Prediction of Mild Cognitive Impairment Conversion Using Auxiliary Information

Xiaofeng Zhu
University of Electronic Science and Technology of China, Chengdu 611731, China.
seanzhuxf@gmail.com

Abstract

In this paper, we propose a new feature selection method to exploit the issue of High Dimension Low Sample Size (HDLSS) for the prediction of Mild Cognitive Impairment (MCI) conversion. Specially, by regarding the Magnetic Resonance Imaging (MRI) information of MCI subjects as the target data, this paper proposes to integrate auxiliary information with the target data in a unified feature selection framework for distinguishing progressive MCI (pMCI) subjects from stable MCI (sMCI) subjects, *i.e.,* the MCI conversion classification for short in this paper, based on their MRI information. The auxiliary information includes the Positron Emission Tomography (PET) information of the target data, the MRI information of Alzheimer's Disease (AD) subjects and Normal Control (NC) subjects, and the ages of the target data and the AD and NC subjects. As a result, the proposed method jointly selects features from the auxiliary data and the target data by taking into account the influence of outliers and aging of these two kinds of data. Experimental results on the public data of Alzheimer's Disease Neuroimaging Initiative (ADNI) verified the effectiveness of our proposed method, compared to three state-of-the-art feature selection methods, in terms of four classification evaluation metrics.

1 Introduction

Alzheimer's Disease (AD) is a chronic neurodegenerative disease to slowly and gradually worsen human brain over time, and has been becoming the most common dementia in many countries. By considering the development and prevalence of AD, the early stage of AD pathology, *i.e.,* Mild Cognitive Impairment (MCI), has been demonstrated to be the optimal stage that clinical treatments could be effectively investigated to prevent MCI conversion [Spasov *et al.*, 2019; Zhu *et al.*, 2016]. Clinically, it is essential to conduct MCI conversion classification by distinguishing pMCI subjects (who possibly progress to AD) from sMCI subjects (who may remain stable in the progress of MCI for a long time) based on their neuroimaging data [Forlenza *et al.*, 2019; Weiner *et al.*, 2017].

It is very challenging to differentiate the pMCI subjects from the sMCI subjects in an individual level based on their neuroimaging data due to the following reasons. First, the pMCI has small inter-group difference from the sMCI so that many studies of AD diagnosis integrate the pMCI with the sMCI as a single category, *i.e.,* MCI. Second, there is high intra-group variations for either pMCI subjects or sMCI subjects. That is, different subjects in the same sub-category (*i.e.,* either the pMCIs or the sMCIs) have high intra-group variations, which makes difficult construct robust classification models. Last but not least, the number of MCI subjects is small, but the dimensions of their features are usually high. In this case, the issue of High Dimension Low Sample Size (HDLSS) is often found so that the MCI conversion classification easily results in the problem of curse of dimensionality. As a result, previous classification models are usually affected by redundant features and subject-level noise. Hence, it is very vital to investigate informative and discriminative patterns to address above issues for achieving high diagnosis accuracy.

In the literature, machine learning techniques based on neuroimaging data, such as Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and CereboSpinal Fluid (CSF), have been proposed to address a part of above issues [Weiner *et al.*, 2017]. For example, [Zhu *et al.*, 2017] investigated a joint regression and classification model on both the MRI data and the PET data to conduct the MCI conversion classification. [Zhu *et al.*, 2014] considered the influence of ages (*e.g.,* taking the ages of the subjects as a feature) to detect the association between MRI data and genetic data, with the assumption that brain atrophy may be influenced by the normal aging as well as the AD [Moradi *et al.*, 2015]. [Wang *et al.*, 2017] first assumed that the relationship between the AD and the Normal Control (NC) is similar to the relationship between the pMCI (*i.e.,* AD-like) and the sMCI (*i.e.,* NC-like), and then employed the information of AD and NC to improve the robustness of the MCI conversion classification.

Previous machine learning techniques for the MCI conversion classification have the following common characteristics, *i.e.,* the auxiliary information is sequentially or partly used for constructing a prediction model. Normally, by regarding the MRI information of MCI subjects as the target data, the auxiliary information includes the ages of the target data, the MRI information of AD subjects and NC subjects, the PET and genetic information of the target data, *etc.* Moreover, each kind of auxiliary information is heterogenous to others (including other auxiliary information and the target data) as they are with different data structures and data distributions. Hence, integrating these information for the MCI conversion classification is complex as well as practical for the AD study.

In this paper, we consider the MRI data of the MCI subjects as the target data to propose a new sparse feature selection model for conducting the MCI conversion classification. Specifically, our proposed model integrates all kinds of auxiliary data with the target data for feature selection, it thus makes the best use of auxiliary information to improve the robustness of the classification on the target data. In our method, the auxiliary information includes the ages of both the target data and the auxiliary data (*i.e.,* the AD subjects and NC subjects), the PET information of the target data as the ADNI dataset only provided PET data for some subjects and provided MRI data for all subjects, and the MRI information of AD subjects and NC subjects. To do this, our proposed method includes three key factors, *i.e.,* (1) *feature selection*: we use all kinds of auxiliary data to select informative features of the target data; (2) *outlier influence reduce*: our formulation is robust to outliers for both the auxiliary data and the target data; (3) *aging effect removal*: we regard the ages of the subjects as one feature firstly, and then integrate it with neuroimaging features to form the feature matrix of our prediction model. Finally, we use our proposed feature selection model to select features from the target data, and then use a linear Support Vector Machine (SVM) to conduct the MCI conversion classification.

Compared to previous studies of the MCI conversion classification, the main contributions of our proposed method are summarized as follows.

