ \text{s.t.}\mathbf{P}\geq 0}} \sum_{m=1}^{M} \|\mathbf{Y} + \mathbf{B} \odot \mathbf{P} - \boldsymbol{x_m}\mathbf{W}\|_F^2 + \lambda \text{Tr}\Big(\mathbf{W}^{\boldsymbol{T}}\boldsymbol{L_D}\mathbf{W}\Big) + \lambda_1 \text{Tr}\Big(\mathbf{W}^{\boldsymbol{T}}\boldsymbol{L_f}\mathbf{W}\Big) + \lambda_2 \text{Tr}\Big((\mathbf{X}\mathbf{W})^{\boldsymbol{T}}\boldsymbol{L_s}\mathbf{X}\mathbf{W}\Big) \quad (11)$$

where $\boldsymbol{L_D}$ is the Laplacian graph built based on a diagonal matrix of $\mathbf{W}$.

Let $\widehat{\mathbf{Y}} = \mathbf{Y} + \mathbf{B} \odot \mathbf{P}$, we can get the optimal solution of $\mathbf{W}$ with fixed $\mathbf{P}$, which is obtained by taking the derivative of the objective function with respect to $\mathbf{W}$ and set it to 0. After derivation, we can have the following solution:

$$\mathbf{X}^{\boldsymbol{T}}\mathbf{X} - \mathbf{X}\mathbf{Y}^{\boldsymbol{T}} + \lambda\boldsymbol{L_D}\mathbf{W} + \lambda_1\boldsymbol{L_f}\mathbf{W} + \lambda_2\mathbf{X}^{\boldsymbol{T}}\boldsymbol{L_s}\mathbf{X}\mathbf{W} = 0. \quad (12)$$

We can further formulate this equation as

$$\Big(\mathbf{X}^{\boldsymbol{T}}\mathbf{X} + \lambda\boldsymbol{L_D} + \lambda_1\boldsymbol{L_f} + \lambda_2\mathbf{X}^{\boldsymbol{T}}\boldsymbol{L_s}\mathbf{X}\Big)\mathbf{W} = \mathbf{X}\mathbf{Y}^{\boldsymbol{T}}. \quad (13)$$

This equation is solvable in the closed form and rewritten as

$$\boldsymbol{AW} = \boldsymbol{Q} \quad (14)$$

where $\mathbf{A} = \mathbf{X}^{\boldsymbol{T}}\mathbf{X} + \lambda\boldsymbol{L_D} + \lambda_1\boldsymbol{L_f} + \lambda_2\mathbf{X}^{\boldsymbol{T}}\boldsymbol{L_s}\mathbf{X}$, $\mathbf{Q}$ is $\mathbf{X}\widehat{\mathbf{Y}}^{\boldsymbol{T}}$, and $\mathbf{W}$ can be obtained by solving this equation.

Similarly, we can get the optimal $\mathbf{P}$ by fixing $\mathbf{W}$. When $\mathbf{W}$ is fixed and multimodal data is integrated in $\mathbf{X}$, $\mathbf{P}$ is solved by optimizing the following algorithm:

$$\min_{\substack{\mathbf{P}\\ \text{s.t.}P\geq 0}} \|\mathbf{Y} + \mathbf{B} \odot \mathbf{P} - \mathbf{X}\mathbf{W}\|_F^2. \quad (15)$$

Let $\mathbf{T} = \mathbf{X}W - \mathbf{Y}$, $\mathbf{P}$ can be obtained by solving the problem

$$\min_{\substack{\mathbf{P}\\ \text{s.t.}\mathbf{P}\geq 0}} \|\mathbf{T}{-}\mathbf{B} \odot \mathbf{P}\|_F^2. \quad (16)$$

Based on the matrix theory, the Frobenius norm of the matrix is treated as an element by element case, we simplify the problem into $S \times C$ subproblems, and sovle the following problem to get P_{ij}:

$$\min_{\substack{P_{ij}\\ \text{s.t.}P_{ij}\geq 0}} \left\|T_{ij}{-}B_{ij} \odot P_{ij}\right\|_F^2. \quad (17)$$

Since $B_{ij} = 1$, we can get $\|T_{ij} - B_{ij}P_{ij}\|^2 = \|B_{ij}P_{ij} - T_{ij}\|^2$, and hence the optimization problem is reformulated as

$$P_{ij} = \max\big(B_{ij}T_{ij}, 0\big). \quad (18)$$

As a result, we can obtain $\mathbf{P}$ by

$$\mathbf{P} = \max(\mathbf{B} \odot \mathbf{T}, 0). \quad (19)$$

Based on the aforementioned mathematical derivation, we can solve the optimization problem in an iterative way. Since $\boldsymbol{L_f}$, $\boldsymbol{L_s}$, and $\boldsymbol{L_D}$ are obtained from $\mathbf{W}$ and dependent on $\mathbf{W}$, an iterative optimization is proposed to obtain the global solutions $\mathbf{W}$ efficiently. The solution of obtaining $\mathbf{W}$ is summarized in Algorithm 1. The iterative optimization method updates $\mathbf{W}$ until the convergence of objective function has reached. The detail of the convergence analysis can be found in the Appendix. Based on the obtained $\mathbf{W}$, we can obtain the selected top ranked features via the relational regularized discriminative learning.

IV. Experimental Results

A. Experimental Setup

To show the effectiveness of our proposed method, we carry out our experiments based on the ADNI database (http://adni.loni.usc.edu/). A total of 805 subjects [including 226 AD patients, 393 MCI, and 186 normal controls (NCs)] are used from ADNI database. Three hundred and ninety three MCI patients are comprised of 167 progressive MCI (pMCI) patients (who will progress to AD in 18 months) and 226 stable MCI (sMCI) patients (whose symptom are stable and will not progress to AD in 18 months). Out of these 805 subjects, more than half of them do not have PET data, and hence the single PET modality does not use for performance comparison. We perform the experiments and construct a new matrix $\mathbf{X}$ including all modalities such as, MRI, PET, and CSF (MPC) to enhance the diagnostic accuracy via complementary features. Since multiple modalities get better performance than single modality (i.e., MRI) due to more available information, we only report the performance with multimodal data (i.e., MPC) here. We obtain the average GM volumes of each ROI extracted from multimodal data as feature. In this paper, we jointly classify different classes (e.g., AD, MCI, and NC) and predict the four clinical scores (e.g., ADAS-Cog, MMSE, CDRSOB, and CDR-Global).

