Alzheimer's Brain Network Analysis Using Sparse Learning Feature Selection

Lixin Cui[1], Lichi Zhang[2⋆], Lu Bai[1⋆⋆], Yue Wang[1], and Edwin R. Hancock[3]

[1]Central University of Finance and Economics, Beijing, China
[2]Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]University of York, York, UK

Abstract. Accurate identification of Mild Cognitive Impairment (MCI) based on resting-state functional Magnetic Resonance Imaging (RS-fMRI) is crucial for reducing the risk of developing Alzheimer's disease (AD). In the literature, functional connectivity (FC) is often used to extract brain network features. However, it still remains challenging for the estimation of FC because RS-fMRI data are often high-dimensional and small in sample size. Although various Lasso-type sparse learning feature selection methods have been adopted to identify the most discriminative features for brain disease diagnosis, they suffer from two common drawbacks. First, Lasso is instable and not very satisfactory for the high-dimensional and small sample size problem. Second, existing Lasso-type feature selection methods have not simultaneously encapsulate the joint correlations between pairwise features and the target, the correlations between pairwise features, and the joint feature interaction into the feature selection process, thus may lead to suboptimal solutions. To overcome these issues, we propose a novel sparse learning feature selection method for MCI classification in this work. It unifies the above measures into a minimization problem associated with a least square error and an Elastic Net regularizer. Experimental results demonstrate that the diagnosis accuracy for MCI subjects can be significantly improved using our proposed feature selection method.

Keywords: Alzheimer's Disease, Feature Selection, Elastic Net

1 Introduction

Alzheimer's disease (AD) is the most common form of dementia in old people, which severely interferes with their daily life and may eventually cause death [3]. Effective and accurate diagnosis of AD at its early stage may possess crucial significance in preventing progression of detrimental symptoms [3]. Recently, the identification of MCI subjects is important for reducing the risk of developing AD and has attracted much attention recently [11]. However, it is very challenging to identify MCI subjects due to its mild clinical symptoms.

⋆ Co-First Author
⋆⋆ Corresponding Author: bailucs@cufe.edu.cn

In the literature, MCI is generally believed to be associated with a disconnection syndrome within brain networks. Therefore, constructing brain functional connectivity (FC) networks based on the resting-state fMRI (RS-fMRI) BOLD signals of various brain regions has become promising for MCI classification. In this paper, we use a sliding window approach [9] to partition the RS-fMRI BOLD signal from each voxel into multiple overlapping segments, in order to capture the time-varying interactions between different ROIs and obtain a series of dynamic FC networks. We then extract the corresponding FC features for the subsequent brain network analysis. However, the number of the extracted features is much larger than that of the MCI subjects, and more importantly, many features may be irrelevant to the classification task, thus leading to the overfitting problem.

In pattern recognition and machine learning, feature selection are powerful tools for identifying the most salient features from the original feature space and alleviating the overfitting problem [10]. In this regard, various feature selection methods have been widely applied to detect the most discriminative features for AD prediction. In some early works, Chyzhyk et al. [4] proposed an evolutionary wrapper feature selection using Extreme Learning Machines to determine the most salient features for AD diagnosis. However, wrapper methods are often computational burdensome and the results are biased depending on the classifier [6]. To overcome these issues, many efforts have been devoted to developing LASSO-type feature selection methods for AD diagnosis. For instance, Suk et al. [7] utilized a group sparse representation along with a structural equation model to estimate FC from RS-fMRI. Wee et al. [9] proposed a fused sparse learning algorithm for early MCI identification. Chen et al. [3] developed a two-stage feature selection procedure to select a subset of the original features for MCI classification. However, existing LASSO-type feature selection methods for MCI classification suffer from two common limitations. First, LASSO shows instability and is not very satisfactory for high-dimensional small sample size problem. Second, existing Lasso-type feature selection methods have not simultaneously encapsulate the joint correlations between pairwise features and the target, the correlations between pairwise features, and the joint feature interaction into the feature selection process, thus may lead to suboptimal solutions.