- Our proposed method considers three kinds of auxiliary information to jointly select features in a unified framework. It is noteworthy that previous methods [Eskildsen *et al.*, 2013; Ye *et al.*, 2011; Cheng *et al.*, 2015; Moradi *et al.*, 2015] only used a part of them sequentially. We argue that auxiliary information can provide complementary information to the target data in some ways, while utilizing a part of all auxiliary information separately and sequentially is limited for the MCI conversion classification.
- Our proposed method conducts feature selection by taking into account the influence of outliers in both the auxiliary data and the target data as well as the relationship between the auxiliary data and the target data. By contrast, [Moradi *et al.*, 2015] used the auxiliary data for feature selection only, while [Cheng *et al.*, 2015] used both the target data and the auxiliary data, but not taking the influence of their outliers into account.
- Our proposed method integrates three kinds of auxiliary data with the target data to remove their heterogeneity. The experimental results on the public data of Alzheimer's Disease Neuroimaging Initiative (ADNI) verified the effectiveness of our proposed method, compared to the state-of-the-art feature selection methods. Obviously, our proposed framework can be easily applied for other methods for the MCI conversion classification [Zhu *et al.*, 2014; Moradi *et al.*, 2015] and previous methods of the AD study [Zhu *et al.*, 2017].

2 Method

In this paper, we denote matrices, vectors, and scalars, respectively, by boldface uppercase letters, boldface lowercase letters, and normal italic letters. Specifically, we denote the MRI feature matrix, the PET feature matrix, and the label matrix, respectively, for n_t subjects of pMCI and sMCI, as $\mathbf{X}_t \in \mathbb{R}^{n_t \times d}$, $\mathbf{X}_p \in \mathbb{R}^{n_t \times d}$, and $\mathbf{Y}_t \in \{0,1\}^{n_t \times c_t}$, where d denotes the number of features and c_t is the class number of the target data. We also denote the MRI feature matrix and the label matrix, of n_a subjects of AD and NC, respectively, as $\mathbf{X}_a \in \mathbb{R}^{n_a \times d}$ and $\mathbf{Y}_a \in \{0,1\}^{n_a \times c_a}$, where c_a is the class number of the auxiliary data. We further denote the age factors of the target data and the auxiliary data, as $\mathbf{x}_{tg} \in \mathbb{R}^{n_t}$ and $\mathbf{x}_{ag} \in \mathbb{R}^{n_a}$, respectively. In this work, we only focus on the binary classification problem, *i.e.,* $c_t = c_a = 2$. However, it is straightforward to extend our proposed framework to a multi-class classification problem.

2.1 Feature Selection on Target Data

Given the target data $\mathbf{X}_t$ and its corresponding label information matrix $\mathbf{Y}_t$, a robust sparse regression method linearly estimates a coefficient matrix $\mathbf{W}_t \in \mathbb{R}^{d \times c_t}$ by optimizing the following objective function:

$$\min_{\mathbf{W}_t} \|\mathbf{Y}_t - \mathbf{X}_t\mathbf{W}_t\|_{2,1} + \lambda\|\mathbf{W}_t\|_{2,1} \tag{1}$$

where λ is the non-negative tuning parameter. The $\ell_{2,1}$-norm loss function, *i.e.,* a robust loss function in first term of Eq. (1), makes Eq. (1) robust against the subject-level outliers. Specifically, each row of $(\mathbf{Y}_t - \mathbf{X}_t\mathbf{W}_t)$ in Eq. (1) corresponds to the prediction residual of one subject. Under the $\ell_{2,1}$-norm operation, the residual values of each row (*i.e.,* subject) are combined via ℓ_2-norm, *i.e.,* the square root of the sum of the squares, and thus are less affected by the outliers, compared to the least square loss function [Lei and Zhu, 2018]. The $\ell_{2,1}$-norm regularization term on $\mathbf{W}_t$ penalizes $\mathbf{W}_t$ by encouraging the row sparsity, *i.e.,* all elements of some rows of $\mathbf{W}_t$ are all zeros, to select the corresponding features in $\mathbf{X}_t$.

2.2 Feature Selection on Target and Auxiliary Data

Using Eq. (1) directly on the target data (*i.e.,* the MRI features of the pMCI and sMCI subjects) for the MCI conversion classification could still be ineffective due to the limited training data. To circumvent the lack of training samples, recent studies [Coupé *et al.*, 2012; Moradi *et al.*, 2015; Ye *et al.*, 2011; Young *et al.*, 2013] exploited auxiliary information from non-target groups, *e.g.,* AD and NC subjects. The rationale of using such auxiliary data is that in terms of the AD pathological spectrum, *i.e.,* the sMCI is closer to the NC while the pMCI is closer to the AD. Thus, the features that are informative for the AD/NC separation could be also useful for the pMCI/sMCI separation, *i.e.,* the MCI conversion classification [Coupé *et al.*, 2012; Young *et al.*, 2013]. In this paper, we also utilize such auxiliary data for feature selection. However, unlike previous methods [Ye *et al.*, 2011; Young *et al.*, 2013] that mostly first learned a classification model over only the auxiliary data and then transferred the learned model to build a target-oriented model, we devise a novel sparse feature selection model that jointly exploits both the target data and the auxiliary data.

