Matrix-Similarity Based Loss Function and Feature Selection for Alzheimer's Disease Diagnosis

Xiaofeng Zhu, Heung-Il Suk, and Dinggang Shen*
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, USA
{xiaofeng,hsuk,dinggang_shen}@med.unc.edu

Abstract

Recent studies on Alzheimer's Disease (AD) or its prodromal stage, Mild Cognitive Impairment (MCI), diagnosis presented that the tasks of identifying brain disease status and predicting clinical scores based on neuroimaging features were highly related to each other. However, these tasks were often conducted independently in the previous studies. Regarding the feature selection, to our best knowledge, most of the previous work considered a loss function defined as an element-wise difference between the target values and the predicted ones. In this paper, we consider the problems of joint regression and classification for AD/MCI diagnosis and propose a novel matrix-similarity based loss function that uses high-level information inherent in the target response matrix and imposes the information to be preserved in the predicted response matrix. The newly devised loss function is combined with a group lasso method for joint feature selection across tasks, i.e., clinical scores prediction and disease status identification. We conducted experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, and showed that the newly devised loss function was effective to enhance the performances of both clinical score prediction and disease status identification, outperforming the state-of-the-art methods.

1. Introduction

Alzheimer's Disease (AD) is the most common form of dementia that often appears in the persons aged over 65. Brookmeyer *et al*. showed that there are 26.6 million AD patients worldwide and 1 out of 85 people will be affected by AD by 2050 [1]. Thus, for timely treatment that might be effective to slow the progression, it's of great importance for early diagnosis of AD and its prodromal stage Mild Cognitive Impairment (MCI).

For the last decade, machine learning techniques have been successfully used to analyze complex patterns in neuroimaging data, especially, for AD/MCI diagnosis [5, 20, 21, 25]. The previous imaging studies mostly focused on developing classification models (*e.g*. [8, 11, 17]) to identify clinical labels such as AD, MCI, and Normal Control (NC). Recently, regression models have also been investigated to predict clinical scores such as Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) and Mini-Mental State Examination (MMSE) from individual Magnetic Resonance Imaging (MRI) and/or Positron Emission Tomography (PET) scans [15].

Unlike those studies that focused on only one of the tasks [8, 17], there have been also efforts to tackle both tasks simultaneously in a unified framework. For example, Zhang and Shen [24] proposed a method of joint feature selection for both disease diagnosis and clinical scores prediction, and showed that the features used for these tasks are highly related. In line with their work and for better understanding of the underlying mechanism of AD, our interest in this paper is to predict clinical scores and to identify disease status jointly and here we call it as a Joint Regression and Classification (JRC) problem.

In the computer-aided AD diagnosis, the available sample size is usually much smaller than the feature dimensionality. For example, the sample size used in [8, 11] was as small as 103 (*i.e*., 51 AD and 52 NC), while the feature dimensionality (including MRI features 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 the over-fitting issue although the number of intrinsic features may be very low [22, 26, 28]. To circumvent this challenging problem, Wang *et al*. used features of the predefined Region-Of-Interests (ROIs) of medial temporal lobe structures, medial and lateral parietal, and prefrontal cortical areas in predicting memory scores and discriminating between AD and NC [19].

Rather than predefining the ROIs based on the prior knowledge, it is preferable to select the informative features in a data-driven manner. In this respect, Zhang and Shen embedded an $\ell_{2,1}$-norm regularizer into the sparse re-

*corresponding author.

gression model, thus formulating a multi-task learning [24]. Recent studies demonstrated that the consideration of the manifold of the data can further boost the power of feature selection methods [27, 29]. However, to our best knowledge, the previous methods used mostly a loss function defined as sum of the element-wise difference between target values and predicted ones, and considered only the manifold of feature observations, not that of the target variables. Furthermore, none of the previous methods utilized manifold-based feature selection for the JRC problem.

In this paper, we propose a new loss function that uses high-level information inherent in the observations, and combine it with a group lasso [23] for joint sparse feature selection in the JRC problem. Specifically, we define a loss function as *matrix similarity* and impose the high-level information in the target response matrix to be preserved in the predicted response matrix. For the high-level information, we use the relations between samples and the relations between response variables, each of which we call as '*sample-sample relation*' and '*variable-variable relation*'. Hereafter, each column and each row of a matrix correspond, respectively, to one sample and one response variable. In our work, a sample in a response matrix consists of clinical scores and a class label, and each of the clinical scores or a class label is considered as a response variable. By utilizing the high-level information inherent in the target response matrix and imposing it to be preserved in the predicted response matrix, we define a more sophisticated loss function, which affects feature selection, and thus helps enhance the prediction and classification performances in AD/MCI diagnosis.