To effectively tackle the issues of existing Lasso-type sparse learning feature selection methods, we propose a new feature selection method, i.e., Tripple-EN for MCI classification. We commence by defining three new information theoretic criteria to measure: 1)the relevancy of pairwise features in relation to the target feature, 2) the redundancy of pairwise features and 3)joint feature interaction. With these measures to hand, we formulate the corresponding feature subset selection problem as a least square regression model associated with an elastic net regularizer to simultaneously maximize relevancy, minimize redundancy, and maximize joint interaction of the selected features. An iterative optimization algorithm based on the alternating direction method of multipliers (ADMM) [1] is proposed to solve the optimization problem.

The advantages of the proposed method are twofold. First, it encapsulates the pairwise feature relevancy, feature redundancy and joint feature interaction into a unified learning model to improve the performance of feature selection. Second, by using the elastic net regularizer, the proposed method can ensure sparsity and also promote a

grouping effect of the features. Figure 1 shows an overview of the framework of this paper, which consists of the following steps: (1) constructing brain FC networks using a sliding window strategy, (2) identifying the most discriminative features using a new sparse learning feature selection method, and (3) implementing classification following the C-SVM method. Details of each step are illustrated in the following sections.

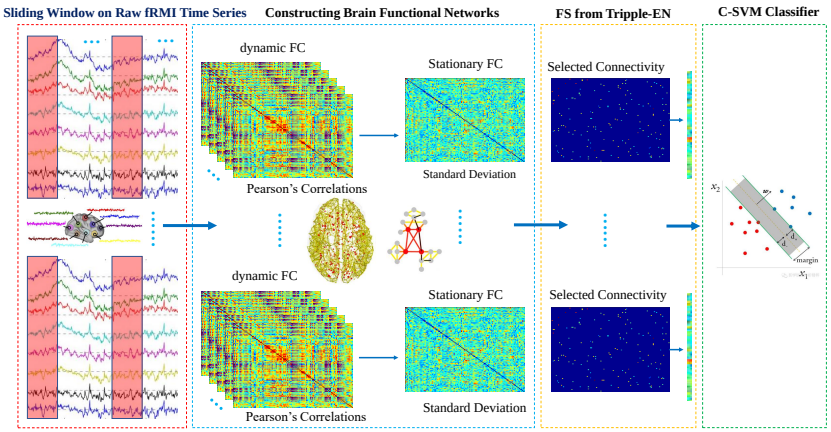


Fig. 1. Framework of this paper.

This paper is organized as follows. Section 2 introduces the construction of the functional connectivity networks from brain networks. Section 2 illustrates the proposed sparse learning feature selection method for MCI classification. Finally, Section 5 concludes this paper.

2 Constructing Functional Connectivity Networks

In this section, we will introduce how to construct the functional connectivity networks, which mainly consists of two steps, i.e., generating functional networks using RS-fMRI and feature extraction.

2.1 Generating FC Networks using RS-fMRI

As in Fig 1, the preprocessed RS-fMRI data was parcellated using the Automated Anatomical Labeling (AAL) atlas with 116 ROIs [3], which are represented by the time series curves of different colors. We use a sliding window approach to partition the RS-fMRI BOLD signal from each voxel into multiple overlapping segments, in order to capture the time-varying interactions between different ROIs. Specifically, denote the total length of image volumes as M and the length of the sliding windows as N. Then, the total number of segments is $K = \lfloor (M - N)/s \rfloor$. On each segment, within the GM, a regional mean BOLD signal is computed by averaging the BOLD time series over all voxels inside each ROI, which reflects the regional neural activity during a short period of time. We use C_{ij}^k to denote the Pearson's correlation coefficients between ROI i and ROI j on the k-th sliding window. Then we can obtain the interregional dynamic FC

(dFC), denoted as $dFC_{ij} = [C_{ij}^1, ..., C_{ij}^k, ..., C_{ij}^K]$, which measures the time-varying interactions of FC between ROI i and ROI j. As shown in Figure 1, we can obtain a series of dynamic time-varying FC networks. Note that, due to the symmetry of Pearson's correlation, the number of dFC is equal to the total number of ROI pairs.