We adopt a tenfold cross-validation algorithm to assess both classification and regression performance. Namely, all samples are divided into ten portions (each portion with a roughly equal size), and samples in one portion are successively chosen as the testing data, and the rest are utilized as the training data. The LibSVM classifier is adopted to train the support vector regression (SVR) and support vector classification (SVC) model by sigmoid kernel [40]. We perform another fivefold inner cross-validation to choose the parameters using a line search method in the prespecified range (e.g., $\lambda \in \{10^{-10}, \ldots, 10^{10}\}$) in the LIBSVM toolbox in each fold. After cross-validation, we choose the parameters with the best performance. To prevent any bias in data partitioning, we

TABLE I
CLASSIFICATION RESULTS (%)

Classifier	Method	ACC	SEN	SPEC	AUC	PPV	NPV	F1	p-Value
AD vs. NC	Lasso	88.87+ 5.63	92.07+ 5.95	84.63+ 6.07	94.75+3.69	87.36+6.71	92.07+5.95	89.60+5.98	<0.0001
	M3T	92.50+ 3.82	95.22+ 3.84	89.37+ 6.69	96.84+2.47	91.35+5.68	95.22+3.84	93.15+3.68	<0.0001
	M3TFS	90.53+ 4.06	93.21+ 5.85	88.28+ 6.59	96.49+2.39	89.93+6.90	93.21+5.85	91.29+4.18	<0.0001
	LSR	90.56+ 4.58	92.46+ 6.50	89.41+ 7.40	96.56+2.50	90.73+7.20	92.46+6.50	91.30+4.41	<0.0001
	DLSR	90.78+ 4.27	93.50+ 5.44	88.21+ 6.74	96.45+2.44	89.94+6.82	93.50+5.44	91.51+4.53	<0.0001
	RLSR	92.73+ 3.60	96.72+ 4.89	87.44+ 9.16	96.76+2.56	90.58+5.61	96.72+4.89	93.39+3.51	<0.0001
	RDLSR	94.42+2.81	96.86+3.81	90.63+6.06	97.76+1.82	**92.48+4.03**	96.86+3.81	94.57+3.17	<0.0001
	R2DLSR	**94.68+2.93**	**97.90+2.89**	**91.38+5.91**	**97.92+1.88**	92.30+5.48	**97.90+2.89**	**94.93+3.35**	<0.0001
MCI vs. NC	Lasso	76.12+ 4.88	57.67+ 15.73	86.23+ 8.75	79.38+6.11	73.46+12.04	57.67+15.73	62.35+10.72	<0.0001
	M3T	72.86+ 5.74	54.54+ 7.68	83.86+ 7.28	78.97+6.93	66.11+13.15	54.54+7.68	59.09+7.40	<0.0001
	M3TFS	72.67+ 4.72	53.85+ 7.50	82.82+ 7.88	78.54+6.50	65.20+11.77	53.85+7.50	58.38+7.42	<0.0001
	LSR	73.99+ 5.04	57.62+ 9.21	83.17+ 4.73	78.64+6.72	65.77+10.16	57.62+9.21	61.12+8.72	<0.0001
	DLSR	74.46+ 5.32	58.42+ 6.82	83.74+ 5.69	80.56+6.56	67.07+11.51	58.42+6.82	62.13+7.38	<0.0001
	RLSR	75.07+ 4.42	55.71+ 10.63	85.73+ 6.81	79.04+6.53	70.05+11.93	55.71+10.63	60.98+9.43	<0.0001
	RDLSR	79.68+6.28	63.33+8.22	86.13+7.65	**83.59+7.52**	74.24+10.77	63.33+8.22	68.10+8.40	<0.0001
	R2DLSR	**80.32+6.04**	**64.35+10.65**	**86.67+6.62**	82.60+7.31	**74.71+11.08**	**64.35+10.65**	**68.28+8.30**	<0.0001
pMCI vs. sMCI	Lasso	67.20+ 5.63	42.08+ 29.71	83.03+ 16.25	66.53+11.37	58.41+32.92	42.08+29.71	44.83+28.96	<0.0001
	M3T	67.22+ 5.98	44.85+ 16.16	83.73+ 10.36	68.26+9.30	70.42+15.70	44.85+16.16	51.87+14.99	<0.0001
	M3TFS	66.96+ 4.04	50.88+ 10.43	78.87+ 7.83	65.22+6.30	64.57+8.84	50.88+10.43	56.01+7.67	<0.0001
	LSR	65.97+ 7.05	40.76+ 22.13	84.07+ 8.92	65.70+8.14	64.46+26.35	40.76+22.13	46.78+22.99	<0.0001
	DLSR	68.19+ 3.66	47.92+ 9.06	83.63+ 8.45	69.78+6.08	68.52+13 .28	47.92+9.06	55.53+8.35	<0.0001
	RLSR	67.27+ 7.83	40.27+ 28.53	85.59+ 12.40	64.89+10.03	58.95+33.27	40.27+28.53	44.35+29.05	<0.0001
	RDLSR	72.77+5.55	46.20+21.56	87.54+8.35	74.53+10.76	71.33+27.50	46.20+21.56	54.46+23.01	<0.0001
	R2DLSR	**74.58+6.12**	**51.31+23.81**	**88.72+7.98**	**74.58+9.53**	**71.65+27.05**	**51.31+23.81**	**58.01+23.36**	<0.0001

repeat the process 10 times. We report the final performance after averaging the results of the repeated cross-validations.

For each binary classification task, the extracted ROI features are mapped to the target space to get a transformation matrix for support vector machine (SVM) learning. In each set of experiment, the regression performance is measured in terms of correlation (Corr) and normalized root mean square error (NRMSE) of the actual and predicted clinical scores [4]. The same definition of the quantitative measurements as [6] for classification is utilized to evaluate the diagnosis performance. The performance metrics include classification accuracy (the disease status of subjects is correctly classified as the actual disease status of the subjects in each class) (ACC), sensitivity (SEN), specificity (SPEC), balanced accuracy (BAC), positive predicted value (PPV), negative predictive value (NPV), and F1 measure (F1). The receiver operating characteristics curve (RoC) and the area under RoC (AUC) are also utilized as the performance metrics.

Our proposed R2DLSR feature selection method is adopted to select the most discriminative features. In the following section, RDLSR denotes our proposed method with only feature relation regularization, and RLSR denotes our proposed method without any discriminative learning. Also, we select the following methods for comparison since they are related with our main contribution: discriminative learning and relational regularization.

1) *Lasso [41]:* Lasso is one of the most widely used feature selection method.
2) *DLSR [31]:* Essentially, the proposed method is an extension of DLSR work and uses the same dragging technique.
3) *M3T [2]:* M3T is a relational-regularization related algorithm, which is a special case of the proposed method with the regularization terms setting to zeros.
4) *M2TFS [4]:* M2TFS fuses all the information using a multitask learning algorithm based on multimodal data, which is an enhanced feature selection of M3T via relational regularization.