2. Proposed Method

In Fig. 1, we present a schematic diagram of the proposed method to predict both clinical scores and a class label using neuroimaging data. Given MRI, PET, and CerebroSpinal Fluid (CSF) data, we first extract features from MRI and PET, while using the CSF biomarkers as CSF features. We then construct a feature matrix $\mathbf{X}$ with a concatenation of multi-modal features at each column, and the corresponding response matrix $\mathbf{Y}$ with a concatenation of clinical scores (*e.g.*, ADAS-Cog, MMSE) and a class label at each column. With our new loss function and a group lasso method, we select features jointly used to represent clinical scores and a class label. By using the training samples but with only the selected features, we build clinical scores regression models and a clinical label identification model with Support Vector Regression (SVR) and Support Vector Classification (SVC), respectively.

2.1. 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 column are denoted as $\mathbf{x}^i$ and $\mathbf{x}_j$, respectively. We denote a Frobenius norm and an $\ell_{2,1}$-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 $\|\mathbf{X}\|_{2,1} = \sum_i \|\mathbf{x}^i\|_2$, respectively. We further denote a transpose, a trace, and an inverse of a matrix $\mathbf{X}$ as $\mathbf{X}^T$, $tr(\mathbf{X})$, and $\mathbf{X}^{-1}$, respectively.

2.2. Matrix-Similarity based Loss Function

Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ and $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n] \in \mathbb{R}^{c\times n}$, where *n*, *d*, and *c* denote the numbers of samples (or subjects)[1], feature variables, and response variables, respectively. In our work, the response variables correspond to ADAS-Cog, MMSE, and a class label. We assume that the response variables can be represented by a weighted linear combination of the features as follows:

$$\mathbf{Y} \approx \mathbf{W}^T\mathbf{X} = \hat{\mathbf{Y}}$$

where $\mathbf{W} \in \mathbb{R}^{d\times c}$ is a regression coefficient matrix. By regarding the prediction of each response variable as a task and constraining the same features to be used across tasks, we can formulate a multi-task learning with a group lasso [23] as follows:

$$\min_{\mathbf{W}} f(\mathbf{W}) + \lambda\|\mathbf{W}\|_{2,1} \tag{1}$$

where $f(\mathbf{W})$ is a loss function depending on $\mathbf{W}$ and λ is a sparsity control parameter. Note that each element in a column $\mathbf{w}_k$ of $\mathbf{W}$ assigns a weight to each of the observed features in representing the k-th response variable. The $\ell_{2,1}$-norm regularizer $\|\mathbf{W}\|_{2,1}$ penalizes all coefficients in the same row of $\mathbf{W}$ together for joint feature selection or unselection in predicting the response variables. Specifically, the ℓ_2-norm regularizer enforces the selection of the same features across tasks, and the ℓ_1-norm imposes the feature sparseness in the linear combination. Note that later in our experiments for clinical score prediction and clinical label classification, we extract one feature from each Region-Of-Interest (ROI) of the brain in MRI or PET, thus this $\ell_{2,1}$-norm has the effect of selecting ROIs that are highly relevant to the prediction of both clinical scores and a class label.

With regard to the loss function in Eq. (1), the most commonly used metric in the literature is the element-wise distance between the target response matrix $\mathbf{Y}$ and the predicted response matrix $\hat{\mathbf{Y}}$ as follows:

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

[1] In this work, we have one sample per subject.

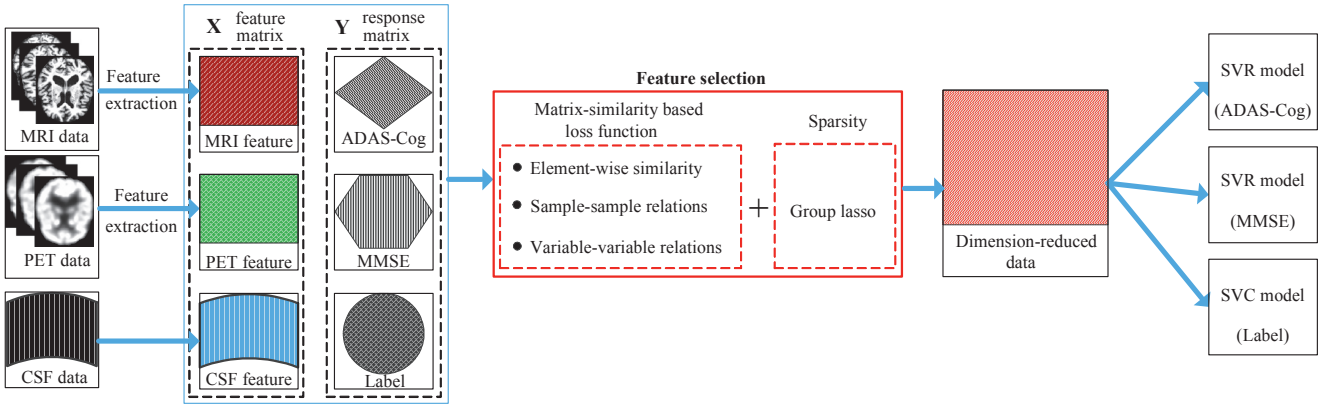


Figure 1. The framework of the proposed method.