2.2 Feature Extraction

To extract the features for further analysis, we calculate the standard deviation of a series of dynamic FC networks and obtain one stationary FC network for each subject. Specifically, the corresponding FC network for a series of dynamic time varying networks is obtained by calculating the standard deviation as $\sqrt{\frac{\sum_{k=1}^{K}(C_{ij}^k-\mu)^2}{K}}$, where μ is the mean value of C_{ij}^k. With these FC networks to hand, a total of 6670 features was generated. As shown in Figure 1, for a series of dynamic time-varying FC networks, we can construct a stationary FC network for each subject, with each node representing a specific ROI and each edge representing the corresponding connection between pairwise ROIs, which incorporates the information from a series of dynamic time-varying FC networks.

3 The Proposed Sparse Learning Feature Selection for MCI Classification

In this section, we focus on the proposed sparse learning feature selection method for identifying the most discriminative FC features. We commence by introducing the proposed information theoretic criteria for measuring the joint relevance (significance) of different pairwise feature combinations with respect to target labels, the redundancy of pairwise features, and the joint feature interaction, respectively. Based on these measures, we develop the corresponding optimization model for feature selection and sparse learning. Finally, an iterative optimization algorithm based on ADMM is proposed to solve the feature selection problem and identify the most discriminative feature subset.

3.1 Proposed Information Theoretic Criteria

Feature Relevancy. For a set of N features $\mathbf{f}_1, \ldots, \mathbf{f}_i, \ldots, \mathbf{f}_N$ and the associated target feature $\mathbf{Y}$, the relevancy degree of each feature pair $\{\mathbf{f}_i, \mathbf{f}_j\}$ in relation to the target feature is estimated through Pearson's correlation coefficients, which is defined as

$$W_{(\mathbf{f}_i,\mathbf{f}_j)} = Cor(\mathbf{f}_i, \mathbf{Y}) \times Cor(\mathbf{f}_j, \mathbf{Y}). \tag{1}$$

where Cor is the Pearson's correlation measure. The first term $Cor(\mathbf{f}_i, \mathbf{Y})$ measures the relevance of feature $\mathbf{f}_i$ with respect to the target. Similarly, the second term is the corresponding relevance of feature $\mathbf{f}_j$ with respect to the target. Therefore, $W_{(\mathbf{f}_i,\mathbf{f}_j)}$ is large if and only if both $Cor(\mathbf{f}_i, \mathbf{Y})$ and $Cor(\mathbf{f}_j, \mathbf{Y})$ are large (i.e., both $\mathbf{f}_i$ and $\mathbf{f}_j$ are informative themselves with respect to the target).

Additionally, it is desirable that strongly correlated features should not be in the model together, i.e., the selected features should be less redundant. Therefore, we propose the following criterion to measure the redundancy of pairwise features.

Feature Redundancy. For a set of N features $\mathbf{f}_1, \ldots, \mathbf{f}_i, \ldots, \mathbf{f}_N$, the redundancy of the feature pair $\{\mathbf{f}_i, \mathbf{f}_j\}$ is calculated as

$$U_{(\mathbf{f}_i, \mathbf{f}_j)} = Cor(\mathbf{f}_j, \mathbf{f}_j) \tag{2}$$

where Cor is the Pearson's correlation measure.