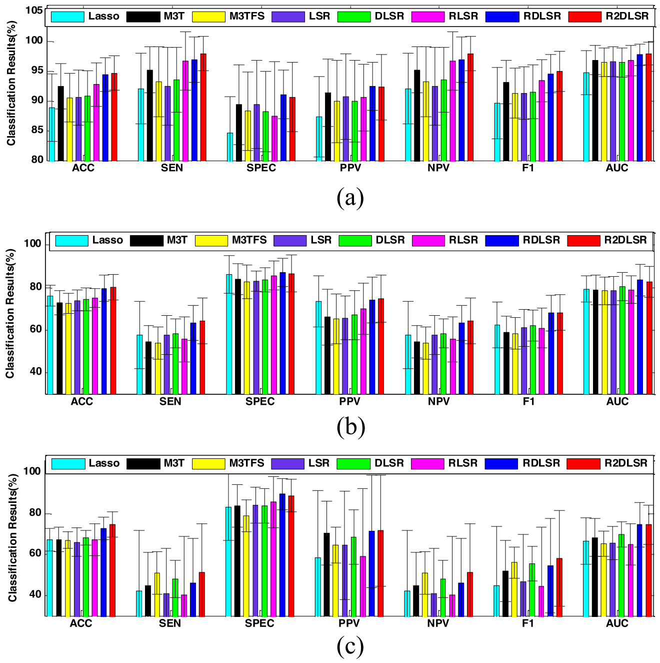
B. Classification Results

Table I shows the binary classification results (note that the boldface denotes the best performance in that column). From Table I, it is observed that the proposed algorithm gets quite appealing classification performance in three classification problems using seven performance metrics. Specifically, our R2DLSR method consistently outperforms the selected algorithms in AD versus MCI, MCI versus NC, and pMCI versus sMCI in most scenarios. The best classification performance in our method for AD versus NC, MCI versus NC, pMCI and sMCI are 94.68%, 80.32%, and 74.58%, respectively. The best classification performance is mainly obtained by our proposed R2DLSR method and superior to others in most cases. This is mainly because that relationship guided feature selection method boosts the AD/MCI classification performance. Compared with the traditional DLSR method, the proposed R2DLSR and RDLSR method with the regularization terms have achieved better performance. The superiority is mainly because our method makes use of the relational information for regularization. We can observe that our proposed R2DLSR method is quite remarkable for AD/MCI diagnosis since it consistently obtains the best performance in most scenarios. It can be concluded that the proposed regularization terms and discriminative learning by ε-dragging are quite promising to find the class-discriminative features. Overall, we can achieve the competing performance by the proposed algorithm.

Fig. 4 plots the results of various methods in three classification tasks. Fig. 5 shows the corresponding ROC curves. From Figs. 4 and 5, we can see that our R2DLSR method

TABLE II
REGRESSION RESULTS

	CORR				NRMSE			
Classifier	Method	ADAS-Cog	MMSE	CDRSOB	CDR-Global	ADAS-Cog	MMSE	CDRSOB	CDR-Global
AD vs. NC	Lasso	0.70+ 0.08	0.65+ 0.08	0.73+ 0.08	0.75+ 0.08	0.74+ 0.14	0.73+ 0.09	0.69+ 0.08	0.69+ 0.11
	M3T	0.70+ 0.09	0.67+ 0.08	0.76+ 0.06	0.76+ 0.06	0.70+ 0.15	0.72+ 0.08	0.66+ 0.07	0.68+ 0.09
	M3TFS	0.71+ 0.10	0.65+ 0.09	0.76+ 0.06	0.76+ 0.07	0.70+ 0.14	0.74+ 0.09	0.67+ 0.09	0.69+ 0.09
	LSR	0.70+ 0.08	0.65+ 0.09	0.75+ 0.06	0.77+ 0.06	0.71+ 0.14	0.74+ 0.08	0.67+ 0.07	0.68+ 0.09
	DLSR	0.71+ 0.08	0.68+ 0.09	0.78+ 0.04	0.77+ 0.06	0.69+ 0.14	0.70+ 0.08	0.64+ 0.08	0.67+ 0.09
	RLSR	0.73+ 0.07	0.70+0.08	0.78+ 0.05	0.77+ 0.06	0.69+ 0.13	0.70+ 0.07	0.66+ 0.08	0.67+ 0.10
	RDLSR	0.74+ 0.08	**0.72+0.08**	**0.80+ 0.04**	**0.80+ 0.05**	0.68+ 0.14	0.67+0.13	0.62+0.07	0.62+0.07
	R2DLSR	**0.75+0.07**	0.72+0.09	0.79+0.05	0.80+0.06	**0.68+0.12**	**0.67+0.08**	**0.63+0.07**	**0.63+0.08**
MCI vs. NC	Lasso	0.46+ 0.11	0.36+ 0.10	0.41+ 0.10	0.47+ 0.11	0.90+ 0.07	0.95+ 0.09	0.96+ 0.10	0.93+ 0.08
	M3T	0.49+ 0.11	0.37+ 0.10	0.38+ 0.10	0.46+ 0.13	0.88+ 0.08	0.95+ 0.09	0.97+ 0.11	0.91+ 0.08
	M3TFS	0.47+ 0.13	0.36+ 0.08	0.38+ 0.11	0.45+ 0.12	0.90+ 0.08	0.96+ 0.09	0.97+ 0.12	0.92+ 0.07
	LSR	0.45+ 0.13	0.37+ 0.10	0.38+ 0.12	0.45+ 0.13	0.89+ 0.08	0.95+ 0.08	0.97+ 0.11	0.93+ 0.07
	DLSR	0.49+ 0.14	0.39+ 0.08	0.40+ 0.10	0.52+ 0.11	0.88+ 0.07	0.94+ 0.08	0.96+ 0.12	0.88+ 0.07
	RLSR	0.51+ 0.12	0.42+ 0.08	0.44+ 0.08	0.48+ 0.12	0.87+ 0.08	0.93+ 0.09	0.94+ 0.10	0.91+ 0.07
	RDLSR	0.51+ 0.13	0.42+ 0.08	0.45+ 0.11	0.52+ 0.12	0.85+ 0.05	0.92+ 0.08	0.93+ 0.10	**0.85+ 0.08**
	R2DLSR	**0.51+0.12**	**0.44+0.07**	**0.45+0.11**	**0.53+0.12**	**0.85+0.08**	**0.92+0.08**	**0.93+0.10**	0.86+0.07
pMCI vs. sMCI	Lasso	0.34+ 0.13	0.30+ 0.10	0.30+ 0.07	NaN	0.94+ 0.13	0.98+ 0.07	0.99+ 0.13	0.35+ 0.95
	M3T	0.40+ 0.11	0.32+ 0.10	0.29+ 0.06	NaN	0.92+ 0.14	0.98+ 0.05	1.01+ 0.13	0.36+ 0.95
	M3TFS	0.40+ 0.11	0.34+ 0.06	0.31+ 0.09	NaN	0.93+ 0.14	0.98+ 0.06	1.01+ 0.13	0.36+ 0.95
	LSR	0.40+ 0.11	0.32+ 0.07	0.27+ 0.07	NaN	0.92+ 0.15	0.98+ 0.05	1.00+ 0.13	0.35+ 0.95
	DLSR	0.39+ 0.11	0.32+ 0.07	0.27+ 0.06	NaN	0.92+ 0.13	0.97+ 0.06	1.00+ 0.12	0.35+ 0.95
	RLSR	0.47+ 0.12	0.33+ 0.08	0.33+ 0.07	NaN	0.92+ 0.15	0.95+ 0.05	0.98+ 0.12	0.35+ 0.95
	RDLSR	0.48+0.09	0.38+0.08	0.33+0.11	NaN	0.92+0.13	0.94+0.06	0.98+0.11	0.35+0.95
	R2DLSR	**0.50+0.11**	**0.38+0.07**	0.**34+0.10**	NaN	**0.91+0.11**	**0.92+0.05**	**0.97+0.11**	**0.35+0.93**