This element-wise loss function has been successfully used in many objective functions in the literature [18, 23, 24]. From a matrix similarity point of view, Eq. (2) measures the similarity with the sum of the element-wise differences between matrices. Note that, in this case, the lower the score is, the more similar they are. However, we believe that there exists additional information inherent in the matrices, which we can use in measuring the similarity, such as the relations between any pair of columns and the relations between any pair of rows. In our case, the columns and the rows correspond, respectively, to samples and response variables. Ideally, besides the element-wise values, those relations in the target response matrix $\mathbf{Y}$ should be also preserved in the predicted response matrix $\hat{\mathbf{Y}}$. By imposing the higher-level information to be matched between two matrices, we can find an optimal regression matrix $\mathbf{W}$ that helps accurately predict the target response values, and thus select useful features. The selected features can be finally used in predicting clinical scores and a class label of a testing sample.

To better characterize the newly devised loss function, we explain them in terms of a graph matching. We illustrate the sample-sample (a pair of columns) relations, *e.g.*, $(\mathbf{y}_i - \mathbf{y}_j)$ or $(\hat{\mathbf{y}}_i - \hat{\mathbf{y}}_j)$, and the variable-variable (a pair of rows) relations, *e.g.*, $(\mathbf{y}^k - \mathbf{y}^l)$ or $(\hat{\mathbf{y}}^k - \hat{\mathbf{y}}^l)$, by means of a graph in Fig. 2(a) and Fig. 2(b), respectively. In Fig. 2(a), a node represents one sample, *i.e.*, a column vector $\mathbf{y}_i$ or $\hat{\mathbf{y}}_i$ in the respective matrices, an edge in a graph denotes the relation between the connected nodes, and different colors denote different class labels. In the graph, the samples of the same class would have a small distance, whereas the samples of different classes would have a large distance. In Fig. 2(b), a node represents a set of observations for a response variable, *i.e.*, a row vector in the respective matrices, and an edge denotes the relation between nodes.

As mentioned above, we impose these relational properties in a target response matrix, now represented by a graph, to be preserved in the respective graph for the predicted response matrix as follows:

$$\begin{aligned} G^S_{\mathbf{Y}} &\approx G^S_{\hat{\mathbf{Y}}} \\ G^V_{\mathbf{Y}} &\approx G^V_{\hat{\mathbf{Y}}} \end{aligned}$$

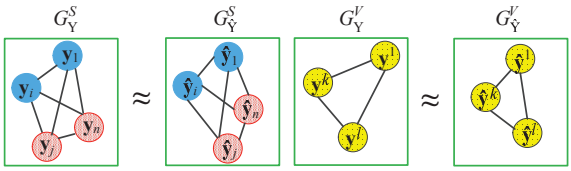
where $G^S_{\mathbf{Y}}$ and $G^S_{\hat{\mathbf{Y}}}$ denote, respectively, graphs representing the sample-sample relations for the target response matrix $\mathbf{Y}$ and the predicted response matrix $\hat{\mathbf{Y}}$, and $G^V_{\mathbf{Y}}$ and $G^V_{\hat{\mathbf{Y}}}$ denote, respectively, graphs representing the variable-variable relations for the target response matrix $\mathbf{Y}$ and the predicted response matrix $\hat{\mathbf{Y}}$. Hereafter, we call the graphs representing the sample-sample relations and the variable-variable relations as 'S-graph' and 'V-graph', respectively. We formulate the problem of matching two graphs, *i.e.*, S-graph and V-graph, as follows:

$$\begin{aligned} M_S &= \sum_{i,j=1}^{n} \|(\mathbf{y}_i - \mathbf{y}_j) - (\hat{\mathbf{y}}_i - \hat{\mathbf{y}}_j)\|_2^2 \qquad (3) \\ &= \sum_{i,j=1}^{n} \|(\mathbf{y}_i - \mathbf{y}_j) - (\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j)\|_2^2 \\ M_V &= \sum_{k,l=1}^{c} \|(\mathbf{y}^k - \mathbf{y}^l) - (\hat{\mathbf{y}}^k - \hat{\mathbf{y}}^l)\|_2^2 \qquad (4) \\ &= \sum_{k,l=1}^{c} \|(\mathbf{y}^k - \mathbf{y}^l) - ((\mathbf{w}_k)^T\mathbf{X} - (\mathbf{w}_l)^T\mathbf{X})\|_2^2 \end{aligned}$$

where M_S and M_V denote, respectively, the graph matching scores between $G^S_{\mathbf{Y}}$ and $G^S_{\hat{\mathbf{Y}}}$, and between $G^V_{\mathbf{Y}}$ and $G^V_{\hat{\mathbf{Y}}}$. By introducing these newly devised graph matching terms into the loss function of Eq. (2), our new loss function becomes as follows:

$$f(\mathbf{W}) = \|\mathbf{Y} - \mathbf{W}^T\mathbf{X}\|_F^2 + \alpha_1 M_S + \alpha_2 M_V \qquad (5)$$

where α_1 and α_2 denote, respectively, the control parameters for the terms. Compared to the conventional element-wise loss function in Eq. (2), the proposed function additionally considers two graph matching regularization terms.