Joint Feature Interaction. We propose to use the following criterion to measure the joint interaction of different pairwise feature combinations with respect to target labels. For a set of N features $\mathbf{f}_1, \ldots, \mathbf{f}_i, \ldots, \mathbf{f}_j, \ldots, \mathbf{f}_N$ and the associated continuous target feature $\mathbf{Y}$, the joint interaction degree of the feature pair $\{\mathbf{f}_i, \mathbf{f}_j\}$ is

$$V_{\mathbf{f}_i, \mathbf{f}_j} = \frac{Cor(\mathbf{f}_i, \mathbf{Y}) + Cor(\mathbf{f}_j, \mathbf{Y})}{Cor(\mathbf{f}_j, \mathbf{f}_j)}, \tag{3}$$

where Cor is the Pearson's correlation measure. The above measure consists of three terms. The terms $Cor(\mathbf{f}_i, \mathbf{Y})$ and $Cor(\mathbf{f}_j, \mathbf{Y})$ are the relevance degrees of individual features $\mathbf{f}_i$ and $\mathbf{f}_j$ with respect to the target feature $\mathbf{Y}$, respectively. The term $Cor(\mathbf{f}_j, \mathbf{f}_j)$ measures the relevance between the feature pair $\{\mathbf{f}_i, \mathbf{f}_i\}$. Therefore, $V_{\mathbf{f}_i, \mathbf{f}_j}$ is large if and only if both $Cor(\mathbf{f}_i, \mathbf{Y})$ and $Cor(\mathbf{f}_j, \mathbf{Y})$ are large (i.e., both $\mathbf{f}_i$ and $\mathbf{f}_j$ are informative themselves with respect to the target feature representation $\mathbf{Y}$) and $Cor(\mathbf{f}_j, \mathbf{f}_j)$ is small (i.e., $\mathbf{f}_i$ and $\mathbf{f}_j$ are not correlated).

Furthermore, based on the proposed information theoretic measures, we construct three interacted matrices denoted as $\mathbf{W}$, $\mathbf{U}$, and $\mathbf{V}$ respectively. Specifically, each element $W_{i,j} \in \mathbf{W}$ represents the joint relevancy between a feature pair $\{\mathbf{f}_i, \mathbf{f}_j\}$ based on Eq.(1). Likewise, each element $U_{i,j} \in \mathbf{U}$ represents the redundancy between a feature pair $\{\mathbf{f}_i, \mathbf{f}_j\}$ based on Eq.(2). Moreover, each element $V_{i,j} \in \mathbf{V}$ represents the joint interaction between a feature pair $\{\mathbf{f}_i, \mathbf{f}_j\}$ based on Eq.(3). Given $\mathbf{W}$, $\mathbf{U}$, $\mathbf{V}$ and the N-dimensional feature indicator vector β, where β_i represents the coefficient for the i-th feature, we can identify the informative feature subset by solving the following optimization problem to ensure maximum joint relevancy, minimum redundancy, and maximum joint interaction of the selected features,

$$\begin{aligned} &\max f(\beta) = \max_{\beta \in \Re^N} \nu \beta^T \mathbf{W} \beta - \omega \beta^T \mathbf{U} \beta + \sigma \beta^T \mathbf{V} \beta, \\ &s.t. \quad \beta \in \Re^N, \beta \geq 0. \end{aligned} \tag{4}$$

where ν, ω and σ are the corresponding tuning parameters.

3.2 A Novel Sparse Learning Feature Selection Approach

Our discriminative feature selection approach is motivated by the purpose to ensure maximum joint relevancy, minimum redundancy, and maximum joint interaction of the selected features. In addition, it should simultaneously promote a sparse solution and a grouping effect of the highly correlated features. Therefore, we unify the minimization problem of Eq.(4) with the elastic net regression framework and propose the sparse learning feature selection method as

$$\min_{\beta \in \Re^N} \frac{1}{2} \|\mathbf{y}^T - \beta^T \mathbf{X}\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 - \lambda_3 \beta^T \mathbf{W} \beta + \lambda_4 \beta^T \mathbf{U} \beta - \lambda_5 \beta^T \mathbf{V} \beta, \tag{5}$$

where λ_1 and λ_2 are the tuning parameters for elastic net, λ_3, λ_4, and λ_5 are the tuning parameters for the relevancy matrix $\mathbf{W}$, the redundancy matrix $\mathbf{U}$, and the joint interaction matrix $\mathbf{V}$, respectively. The first term in Eq.(5) is the least square error term, the second term and the third term encourage sparsity and also promote a grouping effect of the selected feature as in the elastic net model. The fourth term guarantees maximum joint relevancy of selected features. The fifth term ensures minimum redundancy among selected features. Finally, the last term ensures that the selected features are jointly more interacted with the target class.