(a)

(b)

(c)

Fig. 4. Bar chart of three binary classification results. (a) Classification results of AD versus NC. (b) Classification results of MCI versus NC. (c) Classification results of pMCI versus sMCI.

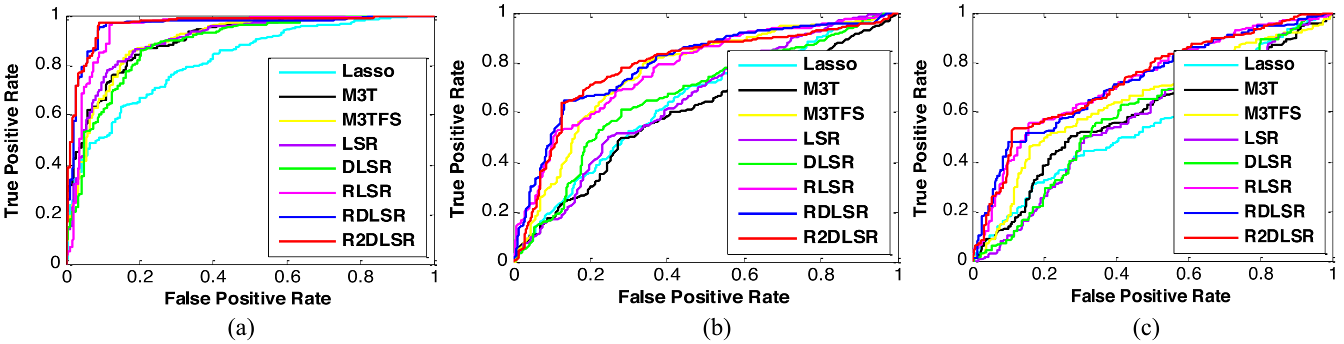
achieves better classification performance than the selected methods. Particularly, R2DLSR achieves the best sensitivity in AD versus NC classification, which indicates that our proposed R2DLSR method can effectively identify AD patients. High sensitivity values indicate high confidence in disease diagnosis, which is potentially quite useful in the real-world applications. Therefore, from a clinical point of view, R2DLSR is less likely to misdiagnose subjects with diseases, in comparison to those listed methods. From the RoC curves, R2DLSR is obviously superior to all other methods in terms of the three classification tasks. Also, our method with discriminative learning generally achieves better results compared with the rest methods, which indicates that the discriminative learning is able to boost the classification and regression performance. The primary explanation is that the rich geometric information of different classes is appealing for classification performance enhancement.

C. Regression Results

Table II summarizes the clinical scores, such as ADAS-Cog, MMSE, CDRSOB, and CDR-Global regression results. We also provide the scatter plots of the original result and prediction results of various scores, as can be seen in Fig. 6. We observe that our proposed method outperforms the competing methods in terms of the regression performance.

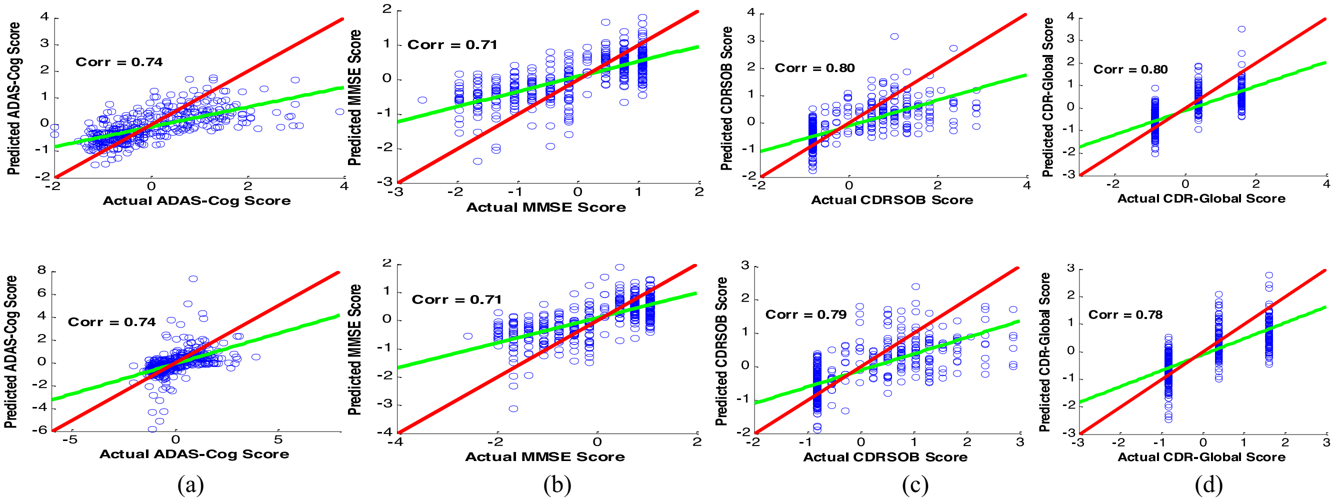
In the regression with MRI for AD versus NC, our proposed R2DLSR method shows the best CORR of 0.75 for ADAS-Cog and 0.72 for MMSE, 0.80 for CDRSOB and 0.80 for CDR-Global. In the regression for MCI versus NC, our method also achieves the best CORR of 0.51 for ADAS-Cog, 0.44 for MMSE, 0.45 for CDRSOB and 0.53 for CDR-Global, and the best NRMSE of 0.85 for ADAS-Cog, 0.92 for MMSE, 0.93 for CDRSOB and 0.86 for CDR-Global. For the case of pMCI versus sMCI, our method achieves the best CORR of 0.50 for ADAS-Cog, 0.38 for MMSE, and 0.34 for CDR-Global, and the best NRMSE of 0.92 for ADAS-Cog, 0.95 for MMSE, 0.98 for CDRSOB and 0.35 for CDR-Global. Note that CORR for CDRSOB is NAN for pMCI and sMCI case. The reason is that there are some zeros in these scores and unpredictable.