(a) S-graph matching (b) V-graph matching

Figure 2. An illustration of measuring a matrix similarity by means of a graph matching. For simplicity, we showed only a small number of nodes. (a) Each node represents a column vector of the target or the predicted response matrix, edges represent the distance between nodes, and colors represent class labels. (b) Each node represents a row vector of the target or the predicted response matrix and edges denote the distance between nodes.

It is worth noting that unlike the previous manifold learning methods, *i.e.*, local linear embedding [13], locality preserving projection [7], and high-order graph matching [10], that focused on the sample similarities by imposing nearby samples to be still nearby in the transformed space, the proposed method imposes more strict constraints, *i.e.*, sample-sample relations and variable-variable relations, in finding the optimal regression matrix $\mathbf{W}$.

2.3. Objective Function Optimization

With some mathematical transformations, we can simplify M_S and M_V as follows:

$$M_S = tr(2\mathbf{W}^T\mathbf{X}\mathbf{H}_n\mathbf{X}^T\mathbf{W} - 4\mathbf{Y}\mathbf{H}_n\mathbf{X}^T\mathbf{W}) \quad (6)$$

$$M_V = tr(2\mathbf{X}^T\mathbf{W}\mathbf{H}_c\mathbf{W}^T\mathbf{X} - 4\mathbf{X}^T\mathbf{W}\mathbf{H}_c\mathbf{Y}) \quad (7)$$

where $\mathbf{H}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n(\mathbf{1}_n)^T$ and $\mathbf{H}_c = \mathbf{I}_c - \frac{1}{c}\mathbf{1}_c(\mathbf{1}_c)^T$, $\mathbf{I}_n$ (or $\mathbf{I}_c$) is an identity matrix of size n (or c), and $\mathbf{1}_n$ (or $\mathbf{1}_c$) is a column vector of n (or c) ones. By replacing the graph matching terms M_S and M_V in Eq. (5) with Eq. (6) and Eq. (7), respectively, our objective function can be rewritten as follows:

$$\begin{aligned}\min_{\mathbf{W}} \quad & \|\mathbf{Y} - \mathbf{W}^T\mathbf{X}\|_F^2 \\ & +\alpha_1 tr(2\mathbf{W}^T\mathbf{X}\mathbf{H}_n\mathbf{X}^T\mathbf{W} - 4\mathbf{Y}\mathbf{H}_n\mathbf{X}^T\mathbf{W}) \\ & +\alpha_2 tr(2\mathbf{X}^T\mathbf{W}\mathbf{H}_c\mathbf{W}^T\mathbf{X} - 4\mathbf{X}^T\mathbf{W}\mathbf{H}_c\mathbf{Y}) \\ & +\lambda\|\mathbf{W}\|_{2,1}. \end{aligned} \quad (8)$$

By setting the derivative of the objective function in Eq. (8) with respect to $\mathbf{W}$ to zero, we can obtain an equation of the following form:

$$\mathbf{A}\mathbf{W} + \mathbf{W}\mathbf{B} = \mathbf{C} \quad (9)$$

where $\mathbf{A} = -(\mathbf{X}\mathbf{X}^T)^{-1}(\mathbf{X}\mathbf{X}^T + 2\alpha_1\mathbf{X}\mathbf{H}_n\mathbf{X}^T + \lambda\mathbf{Q})$, $\mathbf{B} = 2\alpha_2\mathbf{H}_c$, $\mathbf{C} = -(\mathbf{X}\mathbf{X}^T)^{-1}(\mathbf{X}\mathbf{Y}^T + 2\alpha_1\mathbf{X}\mathbf{H}_n\mathbf{Y}^T + 2\alpha_2\mathbf{X}\mathbf{Y}^T\mathbf{H}_c)$, and $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with the i-th diagonal element set to

$$q_{ii} = \frac{1}{2\|\mathbf{w}^i\|_2}. \quad (10)$$

Although the objective function in Eq. (8) is convex, due to the non-smooth term of $\|\mathbf{W}\|_{2,1}$, it is not straightforward to find the global optimum. Furthermore, due to the interdependence in computing matrices of $\mathbf{W}$ and $\mathbf{Q}$, it's not trivial to solve Eq. (9). To this end, in this work, we apply an iterative approach to optimize Eq. (9) 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 $\mathbf{Q}(t-1)$ fixed and then update the matrix $\mathbf{Q}(t)$ with the updated matrix $\mathbf{W}(t)$. Due to the limited space, we don't prove the convergence of Algorithm 1, but in a nutshell, according to [29], the error value monotonically decreases in every iteration.