3.3 An Iterative Optimization Algorithm

To solve the optimization problem (5), we develop an iterative optimization algorithm based on ADMM, which uses a decomposition-coordination procedure. By using ADMM, the solutions to small local subproblems are coordinated to find a solution to a large global problem. This algorithm can be viewed as an attempt to blend the benefits of dual decomposition and augmented Lagrangian methods for constrained optimization.

Firstly, we reformulate the proposed feature selection problem into an equivalent constrained problem in the ADMM form,

$$\begin{aligned} &\min_{\beta\in\Re^N} \frac{1}{2}\|\mathbf{y}^T - \beta^T\mathbf{X}\|_2^2 + \lambda_2\|\beta\|_2^2 - \lambda_3\beta^T\mathbf{W}\beta + \lambda_4\beta^T\mathbf{U}\beta - \lambda_5\beta^T\mathbf{V}\beta + \lambda_1\|\gamma\|_1 \\ &s.t. \quad \beta = \gamma, \end{aligned} \tag{6}$$

where γ is an auxiliary variable, which can be regarded as a proxy for vector β. In this way, the objective function can be divided into two separate parts associated with two different variables, i.e., β and γ. This indicates that the hard constrained problem can be solved separately. As in the method of multipliers, we form the augmented Lagrangian function associated with the constrained problem (5) as follows

$$\begin{aligned} L_\rho(\beta,\gamma,z) &= \frac{1}{2}\|\mathbf{y}^T - \beta^T\mathbf{X}\|_2^2 + \lambda_2\|\beta\|_2^2 - \lambda_3\beta^T\mathbf{W}\beta + \lambda_4\beta^T\mathbf{U}\beta - \lambda_5\beta^T\mathbf{V}\beta \\ &+ \lambda_1\|\gamma\|_1 + < \beta-\gamma, z > + \frac{\rho}{2}\|\beta-\gamma\|_2^2, \end{aligned} \tag{7}$$

where $\langle\cdot,\cdot\rangle$ is an Euclidean inner product, z is a dual variable (i.e.,the Lagrange multiplier) associated with the equality constraint $\beta = \gamma$, and ρ is a positive penalty parameter (step size for dual variable update). By introducing an additional variable γ and an additional constraint $\beta - \gamma = 0$, we have simplified the optimization problem (5) by decoupling the objective function into two parts that depend on two different variables. In other words, we can decompose the minimization of $L_\rho(\beta,\gamma,z)$ into two simpler subproblems. Specifically, we solve the original problem (5) by seeking for a saddle point of the augmented Lagrangian by iteratively minimizing $L_\rho(\beta,\gamma,z)$ over β, γ, and z. Then the variables β, γ, and z can be updated in an alternating or sequential fashion, which accounts for the term alternating direction. This updating rule is shown as follows

(1) $\beta^{k+1} = \arg\min_{\beta\in\Re^p} L(\beta,\gamma^k,z^k)$, //$\beta$-minimization

(2) $\gamma^{k+1} = \arg\min_{\beta\in\Re^p} L(\beta^{k+1}, \gamma, z^k)$, //$\gamma$-minimization
(3) $z^{k+1} = z^k + \rho(\beta^{k+1} - \gamma^{k+1})$, //$z$-update

Given the above updating rule, we need to resolve each sub-problem iteratively until the termination criteria is satisfied. We perform the following calculation steps at each iteration.