With the assumption that MRI features selected for the AD/NC classification could be also informative for the MCI conversion classification, we propose to use MRI features of AD and NC subjects (*i.e.,* auxiliary data $\mathbf{X}_a$), to help in selecting MRI features of pMCI and sMCI subjects, as follows:

$$\begin{aligned}\min_{\mathbf{W}_t,\mathbf{W}_a} \; & \|\mathbf{Y}_t - \mathbf{X}_t\mathbf{W}_t\|_{2,1} + \|\mathbf{Y}_a - \mathbf{X}_a\mathbf{W}_a\|_{2,1} \\ & + \lambda_1\|[\mathbf{W}_t, \mathbf{W}_a]\|_{2,1} + \lambda_2\|[\mathbf{W}_t, \mathbf{W}_a]\|_F^2\end{aligned} \tag{2}$$

where $\mathbf{W}_a \in \mathbb{R}^{d\times c_a}$ is a coefficient matrix for the auxiliary data, λ_1 and λ_2 are non-negative tuning parameters. The reason to use $\ell_{2,1}$-norm on the loss function of auxiliary data (*i.e.,* the second term in Eq. (2)) is similar to Eq. (1), *i.e.,* for robustness to outliers.

The $\ell_{2,1}$-norm regularizer on $[\mathbf{W}_t, \mathbf{W}_a] \in \mathbb{R}^{d\times(c_t+c_a)}$ encourages the row-wise joint sparsity. This sparsity constraint encourages the same set of features to be selected for both $\mathbf{X}_t$ and $\mathbf{X}_a$ (recall that $\mathbf{X}_t$ and $\mathbf{X}_a$ denote the feature matrix for the target and auxiliary data, respectively). With the sparsity regularization term $\|[\mathbf{W}_t, \mathbf{W}_a]\|_{2,1}$, the useful features are kept by satisfying the AD/NC separation constraint (via $\mathbf{W}_a$) and the MCI conversion separation constraint (via $\mathbf{W}_t$), simultaneously. The jointly learned model is more robust than the individual models of either only satisfying the AD/NC separation constraint (via $\mathbf{W}_a$) [Coupé *et al.*, 2012; Moradi *et al.*, 2015] which does not consider the pathological difference in the pMCI and sMCI subjects, or only satisfying the MCI conversion separation constraint (via $\mathbf{W}_t$) in [Ye *et al.*, 2011; Young *et al.*, 2013] which has been reported to have limited performance due to the small number of subjects.

The Frobenius norm on $[\mathbf{W}_t \; \mathbf{W}_a]$ in the fourth term of Eq. (2) is used to provide a grouping effect, which tends to select highly correlated features together, by countering for some weaknesses of the sparsity constraint [Zou and Hastie, 2005]. By considering the time-consuming issue of the parameter tuning, we change Eq. (2) to the following objective function which makes the fourth term of Eq. (2) still provide a grouping effect and reduce the model complexity.

$$\begin{aligned}\min_{\mathbf{W}_t,\mathbf{W}_a} \; & \|\mathbf{Y}_t - \mathbf{X}_t\mathbf{W}_t\|_{2,1} + \|\mathbf{Y}_a - \mathbf{X}_a\mathbf{W}_a\|_{2,1} \\ & + \lambda_1\|[\mathbf{W}_t, \mathbf{W}_a]\|_{2,1} + \|[\mathbf{W}_t, \mathbf{W}_a]\|_F\end{aligned} \tag{3}$$

Different from Eq. (2), Eq. (3) does not need to tune the parameter λ_2 manually and changes the square operation of the Frobenius norm to the Frobenius norm only. To solve the optimization problem of Eq. (3), *i.e.,* optimizing either the variable $\mathbf{W}_t$ or the variable $\mathbf{W}_a$, we could compute the derivatives of the Frobenius norm in Eq. (3) to iteratively optimize the following problems [Zhu *et al.*, 2018].

$$\begin{cases}\min\limits_{\mathbf{W}_t,\mathbf{W}_a} \; \|\mathbf{Y}_t - \mathbf{X}_t\mathbf{W}_t\|_{2,1} + \|\mathbf{Y}_a - \mathbf{X}_a\mathbf{W}_a\|_{2,1} \\ \qquad + \lambda_1\|[\mathbf{W}_t, \mathbf{W}_a]\|_{2,1} + \lambda_2\|[\mathbf{W}_t, \mathbf{W}_a]\|_F^2 & \text{(4a)} \\ \lambda_2 = \dfrac{1}{2\|[\mathbf{W}_t, \mathbf{W}_a]\|_F} & \text{(4b)}\end{cases}$$

In the multiple-modality AD study, it has shown that the PET data and the MRI data could provide complementary information to each other [Cheng *et al.*, 2015; Moradi *et al.*, 2015; Young *et al.*, 2013]. In this paper, we use the PET data of the pMCI and sMCI subjects, *i.e.,* $\mathbf{X}_p$, as another kind of auxiliary data, to help learn the coefficient matrix $\mathbf{W}_t$ of the target data. More specifically, we constrain the predicted values from the PET data and the MRI data to be close to each other, as both modalities share the same label information. As a result, we have the following objective function

$$\min_{\mathbf{W}_t,\mathbf{W}_p} \|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1} + \lambda_3\|\mathbf{W}_p\|_{2,1} \tag{5}$$

where λ_3 is a nonnegative tuning parameter and $\mathbf{W}_p \in \mathbb{R}^{d\times c_t}$ is a coefficient matrix to the PET data. Note that $\mathbf{X}_t\mathbf{W}_t$ and $\mathbf{X}_p\mathbf{W}_p$ are the predictions of the label matrix using the MRI and PET data, respectively. Thus, their difference, measured by the summation of element-wise similarity, should be as small as possible.