Algorithm 1: Pseudo code of solving Eq. (8).

Input: $\mathbf{X} \in \mathbb{R}^{d\times n}$, $\mathbf{Y} \in \mathbb{R}^{c\times n}$, $\alpha_1, \alpha_2, \lambda$;
Output: $\mathbf{W}$;
Initialize $t = 0$, $\mathbf{Q}(t)$ as a random diagonal matrix;
repeat
  Update $\mathbf{W}(t+1)$ by solving Eq. (9)[2];
  Update $\mathbf{Q}(t+1)$ via Eq. (10);
  $t = t+1$;
until Eq. (8) converges;

2.4. Feature Selection and Model Training

Due to the use of an $\ell_{2,1}$-norm regularizer in our objective function, after finding the optimal solution with Algorithm 1, we have some zero (or close to zero) row vectors in $\mathbf{W}$, whose corresponding features are not useful in joint prediction of clinical scores and a class label. Furthermore, we believe that 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 each row's ℓ_2-norm value, *i.e.*, $\|\mathbf{w}^j\|_2, j \in \{1, \ldots, d\}$, to find K top-ranked rows, and then select the respective features. Note that the selected features are jointly used to predict clinical scores and a class label.

By using training samples but with only the selected features, we then train support vector machines, which have been successfully used in many fields [16, 24]. Specifically, we build two SVR models for ADAS-Cog and MMSE scores prediction, respectively, and a SVC model for a class label identification[3].

[2] In our work, we used the built-in function 'lyap' in MATLAB.
[3] We used the LIBSVM toolbox [2].

3. Experimental Results

We conducted various experiments on the ADNI dataset[4] to compare the proposed method with the state-of-the-art methods as detailed below.

3.1. Experimental Settings

In our experiments, we used baseline MRI, PET, and CSF data obtained from 202 subjects including 51 AD subjects, 52 NC subjects, 43 MCI Converters (MCI-C), and 56 MCI Non-Converters (MCI-NC). We preprocessed the MRI and PET images by performing spatial distortion, skull-stripping, and cerebellum removal, sequentially. We segmented MRI images into gray matter, white matter, and cerebrospinal fluid. We then parcellated MRI images into 93 ROIs based on a Jacob template [9], by means of registering via HAMMER [14]. We finally computed the gray matter tissue volumes of the ROIs as features. For the PET images, we aligned them to their respective MRI images. We obtained 93 gray matter volumes from an MRI image and 93 mean intensities from a PET image and used them as features.

We considered two binary classification problems: AD vs. NC and MCI vs. NC. We used features from MRI, PET, MRI+PET (MP for short), or MRI+PET+CSF (MPC for short) and learned feature selection models with the target responses composed of two clinical scores and one class label. We then trained regression models and a classification model using the training samples with only the selected features.

For the quantitative performance evaluation, we employed the metrics of Correlation Coefficient (CC) and Root Mean Squared Error (RMSE) between the predicted clinical scores and the target clinical scores in regression, and the metrics of classification ACCuracy (ACC), SENsitivity (SEN), SPEcificity (SPE), and the Area Under an receiver operating characteristic Curve (AUC) in classification.

We performed 10-fold cross-validation and repeated the whole process 10 times to avoid the possible bias during dataset partitioning for cross-validation. We reported the averaged performances.

3.2. Competing Methods

In order to show the validity of the proposed method, we compare our method with the following methods.

- Full features based method: We conducted the tasks of regression and classification using the original full features with no feature selection step, and considered the results as baseline. In the following, we denote this method with the suffix "N".
- Single-task based method: We conducted regression and classification tasks separately but selecting features using the objective function in Eq. (8). Although here we used the same original features as the proposed method, we performed the task of regression or classification by selecting the set of features separately. In the following, we use the suffix "S" to represent the type of single-task based method.
- M3T [24]: A Multi-Modal Multi-Task method includes two key steps: (1) multi-task learning for each modality to find relevant features that jointly represent multiple response variables, and (2) a multi-kernel learning to integrate decisions from multiple modalities. It is noteworthy that M3T is a special case of our method by setting $\alpha_1 = \alpha_2 = 0$ in Eq. (8).
- HOGM [10]: A High-Order Graph Matching method uses a sample-sample relation in a matrix and applies an ℓ_1-norm regularization term with a single response variable (*i.e.*, single-task learning).
- M2TFS [8]: A Manifold regularized Multi-Task Feature Selection method selects features by combining the least square loss function with an $\ell_{2,1}$-norm regularizer and a graph regularizer. It then performs multi-modality classification via a multi-task learning, in which each task focuses on each modality. This method is designed only for classification. In our experiments, we considered two versions of M2TFS depending on the way of fusing multi-modality information: M2TFS-C (simple concatenation of multi-modality features into a long vector) and M2TFS-K (combining decisions from modalities through a multi-kernel learning).
- SJCR [19]: A Sparse Joint Classification and Regression method jointly uses a logistic loss function and a least square loss function along with an $\ell_{2,1}$-norm for multi-task feature selection.