(a)**Update** β

In the $(k+1)$-th iteration, in order to update β^k, we need to solve the following sub-problem, where the values of γ^k and z^k are fixed

$$\min_{\beta\in\Re^N} \frac{1}{2}\|\mathbf{y}^T - \beta^T\mathbf{X}\|_2^2 + \lambda_2\|\beta\|_2^2 - \lambda_3\beta^T\mathbf{W}\beta + \lambda_4\beta^T\mathbf{U}\beta - \lambda_5\beta^T\mathbf{V}\beta + \lambda_1\|\gamma\|_1 + < \beta - \gamma^k, z^k > + \frac{\rho}{2}\|\beta - \gamma^k\|_2^2. \quad (8)$$

Let the partial derivative with respect to β be equal to zero, we have

$$\frac{\partial[\min_{\beta\in\Re^N} \frac{1}{2}\|\mathbf{y}^T - \beta^T\mathbf{X}\|_2^2 + \lambda_2\|\beta\|_2^2 - \lambda_3\beta^T\mathbf{W}\beta + \lambda_4\beta^T\mathbf{U}\beta - \lambda_5\beta^T\mathbf{V}\beta]}{\partial\beta} + \frac{\partial[\min_{\beta\in\Re^N} \lambda_1\|\gamma\|_1 + < \beta - \gamma^k, z^k > + \frac{\rho}{2}\|\beta - \gamma^k\|_2^2]}{\partial\beta} = 0, \quad (9)$$

because

$$\begin{cases} \frac{\partial(\frac{1}{2}\|\mathbf{y}^T - \beta^T\mathbf{X}\|_2^2)}{\partial\beta} = -\mathbf{X}\mathbf{y} + \mathbf{X}\mathbf{X}^T\beta \\ \frac{\partial(\lambda_2\|\beta\|_2^2)}{\partial\beta} = \lambda_2\beta \\ \frac{\partial(-\lambda_3\beta^T\mathbf{W}\beta)}{\partial\beta} = -2\lambda_3\mathbf{W}\beta \\ \frac{\partial(\lambda_4\beta^T\mathbf{U}\beta)}{\partial\beta} = 2\lambda_4\mathbf{U}\beta \\ \frac{\partial(-\lambda_5\beta^T\mathbf{V}\beta)}{\partial\beta} = -2\lambda_5\mathbf{V}\beta \\ \frac{\partial<\beta-\gamma^k, z^k>}{\partial\beta} = z^k \\ \frac{\partial(\frac{\rho}{2}\|\beta-\gamma^k\|_2^2)}{\partial\beta} = \rho(\beta - \gamma^k), \end{cases} \quad (10)$$

we have

$$-\mathbf{X}\mathbf{y} + \mathbf{X}\mathbf{X}^T\beta + \lambda_2\beta - 2\lambda_3\mathbf{W}\beta + 2\lambda_4\mathbf{U}\beta - 2\lambda_5\mathbf{V}\beta + z^k + \rho(\beta - \gamma^k) = 0, \quad (11)$$

that is,

$$\beta^{k+1} = (\mathbf{X}\mathbf{X}^T + \lambda_2\mathbf{I} - 2\lambda_3\mathbf{W} + 2\lambda_4\mathbf{U} - 2\lambda_5\mathbf{V} + \rho\mathbf{I})^{-1}(\mathbf{X}\mathbf{y} - z^k + \rho\gamma^k). \quad (12)$$

(b)**Update** γ

Based on the results, assume β_i^{k+1} and z_i^k are fixed, for $i = 1, 2, ..., d$, we update γ_i^{k+1} by solving the following sub-optimization problem

$$\min_{\gamma_i} \lambda_1 \sum_{i=1}^{p} \|\gamma_i\|_1 - \sum_{i=1}^{p} < \gamma_i, z_i^k > + \frac{\rho}{2}\sum_{i=1}^{p} \|\beta_i^{k+1} - \gamma_i\|_2^2, \quad (13)$$

$$\frac{\partial[\min_{\gamma_i} \lambda_1 \sum_{i=1}^{p} \|\gamma_i\|_1 - \sum_{i=1}^{p} < \gamma_i, z_i^k > + \frac{\rho}{2} \sum_{i=1}^{p} \|\beta_i^{k+1} - \gamma_i\|_2^2]}{\partial \gamma_i} = 0. \quad (14)$$