Combining Eq. (3) with Eq. (5), we obtain the following objective function, which learns $\mathbf{W}_t$ with the auxiliary MRI data of AD/NC subjects and the PET data of pMCI/sMCI subjects,

$$\begin{aligned}\min_{\mathbf{W}_t,\mathbf{W}_a,\mathbf{W}_p} \; & \|\mathbf{Y}_t - \mathbf{X}_t\mathbf{W}_t\|_{2,1} + \|\mathbf{Y}_a - \mathbf{X}_a\mathbf{W}_a\|_{2,1} \\ & + \|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1} + \lambda_1\|[\mathbf{W}_t, \mathbf{W}_a]\|_{2,1} \\ & + \|[\mathbf{W}_t, \mathbf{W}_a]\|_F + \lambda_3\|\mathbf{W}_p\|_{2,1}.\end{aligned} \tag{6}$$

2.3 Aging Effect Removal

[Franke *et al.*, 2010; Moradi *et al.*, 2015] showed that both the normal aging and the AD pathology contributed to brain atrophy and it is necessary to remove the aging effect to the brain atrophy before analysis. The first method designed for the aging effect removal fits a linear regression model between the features and the age of NC subjects to obtain a coefficient matrix [Franke *et al.*, 2010; Moradi *et al.*, 2015]. This coefficient matrix denotes how the age affects the feature values. The second method directly fits the model by using both features and the age as covariates [Moradi *et al.*, 2015; Zhu *et al.*, 2014]. Actually, both of them assume that there is linear relationship among the labels, features and ages. Hence, we use the ages of the subjects as one feature in both the target data and the auxiliary data to have our final objective function as follows.

$$\begin{aligned}\min_{\mathbf{W}_t,\mathbf{w}_{tg},\mathbf{W}_a,\mathbf{w}_{ag},\mathbf{W}_p} \; & \|\mathbf{Y}_t - [\mathbf{X}_t, \mathbf{x}_{tg}][\mathbf{W}_t^\top, \mathbf{w}_{tg}^\top]^\top\|_{2,1} \\ & + \|\mathbf{Y}_a - [\mathbf{X}_a, \mathbf{x}_{ag}][\mathbf{W}_a^\top, \mathbf{w}_{ag}^\top]^\top\|_{2,1} \\ & + \|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1} + \lambda_1\|[\mathbf{W}_t, \mathbf{W}_a]\|_{2,1} \\ & + \|[\mathbf{W}_t, \mathbf{w}_{tg}, \mathbf{W}_a, \mathbf{w}_{ag}]\|_F + \lambda_3\|\mathbf{W}_p\|_{2,1}\end{aligned} \tag{7}$$

where $\mathbf{w}_{tg} \in \mathbb{R}^{1\times c_t}$ and $\mathbf{w}_{ag} \in \mathbb{R}^{1\times c_a}$ are coefficient matrices, and λ_1 and λ_3 are non-negative tuning parameters. In Eq.

(7), the fourth and fifth terms help select common useful features for the first two data fitting terms, while the third term imposes label prediction consistency between $\mathbf{X}_t$ and $\mathbf{X}_p$. In addition, the use of the $\ell_{2,1}$-norm loss function helps to learn $\mathbf{X}_t$, $\mathbf{X}_p$, and $\mathbf{X}_a$ by reducing the influence of outliers.

2.4 Optimization

We employ the alternating optimization strategy to solve Eq. (7), by iteratively optimizing each of the parameters (*i.e.,* $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_p$) while fixing the other parameters. We list the pseudo of our method in Algorithm 1 and explain the detail as follows.

i) Update $\mathbf{w}_{tg}$ and $\mathbf{w}_{ag}$ by fixing $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_p$

With the fixed $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_p$, the optimization with respect to the variables $\mathbf{w}_{tg}$ and $\mathbf{w}_{ag}$ are independent to each other. Thus we individually set the derivative of Eq. (7) with respect to $\mathbf{w}_{tg}$ and $\mathbf{w}_{ag}$ to zero to obtain

$$\begin{aligned}\hat{\mathbf{w}}_{tg} &= (\mathbf{x}_{tg}^\top \mathbf{A}\mathbf{x}_{tg} + \lambda_2)^{-1}\mathbf{x}_{tg}^\top \mathbf{A}(\mathbf{Y}_t - \mathbf{X}_t\mathbf{W}_t)\\ \hat{\mathbf{w}}_{ag} &= (\mathbf{x}_{ag}^\top \mathbf{B}\mathbf{x}_{ag} + \lambda_2)^{-1}\mathbf{x}_{ag}^\top \mathbf{B}(\mathbf{Y}_a - \mathbf{X}_a\mathbf{W}_a)\end{aligned} \tag{8}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times n}$, respectively, are diagonal matrices and their respective diagonal elements are defined as $a_{jj} = \frac{1}{2\|(\mathbf{Y}_t - [\mathbf{X}_t, \mathbf{x}_{tg}][\mathbf{W}_t^\top, \mathbf{w}_{tg}^\top]^\top)^j\|_2^2}$ and $b_{jj} = \frac{1}{2\|(\mathbf{Y}_a - [\mathbf{X}_a, \mathbf{x}_{ag}][\mathbf{W}_a^\top, \mathbf{w}_{ag}^\top]^\top)^j\|_2^2}, j = 1, ..., n$.