3.3. Classification Results

Table 1 shows the classification performances of the methods. It is clear that the proposed method outperforms the competing methods in all experiments. Specifically, we observe the following results.

- It is important to conduct feature selection on the high-dimensional features before performing classification. The worst results were obtained by the methods without feature selection, *i.e.*, MRI-N, PET-N, MP-N, and MPC-N. For example, for MRI-based classification as shown in the first block of Table 1, even using a simple feature selection method, *i.e.*, MRI-S, can still increase the classification accuracies by 1.7% and 8.4%,

[4]Please refer to "www.adni-info.org" for details.

compared to MRI-N in AD vs. NC and MCI vs. NC classifications, respectively. Compared to the baseline, our method with MPC improved the classification accuracies by 5.1% and 9.5% in AD vs. NC and MCI vs. NC classifications, respectively.

- It is beneficial to use joint regression and classification framework for feature selection, even only for the task of classification. As shown in Table 1, the proposed method that performed feature selection for joint regression and classification achieved better classification performance than the single-task based methods (MRI-S, PET-S, MP-S, and MPC-S). For example, in MRI-based classification, our method improved the classification accuracies by 2.6% and 3.0% compared to MRI-S based method in AD vs. NC and MCI vs. NC classifications, respectively.

- To fuse multi-modal features helps improve classification performance. In all experiments, the classification performances with multi-modality data such as MP and MPC were better than the same methods with single-modality data such as MRI and PET. Also, the classification performance by MPC was generally better than MP. For example, in the discrimination between AD and NC, the proposed method with MPC achieved the classification accuracy of 95.9%, sensitivity of 95.7%, specificity of 98.6%, and AUC of 98.8%, while the best performance among other competing methods with single-modality data was 93.8% (ACC), 92.3% (SEN), 96.7% (SPE), and 97.9% (AUC), respectively, and the best performance among other competing methods with MP data was 95.3% (ACC), 94.9% (SEN), 98.1% (SPE), and 98.3% (AUC), respectively.

3.4. Regression Results

We also evaluated the regression performances using MRI, PET, MP, and MPC. We presented the results of CCs and RMSEs of all the competing methods in Table 2. From the results, it is clear that the proposed method outperforms all the competing methods, when using any combinations of three types of data.

Specifically, we observe the following: (1) Again, the regression performance of the methods without feature selection (MRI-N, PET-N, MP-N and MPC-N) was much worse than the methods with feature selection. Moreover, our method achieved the best performance compared to the competing methods. (2) Our method with MPC consistently outperformed the same method with MP on each performance measure, although the method with MP already achieved a better performance than our method with a single modality such as MRI and PET. This was also observed for all other competing methods. (3) The multi-task learning for joint feature selection to represent multiple response variables achieved better performances than the single-task learning, same as for the classification task above.

3.5. Summary

From all the experimental results, we found that (1) the proposed matrix-similarity based loss function helped enhance performances of both regression and classification; (2) the proposed method formulated in a joint regression and classification framework was superior to its counterpart that was formulated for regression and classification separately; (3) the multi-modal information fusion helped improve the performances compared to using the unimodal information.

To verify a statistical significance, we performed paired-sample t-tests between results of our method and those of the competing methods. For most of the cases, the p-values were less than 0.001, which means that our method statistically outperformed the competing methods on the tasks of predicting clinical scores (*i.e.*, ADAS-Cog and MMSE) and identifying a class label.

3.6. Most Discriminative Brain Regions

We also investigated the most discriminative regions based on the features selected by the proposed method. Due to the application of a cross-validation technique, the selected features varied across the repeated experiments. We thus found the most discriminative regions based on the selected frequency of each region over the cross-validations. The top 10 selected regions in MCI vs. NC classification with MPC included the following brain areas: amygdala right, hippocampal formation left, hippocampal formation right, entorhinal cortex left, temporal pole left, parahippocampal gyrus left, uncus left, perirhinal cortex left, cunecus left, and temporal pole right. It is noteworthy that the top six-ranked brain regions are known to be highly related to AD and MCI in many previous studies [3, 4, 6, 11, 12, 24].

4. Conclusions

In this paper, we proposed a novel matrix-similarity based loss function. Specifically, we used high-level information inherent in the target response matrix and imposed the information to be preserved in the predicted response matrix. Our objective function for joint feature selection was formulated by combining the newly devised loss function with a group lasso. In our extensive experiments on ADNI dataset, we validated the effectiveness of the proposed method by showing the performance enhancements in both the clinical scores (*i.e.*, ADAS-Cog and MMSE) prediction and the class label identification, outperforming the state-of-the-art methods.