Fig. 6 shows the scatter plot of clinical scores prediction results using CORR metric. R2DLSR gets slightly better performance than RDLSR. The CORR results are quite high and have the same observations as summarized in Table II. In addition, Fig. 7 shows CORR and NRMSE results of



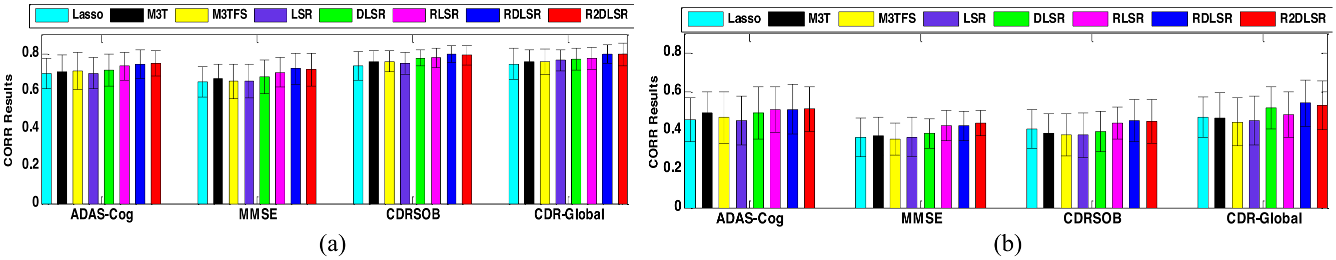
(a) (b) (c)

Fig. 5. RoC curves of (a) AD versus NC, (b) MCI versus NC, and (c) pMCI versus sMCI classification results.



(a) (b) (c) (d)

Fig. 6. Scatter plot of clinical scores prediction results. (a) ADAS-Cog. (b) MMSE. (c) CDRSOB. (d) CDR-Global prediction results. Note that the upper row is R2DLSR method and the bottom row is RDLSR method. Both methods achieve quite promising results.



(a) (b)

Fig. 7. CORR and NRMSE results of clinical score prediction results using various methods. (a) AD versus NC. (b) MCI versus NC. Note that the upper row is R2DLSR method and the bottom row is RDLSR method. Both methods achieve good results.

clinical score prediction results using various methods. From the comparison results, we can observe that our proposed method gets better results than the selected methods. Two regularization terms in the objective function result in better performance than a single regularization. Ultimately, the full utilization of the two relational characteristics achieves the best performance. The method with discriminative learning generally outperforms that without it, which implies that the discriminative learning is able to boost the regression performance by expanding the between-class distance.

D. Competing Method Comparison

We compare our proposed R2DLSR method with the state-of-the-arts competing method based on ADNI database. The AD versus NC, MCI versus NC, and pMCI versus sMCI classification results are shown in Tables III–V. R2DLSR achieves an accuracy of 94.68%, a sensitivity of 97.9%, and a specificity of 91.08% for AD versus NC classification, an accuracy of 80.32%, a sensitivity of 64.35%, and a specificity of 86.67% for MCI versus NC classification, and an accuracy of 74.58%, a sensitivity of 51.31%, and a specificity of 88.71% for pMCI versus sMCI classification. Note that our method outperforms the listed competing methods in most cases. In general, R2DLSR achieves better accuracy, sensitivity, specificity compared with the related algorithms in most scenarios. It is worth noting that our method takes advantage of discriminative learning and relational regularization in a unified framework, and mines the similarity information from features

TABLE III
ALGORITHM COMPARISONS FOR AD VERSUS NC CLASSIFICATION

Algorithm	Subject	Classifier	Modality	ACC (%)	SEN (%)	SPE (%)
Zhang *et al.*[3]	51AD+52NC	SVM	PET+MRI	90.6	90.5	90.7
Zhang *et al.*_[3]	51AD+52NC	SVM	PET+MRI+CSF	93.20	93.0	93.3
Hinrichs *et al.* [19]	48AD+66NC	SVM	MRI + PET	87.6	78.9	93.8
Hinrichs *et al.* [19]	48AD+66NC	SVM	MRI+PET+ CSF+APOE +Cognitive scores	92.4	86.7	96.6
Liu *et al.* [11]	198AD+229N C	SRC ensemble	MRI	90.8	86.32	94.76
Gray *et al.* [42]	37AD+35NC	Random forests	PET+MRI+ CSF+ genetic	89.0	87.9	90.0
Liu *et al.* [13]	51AD+52NC	SVM	PET+MRI	94.37	94.71	94.04
Min *et al.* [43]	97AD+128NC	SVM	MRI	91.64	88.56	93.85
Zhu *et al.* [4]	51AD+52NC	SVC	PET+MRI+CSF	95.9	95.7	98.6
Suk *et al.* [6]	93AD+101NC	MultiModal DBM, SVM	MRI+PET	95.35	94.65	95.22
Ours	226AD+393 MCI+ 186 NC	SVM,SVC	PET+MRI +CSF	94.68	97.9	91.08

TABLE IV
ALGORITHM COMPARISONS FOR MCI VERSUS NC CLASSIFICATION

Algorithm	Subject	Classifier	Modality	ACC (%)	SEN (%)	SPE (%)
Zhang *et al.* [3]	99MCI+52NC	SVM	PET+MRI+ CSF	76.4	81.8	66.0
Liu *et al.* [11]	225MCI+229NC	SRC ensemble	MRI	87.85	85.26	90.4
Gray *et al.* [42]	75MCI+35NC	Random forests	PET+MRI+ CSF+ genetic	74.6	77.5	67.9
Liu *et al.* [13]	99MCI+52NC	SVM	PET+MRI	78.8	84.85	67.06
Zhu *et al.* [4]	99MCI+52NC	SVC	PET+MRI+ CSF	82.0	98.0	60.1
Suk *et al.* [6]	204MCI+101NC	Multimodal DBM,SVM	PET+MRI	85.67	95.37	65.87
Ours	226AD+393 MCI+186 NC	SVM,SVC	PET+MRI +CSF	80.32	64.35	86.67