(ii) Update $\mathbf{W}_t$ by fixing $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_a$, and $\mathbf{W}_p$

Given $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_a$, and $\mathbf{W}_p$, we can rewrite Eq. (7) as follows:

$$\begin{aligned}\min_{\mathbf{W}_t} &\|\mathbf{Y}_t - [\mathbf{X}_t, \mathbf{x}_{tg}][\mathbf{W}_t^\top, \mathbf{w}_{tg}^\top]^\top\|_{2,1}\\ &+ \|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1} + \lambda_1\|[\mathbf{W}_t, \mathbf{W}_a]\|_{2,1}\\ &+ \|[\mathbf{W}_t, \mathbf{w}_{tg}, \mathbf{W}_a, \mathbf{w}_{ag}]\|_F.\end{aligned} \tag{9}$$

By setting the derivative of Eq. (9) with respect to $\mathbf{W}_t$ to zero and solving the resulting equations, we can obtain

$$\hat{\mathbf{W}}_t = \mathbf{G}^{-1}\mathbf{H} \tag{10}$$

where $\mathbf{G} = (\mathbf{X}_t^\top(\mathbf{A} + \mathbf{C})\mathbf{X}_t + \lambda_1\mathbf{U} + \lambda_2\mathbf{I}_d)$ and $\mathbf{H} = (\mathbf{X}_t^\top\mathbf{A}(\mathbf{Y}_t - \mathbf{x}_{tg}\mathbf{w}_{tg}) + \mathbf{X}_t^\top\mathbf{C}\mathbf{X}_p\mathbf{W}_p)$, $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is an identity matrix, and $\mathbf{C} \in \mathbb{R}^{n\times n}$ and $\mathbf{U} \in \mathbb{R}^{d\times d}$ are diagonal matrices and their respective diagonal elements are $c_{jj} = \frac{1}{2\|(\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p)^j\|_2^2}, j = 1, ..., n$ and $u_{kk} = \frac{1}{2\|(\mathbf{W}_t, \mathbf{W}_a)^k\|_2^2}$, $k = 1, ..., d$, respectively.

(iii) Update $\mathbf{W}_a$ by fixing $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, and $\mathbf{W}_p$

With fixed $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, and $\mathbf{W}_p$, Eq. (7) becomes:

$$\begin{aligned}\min_{\mathbf{W}_a} &\|\mathbf{Y}_a - [\mathbf{X}_a, \mathbf{x}_{ag}][\mathbf{W}_a^\top, \mathbf{w}_{ag}^\top]^\top\|_{2,1}\\ &+ \lambda_1\|[\mathbf{W}_t, \mathbf{W}_a]\|_{2,1} + \|[\mathbf{W}_t, \mathbf{w}_{tg}, \mathbf{W}_a, \mathbf{w}_{ag}]\|_F.\end{aligned} \tag{11}$$

By setting the derivative of Eq. (11) with respect to $\mathbf{W}_t$ to zero and solving the equations, we obtain

$$\hat{\mathbf{W}}_a = (\mathbf{X}_a^\top\mathbf{B}\mathbf{X}_t + \lambda_1\mathbf{U} + \lambda_2\mathbf{I}_d)^{-1}\mathbf{X}_a^\top\mathbf{B}(\mathbf{Y}_t - \mathbf{x}_{tg}\mathbf{w}_{tg}). \tag{12}$$

Algorithm 1: The pseudo of solving Eq. (7).

```
Input: $\mathbf{X}_t$, $\mathbf{X}_p$, $\mathbf{Y}_t$, $\mathbf{X}_a$, $\mathbf{Y}_a$, $\mathbf{x}_{tg}$, $\mathbf{x}_{ag}$, $\lambda_1$, and $\lambda_3$;
Output: $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_p$;
Randomly initialize $\mathbf{W}_t$, $\mathbf{w}_{tg}$, $\mathbf{W}_p$, $\mathbf{W}_a$, and $\mathbf{w}_{ag}$;
Initialize $\lambda_2$ via $\lambda_2 = \frac{1}{2\|[\mathbf{W}_t, \mathbf{w}_{tg}, \mathbf{W}_a, \mathbf{w}_{ag}]\|_F}$;
repeat
    Calculate the diagonal matrix $\mathbf{A}$ with
      $a_{jj} = \frac{1}{2\|(\mathbf{Y}_t - [\mathbf{X}_t, \mathbf{x}_{tg}][\mathbf{W}_t^\top, \mathbf{w}_{tg}^\top]^\top)^j\|_2^2}, j = 1, ..., n$;
    Calculate the diagonal matrix $\mathbf{B}$ with
      $b_{jj} = \frac{1}{2\|(\mathbf{Y}_a - [\mathbf{X}_a, \mathbf{x}_{ag}][\mathbf{W}_a^\top, \mathbf{w}_{ag}^\top]^\top)^j\|_2^2}, j = 1, ..., n$;
    Calculate the diagonal matrix $\mathbf{C}$ with
      $c_{jj} = \frac{1}{2\|(\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p)^j\|_2^2}, j = 1, ..., n$;
    Calculate the diagonal matrix $\mathbf{U}$ with
      $u_{kk} = \frac{1}{2\|(\mathbf{W}_t, \mathbf{W}_a)^j\|_2^2}, k = 1, ..., d$;
    Calculate the diagonal matrix $\mathbf{V}$ with
      $v_{kk} = \frac{1}{2\|\mathbf{W}_p^j\|_2^2}, k = 1, ..., d$;
    Update $\mathbf{w}_{tg}$ via Eq. (8);
    Update $\lambda_2$ via $\lambda_2 = \frac{1}{2\|[\mathbf{W}_t, \mathbf{w}_{tg}, \mathbf{W}_a, \mathbf{w}_{ag}]\|_F}$;
    Update $\mathbf{w}_{ag}$ via Eq. (8);
    Update $\lambda_2$ via $\lambda_2 = \frac{1}{2\|[\mathbf{W}_t, \mathbf{w}_{tg}, \mathbf{W}_a, \mathbf{w}_{ag}]\|_F}$;
    Update $\mathbf{W}_t$ via Eq. (10);
    Update $\lambda_2$ via $\lambda_2 = \frac{1}{2\|[\mathbf{W}_t, \mathbf{w}_{tg}, \mathbf{W}_a, \mathbf{w}_{ag}]\|_F}$;
    Update $\mathbf{W}_a$ via Eq. (12);
    Update $\lambda_2$ via $\lambda_2 = \frac{1}{2\|[\mathbf{W}_t, \mathbf{w}_{tg}, \mathbf{W}_a, \mathbf{w}_{ag}]\|_F}$;
    Update $\mathbf{W}_p$ via Eq. (14);
until Eq. (7) converges;
```