We therefore have the following results

$$\gamma_i^{k+1} = \begin{cases} \frac{1}{\rho}(z_i^k + \rho\beta_i^{k+1} - \lambda_1), \text{ if } z_i^k + \rho\beta_i^{k+1} > \lambda_1 \\ \frac{1}{\rho}(z_i^k + \rho\beta_i^{k+1} + \lambda_1), \text{ if } z_i^k + \rho\beta_i^{k+1} < -\lambda_1 \\ 0, \qquad\qquad\qquad \text{if } z_i^k + \rho\beta_i^{k+1} \in [-\lambda_1, \lambda_1] \end{cases} \quad (15)$$

(c)**Update z**

Then, assume β_i^{k+1} and γ_i^{k+1} are fixed, for $i = 1, 2, ..., d$, we update z_i^{k+1} by using the following equation

$$z_i^{k+1} = z_i^k + \rho(\beta_i^{k+1} - \gamma_i^{k+1}). \quad (16)$$

Based on procedures (a), (b), and (c), we summarize the optimization algorithm below

Input: $\mathbf{X}, \mathbf{y}, \beta^0, z^0, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \rho$
Step1: While (not converged), **do**
Step2: Update β^{k+1} according to Eq.(12)
Step3: Update $\gamma_i^{k+1}, i = 1, 2, ..., d$ according to Eq.(15)
Step4: Update $\beta_i^{k+1}, i = 1, 2, ..., d$ according to Eq.(16)
End While
Output: β^*.

Algorithm 1: The iterative optimization algorithm for the proposed Tripple-EN method.

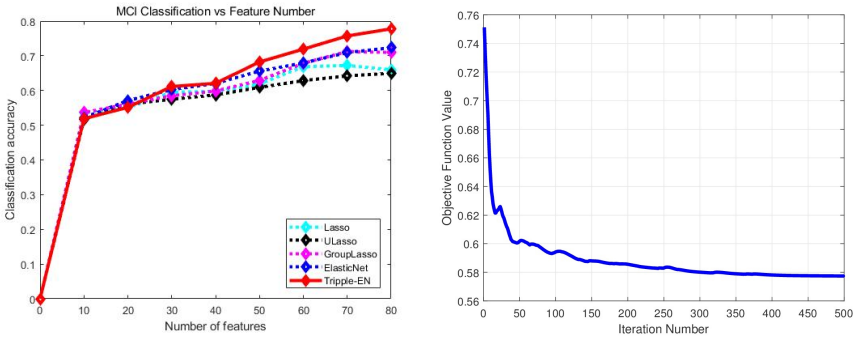
4 Experimental Analysis

We evaluate the performance of the proposed feature selection method for MCI classification on the public available Alzheimer's Disease Neuroimaging Initative (ANDI) database. Specifically, 54 MCI patients and 62 NC subjects were selected from ADNI database. The images of each subject were acquired using a 3.0T Philips scanners at centers in different places. The voxel size is $3.13 \times 3.13 \times 3.13mm^3$. SPM8 software package was applied to preprocess the RS-fMRI data. To evaluate the discriminative capabilities of the information captured by our method, we compare the classification results using the selected features from our method (Mu-InElasticNet) with several state-of-the-art feature selection methods, i.e., a) Lasso [8], b) ULasso [2], c) Group Lasso [5], and d) Elastic Net [12]. For the experiments, due to limited samples, a Leave One Out(LOO) cross-validation associated with C-SVM is applied to benchmark the generalization performance of different methods. Specifically, given N subjects, N-1 subjects are used as training data, and one subject is subsequently evaluated in terms of the classification accuracy. We repeat the procedure L times, and