TABLE V
ALGORITHM COMPARISONS FOR pMCI VERSUS sMCI CLASSIFICATION

Algorithm	Subject	Classifier	Modality	ACC (%)	SEN (%)	SPE (%)
Gray *et al.* [42]	34pMCI +41sMCI	Random forests	PET+MRI+ CSF+genetic	58.0	57.1	58.7
Zhang *et al.*[44]	202MCI	SVM	PET+MRI +CSF	73.9	68.6	73.6
Liu et al. [13]	99MCI	SVM	PET+MRI	67.83	67.83	70.00
Min et al. [43]	117pMCI +117sMCI	SVM	MRI	72.41	72.12	72.58
Zhu *et al.* [4]	99MCI +52NC	SVR, SVC	PET+MRI +CSF	72.6	48.5	94.4
Suk *et al.* [6]	128pMCI +76sMCI	Deep learning	PET+MRI	75.92	48.04	95.23
Cuingnet *et al.*[45]	176pMCI +134sMCI	SVM	Single-atlas	70.40	57.00	78.00
Zhang and Shen [2]	43pMCI +48sMCI	SVM	PET+MRI +CSF	73.9	68.6	73.6
Davatzikos *et al.*[46]	69pMCI +170 sMC	SVM	CSF+MRI	55.8	94.7	37.8
Ours	226AD+393 MCI+186 NC	SVM,SVC	PET+MRI +CSF	74.58	51.31	88.71

and subjects using multiple modalities. Therefore, we achieve a better performance than the method with single technique.

V. DISCUSSION AND LIMITATIONS

In this paper, we demonstrate the efficacy of the proposed method via three binary classifications (e.g., the classifications of AD versus NC, MCI versus NC, and pMCI versus sMCI) and four clinical score prediction tasks. There are discriminative learning and relational-regularization in our proposed R2DLSR framework. It is interesting to know whether discriminative learning and relational regularization strategies can further improve the classification and regression performance. Hence, we provide a comprehensive investigation of the effect on the classification and regression results. In fact, the prior information, such as subject relational information and feature's relation information is quite beneficial to feature selection algorithm as well. Also, the manifold learning-based regularization is able to guide the feature selection steps to localize the informative and discriminative features to further boost AD/MCI classification and regression performance. The proposed techniques enhance both the regression and classification performance compared with the previous study. From the experimental results, it is clear that our method is statistically superior to the related algorithms in terms of both clinical scores (i.e., ADAS-Cog and MMSE) prediction and class label identification. R2DLSR method outperforms the state-of-the-arts method in most scenarios. The primary explanation is that the underlying structure of the original class (e.g., AD, pMCI, sMCI, and NC) is not quite complicated in most cases. Also, our experimental results are consistent with the previous studies.

Despite the promising performance achieved by the proposed method, we still have a few limitations of our method. First, we map the original binary learning problem into a multitask learning problem, and investigate whether the relational information is effective for our task. Second, we take only the feature–feature and subject–subject relational information into consideration. It could be interesting to integrate more complicated relationship learning (i.e., clinical scores' relational information) in a multitask learning framework rather than single and naive machine learning for feature selection.

Since we currently only focus on the ROI features, it is helpful to integrate the visual features using the state-of-the-arts computer vision techniques as well. Different feature extraction algorithms shall discover different characteristics of the AD/MCI subjects. In addition, we can uncover the sharing and common information among different features to facilitate the diagnosis and prognosis for the clinical doctors.

VI. CONCLUSION

In this paper, a discriminative sparse learning framework with multirelation regularization is designed for feature selection. We devise new regularization terms that consider relational information inherent in the observations for regression and classification. A novel objective function considering two new regularization terms and discriminative learning is also developed to jointly discover the internal relationship for AD/MCI classification. In our extensive experiments on the ADNI dataset, we demonstrate the effectiveness of the proposed method by comparing with the state-of-the-art methods for both clinical scores prediction and clinical label identification. The achieved promising results via utilization of the discriminative learning and devised relational regularization terms indicate that it will be beneficial for computer-aided AD diagnosis.

Appendix
Convergence Analysis

In order to prove the convergence of Algorithm 1, we first define the following lemma.

Lemma 1: For any invertible matrices G and its updated matrix $\widetilde{\boldsymbol{G}}$, the following inequality holds:

$$\mathrm{Tr}\left(\widetilde{\boldsymbol{G}}\right)^{\frac{1}{2}} - \frac{1}{2}\mathrm{Tr}\left(\widetilde{\boldsymbol{G}}\boldsymbol{G}^{\frac{1}{2}}\right) \leq \mathrm{Tr}(\boldsymbol{G})^{\frac{1}{2}} - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{G}\boldsymbol{G}^{\frac{1}{2}}\right). \tag{20}$$

Proof: Obviously, $(\|\boldsymbol{G}\|_2 - \|\widetilde{\boldsymbol{G}}\|_2)^2 \leq 0$

$$\left(\|\boldsymbol{G}\|_{\mathbf{2}} - \left\|\widetilde{\boldsymbol{G}}\right\|_2\right)^2 \leq 0 \Rightarrow 2\|\boldsymbol{G}\|_2\left\|\widetilde{\boldsymbol{G}}\right\|_2 - \left\|\widetilde{\boldsymbol{G}}\right\|_2^2 \leq \|G\|_2^2 \tag{21}$$

$$\Rightarrow \left\|\widetilde{\boldsymbol{G}}\right\|_{\mathbf{2}} - \frac{\left\|\widetilde{\boldsymbol{G}}\right\|_{\mathbf{2}}^{\mathbf{2}}}{\mathbf{2}\|\boldsymbol{G}\|_{\mathbf{2}}} \leq \|\boldsymbol{G}\|_{\mathbf{2}} - \frac{\|\boldsymbol{G}\|_{\mathbf{2}}^{\mathbf{2}}}{\mathbf{2}\|\boldsymbol{G}\|_{\mathbf{2}}} \tag{22}$$

$$\Rightarrow \mathrm{Tr}\left(\left\|\widetilde{\boldsymbol{G}}\right\|_2 - \frac{\left\|\tilde{G}\right\|_2^2}{2\|\boldsymbol{G}\|_2}\right) \leq \mathrm{Tr}\left(\|\boldsymbol{G}\|_2 - \frac{\|\boldsymbol{G}\|_2^2}{2\|\boldsymbol{G}\|_2}\right) \tag{23}$$

$$\Rightarrow \mathrm{Tr}\left(\widetilde{\boldsymbol{G}}^{\frac{1}{2}}\right) - \frac{1}{2}\mathrm{Tr}\left(\widetilde{\boldsymbol{G}}\boldsymbol{G}^{\frac{1}{2}}\right) \leq \mathrm{Tr}\left(\boldsymbol{G}^{\frac{1}{2}}\right) - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{G}\boldsymbol{G}^{\frac{\mathbf{1}}{\mathbf{2}}}\right). \tag{24}$$