Table 1. Comparison of classification performances (%) of the competing methods. (ACC: ACCuracy, SEN: SENsitivity, SPE: SPEcificity, and AUC: Area Under the receiver operating characteristic Curve)

Modality	Method	AD vs. NC				MCI vs. NC			
		ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
MRI	MRI-N	89.5	82.7	86.3	95.3	68.3	92.6	39.2	82.5
	MRI-S	91.2	85.9	92.5	96.7	76.7	93.3	37.6	83.7
	HOGM	93.4	89.5	92.5	97.1	77.7	**95.6**	51.4	84.4
	M3T	92.6	87.2	95.9	97.5	78.1	94.5	54.0	83.1
	SJCR	92.5	89.6	96.1	97.5	77.9	95.2	52.7	84.9
	Proposed	**93.8**	**89.7**	**96.7**	**97.9**	**79.7**	95.0	**56.1**	**85.2**
PET	PET-N	86.2	83.5	84.8	94.8	69.0	95.0	30.8	77.9
	PET-S	87.9	85.7	90.9	94.7	73.8	**96.5**	36.2	78.7
	HOGM	91.7	91.1	92.8	95.6	74.7	**96.5**	43.2	79.3
	M3T	90.9	90.5	93.1	96.4	77.2	94.5	44.3	80.5
	SJCR	91.6	83.7	93.6	95.9	73.8	96.4	23.2	80.8
	Proposed	**92.3**	**92.3**	**93.9**	**96.6**	**79.1**	96.1	**47.2**	**81.2**
MP	MP-N	89.7	92.2	85.9	96.1	71.6	96.1	43.9	82.7
	MP-S	90.8	92.6	93.8	96.7	76.3	**97.0**	39.9	83.4
	M2TFS-C	91.0	90.4	91.4	95.0	73.4	76.5	**67.1**	78.0
	M2TFS-K	95.0	**94.9**	95.0	97.0	79.3	85.9	66.6	82.0
	HOGM	95.2	92.8	95.4	97.8	79.5	96.6	58.6	84.6
	M3T	94.0	92.0	96.3	98.0	78.4	95.0	57.7	83.9
	SJCR	93.4	92.7	96.5	96.8	78.2	96.1	54.4	83.1
	Proposed	**95.3**	93.5	**98.1**	**98.3**	**80.2**	96.5	59.7	**85.5**
MPC	MPC-N	90.8	93.1	88.3	96.5	72.5	96.3	47.1	84.1
	MPC-S	92.5	94.1	93.8	97.6	77.1	97.1	47.5	83.9
	HOGM	95.6	94.5	96.9	98.5	80.6	96.7	**64.7**	86.2
	M3T	94.6	93.1	96.4	98.5	80.1	95.2	58.7	84.3
	SJCR	93.9	92.8	96.5	97.0	78.9	96.2	55.9	83.4
	Proposed	**95.9**	**95.7**	**98.6**	**98.8**	**82.0**	**98.0**	60.1	**87.0**

Table 2. Comparison of regression performances of the competing methods. (CC: Correlation Coefficient, RMSE: Root Mean Square Error)

Modality	Method	AD vs. NC				MCI vs. NC			
		ADAS-Cog		MMSE		ADAS-Cog		MMSE	
		CC	RMSE	CC	RMSE	CC	RMSE	CC	RMSE
MRI	MRI-N	0.587	4.96	0.520	2.02	0.329	4.48	0.309	1.90
	MRI-S	0.591	4.85	0.566	1.95	0.347	4.27	0.367	1.64
	HOGM	0.625	**4.53**	0.598	1.91	0.352	4.26	0.371	1.63
	M3T	0.649	4.60	0.638	1.91	0.445	4.27	0.420	1.66
	SJCR	0.652	4.69	0.636	1.90	0.448	4.27	0.425	1.67
	Proposed	**0.661**	4.58	**0.650**	**1.89**	**0.461**	**4.21**	**0.441**	**1.62**
PET	PET-N	0.597	4.86	0.514	2.04	0.333	4.34	0.331	1.70
	PET-S	0.620	4.83	0.593	2.00	0.356	4.26	0.359	1.69
	HOGM	0.600	4.69	0.515	1.99	0.360	**4.21**	0.368	1.67
	M3T	0.647	4.67	0.593	1.92	0.447	4.24	0.432	1.68
	SJCR	0.644	4.65	0.595	1.97	0.446	4.23	0.426	1.68
	Proposed	**0.663**	**4.64**	**0.610**	**1.89**	**0.452**	**4.21**	**0.444**	**1.66**
MP	MP-N	0.626	4.80	0.587	1.99	0.365	4.29	0.335	1.69
	MP-S	0.634	4.83	0.585	1.92	0.359	4.25	0.371	1.67
	M2TFS-C	0.641	4.89	0.636	1.87	0.446	4.25	0.408	1.64
	M2TFS-K	0.645	4.59	0.648	1.82	0.458	4.21	0.415	1.63
	HOGM	0.633	4.64	0.602	1.83	0.364	**4.20**	0.365	1.65
	M3T	0.653	4.61	0.639	1.91	0.450	4.23	0.433	1.64
	SJCR	0.656	4.64	0.643	1.81	0.451	4.19	0.431	1.64
	Proposed	**0.666**	**4.53**	**0.651**	**1.80**	**0.463**	**4.20**	**0.448**	**1.62**
MPC	MPC-N	0.629	4.79	0.588	1.97	0.368	4.29	0.337	1.70
	MPC-S	0.638	4.81	0.599	1.92	0.366	4.25	0.394	1.66
	HOGM	0.639	4.63	0.611	1.81	0.365	4.20	0.368	1.65
	M3T	0.665	4.59	0.663	1.81	0.451	4.22	0.441	1.62
	SJCR	0.658	4.64	0.645	1.81	0.451	4.19	0.433	1.63
	Proposed	**0.668**	**4.47**	**0.685**	**1.78**	**0.470**	**4.16**	**0.456**	**1.59**