(iv) Update $\mathbf{W}_p$ by fixing $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, and $\mathbf{W}_a$

Given $\mathbf{w}_{tg}$, $\mathbf{w}_{ag}$, $\mathbf{W}_t$, and $\mathbf{W}_a$, Eq. (7) becomes:

$$\min_{\mathbf{W}_p} \|\mathbf{X}_t\mathbf{W}_t - \mathbf{X}_p\mathbf{W}_p\|_{2,1} + \lambda_3\|\mathbf{W}_p\|_{2,1}. \tag{13}$$

By setting the derivative of Eq. (13) with respect to $\mathbf{W}_t$ to zero and solving the equations, we obtain

$$\hat{\mathbf{W}}_p = (\mathbf{X}_p^\top\mathbf{C}\mathbf{X}_p + \lambda_3\mathbf{V})^{-1}\mathbf{X}_p^\top\mathbf{C}\mathbf{X}_t\mathbf{W}_t \tag{14}$$

where $\mathbf{V} \in \mathbb{R}^{d\times d}$ is a diagonal matrix whose diagonal elements are defined as $v_{kk} = \frac{1}{2\|(\mathbf{W}_p)^k\|_2^2}, k = 1, ..., d$.

3 Experiments

We evaluated our proposed method by comparing with three state-of-the-art feature selection methods and one baseline method on two data sets in terms of four classification evaluation metrics.

3.1 Data Sets

In this work, we used the ADNI 1 ('`www.adni-info.org`') publicly available on the web for research purposes to generate the binary classification task on two data sets: 1) 'Data1' consisted of 93 AD, 99 NC, 55 pMCI, and 59 sMCI subjects, and 2) 'Data2' consisted of 50 AD, 51 NC, 31 pMCI, and 30 sMCI subjects.

We first preprocessed the MRI and PET images by sequentially applying spatial distortion correction, skull-stripping, and cerebellum removal, followed by segmenting the MRI images into gray matter, white matter, and cerebrospinal fluid, and then warped them into the AAL template to obtain 90 regions. We further aligned the PET images to their respective MRI images. We finally obtained 90 gray matter volumes from a MRI image and 90 mean intensities from a PET image and used them for features.

3.2 Comparison Methods

We defined a baseline model that utilized the original features for classification (thus denoted as 'Original') and also considered other state-of-the-art feature selection methods, namely, general sparsity regularized feature selection (GSR) [Peng and Fan, 2017], Semi-Supervised Learning (SSL) [Moradi *et al.*, 2015], and Domain Transfer Learning (DTL) [Cheng *et al.*, 2015].

The baseline method 'Original' used all target data to perform classification without removing any features. GSR conducted feature selection by optimizing an $\ell_{2,r}$-norm ($0 < r \leq 2$) loss function and an $\ell_{2,p}$-norm ($0 < p \leq 1$) regularization term to reduce the influence of subject-level outliers. In our experiments, we considered to form its two variants: 'GSR-Pre' (using the target data alone) and 'GSR-Aux' (using the auxiliary data of the AD and NC subjects alone [Ye *et al.*, 2011]). The SSL method sequentially performed the aging effect removal and feature selection using the AD and NC subjects. DTL method conducted feature selection using both the target data and the auxiliary data, without taking into account ageing effect removal and the robustness against outliers in the data.

It is noteworthy that all comparison methods did not take age effect removal into account. Hence, we used 'Pro-NoAge' to denote our method without taking aging effect removal into account, *i.e.,* Eq. (6), and 'Proposed' as our proposed model, *i.e.,* Eq. (7).

3.3 Experimental Setting

We repeated the 10-fold cross-validation scheme 100 times on all methods, each of which conducted 5-fold nested cross-validations for model selection. The ranges of parameters of every comparison method were set by strictly following the corresponding literature so that they outputted the best results in our experiments. We used the method of grid search with the search range of $\{10^{-5}, ..., 10^{5}\}$ to conduct model selection in our two proposed methods.

In this paper, we further partitioned each of two data sets into two subsets of the target data, *i.e.,* 'MRI', and 'MRI + PET', respectively, to indicate single modality targets (only MRI features) and multi-modality targets (MRI and PET features).