Here, we need to prove that objective values in (8) and Algorithm 1 will decrease in each iteration and a gobal optimal solution will be obtained after convergence. Let $\widehat{\mathbf{Y}} = \mathbf{Y} + \mathbf{B} \odot \mathbf{P}$, and the loss function in rth iteration is denoted as: $l^{(r)} = \mathrm{Tr}((\widehat{\mathbf{Y}} - \mathbf{XW})^{\boldsymbol{T}}(\widehat{\mathbf{Y}} - \mathbf{XW}))$, $\widetilde{\mathbf{W}}$ is the updated $\mathbf{W}$ in each iteration.$l^{(r+1)} = \mathrm{Tr}((\widehat{\mathbf{Y}} - \mathbf{X}\widetilde{\mathbf{W}})^{\boldsymbol{T}}(\widehat{\mathbf{Y}} - \mathbf{X}\widetilde{\mathbf{W}}))$, it is obvious that, $l^{(r+1)} \leq l^{(r)}$. Algorithm 1 decreases the objective value in the objective function in each iteration and reaches the convergence. Based on Lemma 1, it is easy to prove that loss function is monotonically decreased, namely

$$\mathrm{Tr}\left((\widehat{\mathbf{Y}} - \mathbf{X}\widetilde{\mathbf{W}})^{\boldsymbol{T}}(\widehat{\mathbf{Y}} - \mathbf{X}\widetilde{\mathbf{W}})\right) \leq \mathrm{Tr}\left((\widehat{\mathbf{Y}} - \mathbf{X}\widetilde{\mathbf{W}})^{\boldsymbol{T}}(\widehat{\mathbf{Y}} - \mathbf{X}\widetilde{\mathbf{W}})\right). \tag{25}$$

Based on Algorithm 1, proof in Lemma 1, smoothness property and definition of $\boldsymbol{L}_f, \boldsymbol{L}_s$, we can have

$$\begin{aligned} & l^{(r+1)} + \lambda_1\mathrm{Tr}\left(\widetilde{\boldsymbol{W}}^{\boldsymbol{T}}\boldsymbol{L}_f\widetilde{\boldsymbol{W}}\right) + \lambda_2\mathrm{Tr}\left(\widetilde{\boldsymbol{W}}^{\boldsymbol{T}}\boldsymbol{L}_s\widetilde{\boldsymbol{W}}\right) \\ & \quad \leq l^{(r)} + \lambda_1\mathrm{Tr}\left(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{L}_f\boldsymbol{W}\right) + \lambda_2\mathrm{Tr}\left(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{L}_s\boldsymbol{W}\right). \end{aligned} \tag{26}$$

For the nonsmooth convex form $l_{2,1}$ norm, it is obvious that

$$l^{(r+1)} + \lambda\mathrm{Tr}\left(\widetilde{\mathbf{W}}^{\boldsymbol{T}}\boldsymbol{L}_D\widetilde{\mathbf{W}}\right) \leq l^{(r)} + \lambda\mathrm{Tr}\left(\mathbf{W}^{\boldsymbol{T}}\boldsymbol{L}_D\mathbf{W}\right) \tag{27}$$

$$\Rightarrow \sum_{i=1}^{F}\left(\|\tilde{w}_i\|_2 - \sum_{i=1}^{F}\frac{\|\tilde{w}_i\|_2^2}{2\|\tilde{w}_i\|_2}\right) \leq \sum_{i=1}^{F}\left(\|w_i\|_2 - \sum_{i=1}^{F}\frac{\|w_i\|_2^2}{2\|w_i\|_2}\right) \tag{28}$$

$$\Rightarrow \sum_{i=1}^{F}\left(\|\tilde{w}_i\|_2 - \tilde{w}_i^{\mathbf{T}}\boldsymbol{L}_D\tilde{w}_i\right) \leq \sum_{i=1}^{F}\left(\|w_i\|_2 - w_i{}^{\mathbf{T}}\boldsymbol{L}_D w_i\right) \tag{29}$$

$$\Rightarrow l^{(r+1)} + \lambda_1\sum_{i=1}^{F}\left(\frac{\|\tilde{w}_i\|_2^2}{2\|\tilde{w}_i\|_2}\right) \leq l^{(r)} + \lambda_1\sum_{i=1}^{F}\left(\frac{\|w_i\|_2^2}{2\|w_i\|_2}\right) \tag{30}$$

$$\begin{aligned} & \Rightarrow l^{(r+1)} + \lambda_1\sum_{i=1}^{F}\left(\frac{\|\tilde{w}\|_2^2}{2\|\tilde{w}_i\|_2} - \|\tilde{w}_i\|_2 + \|\tilde{w}_i\|_2\right) \\ & \quad \leq l^{(r)} + \lambda_1\sum_{i=1}^{F}\left(\frac{\|w_i\|_2^2}{2\|w_i\|_2} - \|w_i\|_2 + \|w_i\|_2\right) \end{aligned} \tag{31}$$

which completes the proof. ■

According to the above equations, it is evident that the algorithm monotonically decreases and therefore eventually converges. Since the objective function is convex, a globally optimal solution will be obtained.

When $\mathbf{W}$ is fixed, we can solve the subproblem $\min_{P \geq 0} l(\widetilde{\mathbf{W}}, \mathbf{P})$ to obtain the optimal $l^{(r+1)}$. Considering the convexity of this problem, we can derive that

$$l\left(\widetilde{\mathbf{W}}, \mathbf{P}\right) \leq l(\boldsymbol{W}, \mathbf{P}). \tag{32}$$

Consequently, Algorithm 1 monotonically decreases the objective function in each iteration to get $\boldsymbol{P}_{ij}$

$$\boldsymbol{P}_{ij} = \max\left(\boldsymbol{B}_{ij}\boldsymbol{T}_{ij}, \mathbf{0}\right). \tag{33}$$

As a result, we can obtain $\mathbf{P}$ by

$$\boldsymbol{P} = \max(\boldsymbol{B} \odot \boldsymbol{T}, \mathbf{0}). \tag{34}$$

References

[1] A. Alzheimer's, "2015 Alzheimer's disease facts and figures," *Alzheimer's Dementia*, vol. 11, no. 3, pp. 332–384, 2015.

[2] D. Zhang and D. Shen, "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease," *NeuroImage*, vol. 59, no. 2, pp. 895–907, 2012.

[3] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.

[4] X. Zhu, H.-I. Suk, and D. Shen, "A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis," *NeuroImage*, vol. 100, pp. 91–105, Oct. 2014.