Acknowledgements

This study was supported by National Institutes of Health (EB006733, EB008374, EB009634, AG041721, AG042599, and MH100217). Xiaofeng Zhu was partly supported by the Natural Science Foundation of China (NSFC) under grant 61263035.

References

[1] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and M. H. Arrighi. Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia : the Journal of the Alzheimer's Association*, 3(3):186–191, 2007.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

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

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

[5] Y. Fan, H. Rao, H. Hurt, J. Giannetta, M. Korczykowski, D. Shera, B. B. Avants, J. C. Gee, J. Wang, and D. Shen. Multivariate examination of brain abnormality using both structural and functional MRI. *NeuroImage*, 36(4):1189–1199, 2007.

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

[7] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *NIPS*, pages 1–8, 2005.

[8] B. Jie, D. Zhang, B. Cheng, and D. Shen. Manifold regularized multi-task feature selection for multi-modality classification in Alzheimer's disease. In *MICCAI*, pages 9–16, 2013.

[9] N. J. Kabani. 3D anatomical atlas of the human brain. *NeuroImage*, 7:0700–0717, 1998.

[10] F. Liu, H.-I. Suk, C.-Y. Wee, H. Chen, and D. Shenn. High-order graph matching based feature selection for Alzheimer's disease identification. In *MICCAI*, pages 311–318, 2013.

[11] 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*, 84:466–475, 2014.

[12] 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*, 44(4):1415–1422, 2009.

[13] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

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

[15] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack Jr, J. Ashburner, and R. S. Frackowiak. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *NeuroImage*, 51(4):1405–1413, 2010.

[16] H.-I. Suk and S.-W. Lee. A novel Bayesian framework for discriminative feature extraction in brain-computer interfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):286–299, 2013.

[17] H.-I. Suk and D. Shen. Deep learning-based feature representation for AD/MCI classification. In *MICCAI*, pages 583–590, 2013.

[18] H.-I. Suk, C.-Y. Wee, and D. Shen. Discriminative group sparse representation for mild cognitive impairment classification. In *MLMI*, pages 131–138, 2013.

[19] H. Wang, F. Nie, H. Huang, S. Risacher, A. J. Saykin, and L. Shen. Identifying AD-sensitive and cognition-relevant imaging biomarkers via joint classification and regression. In *MICCAI*, pages 115–123, 2011.

[20] C.-Y. Wee, P.-T. Yap, W. Li, K. Denny, J. N. Browndyke, G. G. Potter, K. A. Welsh-Bohmer, L. Wang, and D. Shen. Enriched white matter connectivity networks for accurate identification of MCI patients. *NeuroImage*, 54(3):1812–1822, 2011.

[21] C.-Y. Wee, P.-T. Yap, D. Zhang, K. Denny, J. N. Browndyke, G. G. Potter, K. A. Welsh-Bohmer, L. Wang, and D. Shen. Identification of MCI individuals using structural and functional connectivity networks. *Neuroimage*, 59(3):2045–2056, 2012.

[22] K. Q. Weinberger, F. Sha, and L. K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *ICML*, pages 17–24, 2004.

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

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

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

[26] X. Zhu, W. Ding, X. Wu, and S. Zhang. Feature selection by joint graph sparse coding. In *SDM*, pages 803–811, 2013.

[27] X. Zhu, Z. Huang, J. Cui, and H. T. Shen. Video-to-shot tag propagation by graph sparse group lasso. *IEEE Transactions on Multimedia*, 13(3):633 – 646, 2013.

[28] X. Zhu, Z. Huang, H. T. Shen, J. Cheng, and C. Xu. Dimensionality reduction by mixed kernel canonical correlation analysis. *Pattern Recognition*, 45(8):3003–3016, 2012.

[29] X. Zhu, Z. Huang, Y. Yang, H. T. Shen, C. Xu, and J. Luo. Self-taught dimensionality reduction on the high-dimensional small-sized data. *Pattern Recognition*, 46(1):215–229, 2013.