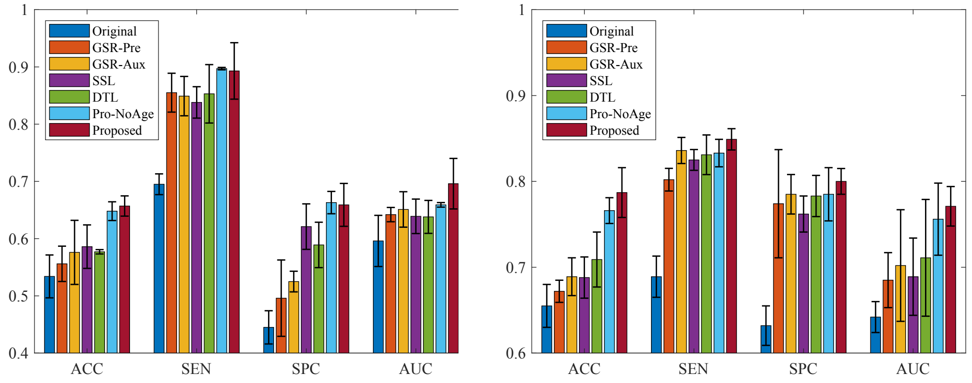
In our experiments, after conducting feature selection, we used the Support Vector Machine (SVM) [Chang and Lin, 2011] to conduct the classification tasks, where we set the parameter C as the range of $C \in \{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$ in the SVM for all methods. We used classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the receiver operating characteristic curve (AUC), to evaluate the classification performance.

3.4 Results Analysis

Figures 1 and 2 listed the classification performance of all methods on the data sets 'Data1' and 'Data2', respectively. We listed our observations as follows.

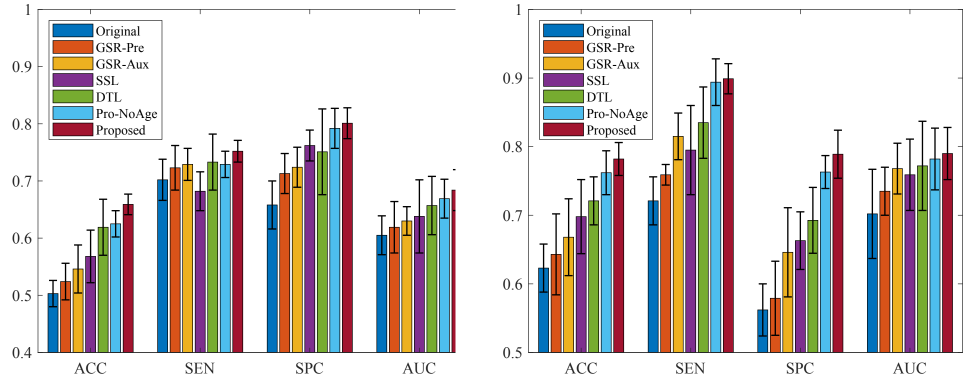
- Our proposed method (*i.e.,* Proposed) achieved the best performance, followed by Pro-NoAge, DTL, SSL, GSR-Aux, GRS-Pre, and Original. Specifically, Proposed improved the classification performance on average by 4.98% compared to the best comparison method, *i.e.,* DTL, while improved on averages by 14.61% compared to the worst comparison method, *i.e.,* Original, in terms of all four evaluation metrics. Moreover, our proposed Pro-NoAge improved the classification performance on average by 3.44%, compared to the best comparison method, *i.e.,* DTL. This demonstrated that either outlier influence reduce or aging effect removal is necessary for the MCI conversion classification.
- All methods achieved larger improvement (in comparison with Original) on Data2, compared to their corresponding improvement on Data1. This may imply that the auxiliary information can improve the prediction ability of the target data, especially when the sample size of the target data is small and the sample size of the auxiliary data are the same
- Proposed outperformed Pro-NoAge method. For example, Proposed improved the classification performance on average by 1.54% compared to Pro-NoAge, in terms of four evaluation metrics. This is consistent with the previous study of [Moradi *et al.*, 2015] that has validated the importance of removing aging effect.
- When regarding the use of the auxiliary data from the AD and NC subjects, GSR-Aux consistently outperformed its counterpart GSR-Pre in all experiments. Specifically, GSR-Aux used the AD and NC subjects to construct the AD/NC classifier to classify the target data, *i.e.,* distinguishing pMCI subjects from sMCI subjects, while GSR-Pre employs the pMCI and sMCI subjects to classify the target data. In our experiments, the AD/NC classifier improves the classification performance by 2.26% in terms of all four evaluation metrics since these two methods select different features to conduct classification tasks. It is noteworthy that the number of features selected by GSR-Pre was larger than the number of features selected by GSR-Aux because GSR-Pre could not capture subtle structure differences among region-of-interests (ROIs) with a limited number of high-dimensional samples.

Finally, we analyzed the selected features by all methods. From our experiments, we could observe that most comparison methods selected similar brain regions as top regions, such as lateral ventricle right, globus palladus left/right, subthalamic nucleus right, uncus right, occipital lobe WM right, nucleus accumbens left, occipital lobe WM left, and fornix right, for the MRI features. Those selected ROIs



(a) MRI (b) MRI + PET

Figure 1: Classification performance of all methods on Data1 with 50 pMCIs and 51 sMCIs.



(a) MRI (b) MRI + PET

Figure 2: Classification performance of all methods on Data2 with 31 pMCIs and 30 sMCIs.

were also verified to be related to AD [Cheng *et al.*, 2015; Misra *et al.*, 2009]. It is noteworthy that our method selected some ROIs more often than the comparison methods, such as parahippocampal gyrus left, hippocampal formation right, middle temporal gyrus left, perirhinal cortex left, temporal pole left, entorhinal cortex left, lateral occipitotemporal gyrus right, hippocampal formation left, amygdala left, parahippocampal gyrus right, middle temporal gyrus right, amygdala right, inferior temporal gyrus right, and lateral occipitotemporal gyrus left [Cheng *et al.*, 2015; Misra *et al.*, 2009]. We believe the selection of those ROIs contributed to enhance the performance in our method. In the mean time, none of comparison methods selected the ROIs of angular gyrus right and postcentral gyrus left from MRI and the ROIs of nucleus accumbens left, lingual gyrus right, and thalamus right from PET.