[5] X. Zhu, H.-I. Suk, L. Wang, S.-W. Lee, and D. Shen, "A novel relational regularization feature selection method for joint regression and classification in AD diagnosis," *Med. Image Anal.*, Nov. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.media.2015.10.008

[6] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, Nov. 2014.

[7] B. Lei, S. Chen, D. Ni, and T. Wang, "Discriminative learning for Alzheimer's disease diagnosis via canonical correlation analysis and multimodal fusion," *Front Aging Neurosci.*, vol. 8, pp. 1–17, May 2016.

[8] L. Huang, Y. Jin, Y. Gao, K.-H. Thung, and D. Shen, "Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest," *Neurobiol. Aging*, vol. 46, pp. 180–191, Oct. 2016.

[9] M. Liu, D. Zhang, and D. Shen, "Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment," *IEEE Trans. Med. Imag.*, vol. 35, no. 6, pp. 1463–1474, Jun. 2016, doi: 10.1109/tmi.2016.2515021.

[10] M. Liu, D. Zhang, and D. Shen, "View-centralized multi-atlas classification for Alzheimer's disease diagnosis," *Human Brain Mapping*, vol. 36, no. 5, pp. 1847–1865, 2015.

[11] M. Liu, D. Zhang, and D. Shen, "Ensemble sparse classification of Alzheimer's disease," *NeuroImage*, vol. 60, no. 2, pp. 1106–1116, 2012.

[12] L. Nie *et al.*, "Modeling disease progression via multisource multitask learners: A case study with Alzheimer's disease," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. PP, no. 99, pp. 1–12, Feb. 2016.

[13] F. Liu, C.-Y. Wee, H. Chen, and D. Shen, "Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification," *NeuroImage*, vol. 84, pp. 466–475, Jan. 2014.

[14] T. Tong *et al.*, "Multiple instance learning for classification of dementia in brain MRI," *Med. Image Anal.*, vol. 18, no. 5, pp. 808–818, 2014.

[15] M. Liu, D. Zhang, and D. Shen, "Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis," *Human. Brain Mapping*, vol. 35, no. 4, pp. 1305–1319, 2014.

[16] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 389–422, 2002.

[17] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[18] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via multi-task learning," *NeuroImage*, vol. 78, pp. 233–248, Sep. 2013.

[19] C. Hinrichs, V. Singh, G. Xu, and S. C. Johnson, "Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population," *NeuroImage*, vol. 55, no. 2, pp. 574–589, 2011.

[20] C. Misra, Y. Fan, and C. Davatzikos, "Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI," *NeuroImage*, vol. 44, no. 4, pp. 1415–1422, 2009.

[21] M. Liu and D. Zhang, "Pairwise constraint-guided sparse learning for feature selection," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 298–310, Jan. 2016.

[22] R. Hong *et al.*, "Image annotation by multiple-instance learning with discriminative feature mapping and selection," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 669–680, May 2014.

[23] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[24] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. PP, no. 99, pp. 1–18, Apr. 2016.

[25] Y. Jin *et al.*, "Identification of infants at high−risk for autism spectrum disorder using multiparameter multiscale white matter connectivity networks," *Human Brain Mapping*, vol. 36, no. 12, pp. 4880–4896, 2015.

[26] X. Zhu, H.-I. Suk, and D. Shen, "A novel multi-relation regularization method for regression and classification in AD diagnosis," in *Proc. MICCAI*, Boston, MA, USA, 2014, pp. 401–408.

[27] B. Lei, S. Chen, D. Ni, and T. Wang, "Joint learning of multiple longitudinal prediction models by exploring internal relations," in *Proc. Mach. Learn. Med. Imag.*, Munich, Germany, 2015, pp. 330–337.

[28] L. Shi, L. Zhao, A. Wong, D. Wang, and V. Mok, "Mapping the relationship of contributing factors for preclinical Alzheimer's disease," *Sci. Rep.*, vol. 5, Jul. 2015, Art. no. 11259.

[29] B. Jie, D. Zhang, B. Cheng, and D. Shen, "Manifold regularized multitask feature learning for multimodality disease classification," *Human Brain Mapping*, vol. 36, no. 2, pp. 489–507, 2015.

[30] B. Lei *et al.*, "Discriminative learning for automatic staging of placental maturity via multi-layer Fisher vector," *Sci. Rep.*, vol. 5, Jul. 2015, Art. no. 12818.

[31] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1738–1754, Nov. 2012.

[32] T. Chen and K.-H. Yap, "Discriminative BoW framework for mobile landmark recognition," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 695–706, May 2014.

[33] L. Wang, X.-Y. Zhang, and C. Pan, "MSDLSR: Margin scalable discriminative least squares regression for multicategory classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2711–2717, Dec. 2016, doi: 10.1109/tnnls.2015.2477826.

[34] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.

[35] Q. Zhang, Y. Tian, Y. Yang, and C. Pan, "Automatic spatial–spectral feature selection for hyperspectral image via discriminative sparse multimodal learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 261–279, Jan. 2015.

[36] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, "Discriminative transfer subspace learning via low-rank and sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 850–863, Feb. 2016.

[37] D. Shen and C. Davatzikos, "HAMMER: Hierarchical attribute matching mechanism for elastic registration," *IEEE Trans. Med. Imag.*, vol. 21, no. 11, pp. 1421–1439, Nov. 2002.

[38] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, Jan. 2001.

[39] X. Zhu, X. Wu, W. Ding, and S. Zhang, "Feature selection by joint graph sparse coding," in *Proc. SDM*, Austin, TX, USA, 2013, pp. 803–811.

[40] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[41] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc.*, vol. 58, no. 1, pp. 267–288, 1996.

[42] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, and D. Rueckert, "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease," *NeuroImage*, vol. 65, pp. 167–175, Jan. 2013.

[43] R. Min, G. Wu, J. Cheng, Q. Wang, and D. Shen, "Multi-atlas based representations for Alzheimer's disease diagnosis," *Human Brain Mapping*, vol. 35, no. 10, pp. 5052–5070, 2014.

[44] D. Zhang, J. Liu, and D. Shen, "Temporally-constrained group sparse learning for longitudinal data analysis," in *Proc. MICCAI*, Nice, France, 2012, pp. 264–271.

[45] R. Cuingnet *et al.*, "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database," *NeuroImage*, vol. 56, no. 2, pp. 766–781, 2011.

[46] C. Davatzikos, P. Bhatt, L. M. Shaw, K. N. Batmanghelich, and J. Q. Trojanowski, "Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification," *Neurobiol. Aging*, vol. 32, no. 12, pp. 2322.e19–2322.e27, 2010.

Authors' photographs and biographies not available at the time of publication.