4 Conclusion

In this paper, we have proposed to use the auxiliary information to improve the diagnostic accuracy in pMCI and sMCI identification. The proposed method used three ways to incorporate the auxiliary data with the target data in a unified framework, *i.e.,* using an $\ell_{2,1}$-norm on the weight matrices (for joint feature selection), using an $\ell_{2,1}$-norm loss function (for outliers robustness), and including the age factor in the feature matrix (for removing aging-effect). Finally, experimental results on ADNI 1 verified the effectiveness of our proposed method, compared to the comparison methods, in terms of classification tasks.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grants No: 61876046 and 61573270), the Guangxi High Institutions Program of Introducing 100 High-Level Overseas Talents, the Strategic Research Excellence Fund at Massey University, and the Marsden Fund of New Zealand (Grant No: MAU1721). The author thanks K. H. Thung and D. Shen for their efforts on the initial draft of this work as well as the Alzheimer's Disease Neuroimaging Initiative for providing the data sets.

References

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines.

ACM Transactions on Intelligent Systems and Technology, 2(3):27, 2011.

[Cheng *et al.*, 2015] Bo Cheng, Mingxia Liu, Daoqiang Zhang, Brent C Munsell, and Dinggang Shen. Domain transfer learning for MCI conversion prediction. *IEEE Transactions on Biomedical Engineering*, 62(7):1805–1817, 2015.

[Coupé *et al.*, 2012] Pierrick Coupé, Simon F Eskildsen, José V Manjón, et al. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. *NeuroImage: Clinical*, 1(1):141–152, 2012.

[Eskildsen *et al.*, 2013] Simon F Eskildsen, Pierrick Coupé, Daniel García-Lorenzo, et al. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *NeuroImage*, 65:511–521, 2013.

[Forlenza *et al.*, 2019] Orestes V Forlenza, Márcia Radanovic, Leda L Talib, and Wagner F Gattaz. Clinical and biological effects of long-term lithium treatment in older adults with amnestic mild cognitive impairment: randomised clinical trial. *The British Journal of Psychiatry*, pages 1–7, 2019.

[Franke *et al.*, 2010] Katja Franke, Gabriel Ziegler, Stefan Klöppel, et al. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50(3):883–892, 2010.

[Lei and Zhu, 2018] Cong Lei and Xiaofeng Zhu. Unsupervised feature selection via local structure learning and sparse learning. *Multimedia Tools and Applications*, 77(22):29605–29622, 2018.

[Misra *et al.*, 2009] Chandan Misra, Yong Fan, and Christos 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.

[Moradi *et al.*, 2015] Elaheh Moradi, Antonietta Pepe, Christian Gaser, et al. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*, 104:398–412, 2015.

[Peng and Fan, 2017] Hanyang Peng and Yong Fan. A general framework for sparsity regularized feature selection via iteratively reweighted least square minimization. In *AAAI*, pages 2471–2477, 2017.

[Spasov *et al.*, 2019] Simeon Spasov, Luca Passamonti, Andrea Duggento, Pietro Lio, Nicola Toschi, Alzheimer's Disease Neuroimaging Initiative, et al. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease. *Neuroimage*, 189:276–287, 2019.

[Wang *et al.*, 2017] Rong Wang, Feiping Nie, Richang Hong, Xiaojun Chang, Xiaojun Yang, and Weizhong Yu. Fast and orthogonal locality preserving projections for dimensionality reduction. *IEEE Transactions on Image Processing*, 26(10):5019–5030, 2017.

[Weiner *et al.*, 2017] Michael W. Weiner, Dallas P. Veitch, Paul S. Aisen, Laurel A. Beckett, Nigel J. Cairns, Robert C. Green, Danielle Harvey, Clifford R. Jack, William Jagust, John C. Morris, Ronald C. Petersen, Andrew J. Saykin, Leslie M. Shaw, Arthur W. Toga, and John Q. Trojanowski. Recent publications from the alzheimer's disease neuroimaging initiative: Reviewing progress toward improved ad clinical trials. *Alzheimer's & Dementia*, 13(4):e1 – e85, 2017.

[Ye *et al.*, 2011] Dong Hye Ye, Kilian M Pohl, and Christos Davatzikos. Semi-supervised pattern classification: application to structural MRI of Alzheimer's disease. In *PRNI*, pages 1–4, 2011.

[Young *et al.*, 2013] Jonathan Young, Marc Modat, Manuel J Cardoso, et al. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2:735–745, 2013.

[Zhu *et al.*, 2014] Hongtu Zhu, Zakaria Khondker, Zhaohua Lu, and Joseph G Ibrahim. Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association*, 109(507):977–990, 2014.

[Zhu *et al.*, 2016] Xiaofeng Zhu, Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen. Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. *IEEE Transactions on Biomedical Engineering*, 63(3):607–618, 2016.

[Zhu *et al.*, 2017] Xiaofeng Zhu, Heung-Il Suk, Li Wang, Seong-Whan Lee, Dinggang Shen, et al. A novel relational regularization feature selection method for joint regression and classification in ad diagnosis. *Medical image analysis*, 38:205–214, 2017.

[Zhu *et al.*, 2018] Xiaofeng Zhu, Hongming Li, and Yong Fan. Parameter-free centralized multi-task learning for characterizing developmental sex differences in resting state functional connectivity. In *AAAI*, 2018.

[Zou and Hastie, 2005] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.