report the averaged classification result. Fig.2(a) exhibits that the C-SVM associated with the proposed method can achieve the best classification accuracy, and the accuracy (y-axis) increases with the increasing number of selected features (x-axis). Moreover, Table.1 shows the best classification accuracy (ACC) for each method associated with the corresponding number of selected features, as well as other four associated indices, i.e., sensitivity (SEN), specificity (SPE), area under the receiver operating characteristic curve (AUC), and F-score. We observe that the proposed method significantly outperforms the remaining methods on all indices. The reason for the effectiveness is that only our method can simultaneously maximize relevancy and minimize redundancy of the selected features. Finally, we also experimentally evaluate the convergence property of the proposed method. Fig.2(b) indicates that the proposed method converges quickly within 50 iterations tend to be stable after 150 iterations.

Table 1. Performance of different methods in MCI classification (NC vs MCI).

Methods	Lasso	ULasso	GroupLasso	ElasticNet	Tripple-EN
ACC	0.6578	0.6842	0.7105	0.7192	**0.7894**
SEN	0.6663	0.6800	0.7143	0.7059	**0.8261**
SPE	0.6783	0.6875	0.7077	0.7143	**0.7647**
AUC	0.6723	0.6821	0.7110	0.7101	**0.7954**
F-score	0.6567	0.6538	0.6796	0.6857	**0.7600**
Feature Numbers	60 features	80 features	70 features	80 features	80 features



(a) Accuracies vs selected number of features (b) Convergence for the optimization

Fig. 2. Experiments for the proposed method.

5 Conclusion

In this paper, we have proposed a novel sparse learning feature selection method for MCI classification for AD diagnosis. Specifically, we devised three information theoretic measures to evaluate feature relevancy, feature redundancy and joint feature interaction. These measures are further encapsulated into the least square regression associated with an elastic net regularizer to simultaneously maximize relevancy, minimize redundancy, and maximize joint interaction of the selected features. Experiments demonstrated the effectiveness of our method on MCI classification tasks.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant no. 61602535 and 61976235), the Program for Innovation Research in Central University of Finance and Economics, and the Youth Talent Development Support Program by Central University of Finance and Economics, No. QYP1908.

References

1. Stephen P. Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
2. Sibao Chen, Chris H. Q. Ding, Bin Luo, and Ying Xie. Uncorrelated lasso. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA*, 2013.
3. Xiaobo Chen, Han Zhang, Lichi Zhang, Celina Shen, Seong-Whan Lee, and Dinggang Shen. Extraction of dynamic functional connectivity from brain grey matter and white matter for mci classification. *Human Brain Mapping*, 38:5019–5034, 2017.
4. Darya Chyzhyk, Alexandre Savio, and Manuel Graña. Evolutionary ELM wrapper feature selection for alzheimer's disease CAD on anatomical brain MRI. *Neurocomputing*, 128:73–80, 2014.
5. L Meier, S Van De Geer, and P Bhlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70(1):53–71, 2008.
6. Tofigh Naghibi, Sarah Hoffmann, and Beat Pfister. A semidefinite programming based search strategy for feature selection with mutual information measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(8):1529–1541, 2015.
7. Heung-Il Suk, Chong-Yaw Wee, Seong-Whan Lee, and Dinggang Shen. Supervised discriminative group sparse representation for mild cognitive impairment diagnosis. *Neuroinformatics*, 13(3):277–295, 2015.
8. Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
9. Chong-Yaw Wee, Sen Yang, Pew-Thian Yap, and Dinggang Shen. Sparse temporally dynamic resting-state functional connectivity networks for early mci identification. *Brain Imaging and Behavior*, 10(2):342–356, 2016.
10. Changqing Zhang, Huazhu Fu, Qinghua Hu, Pengfei Zhu, and Xiaochun Cao. Flexible multi-view dimensionality co-reduction. *IEEE Trans. Image Processing*, 26(2):648–659, 2017.
11. Yu Zhang, Han Zhang, Xiaobo Chen, Mingxia Liu, Xiaofeng Zhu, Seong-Whan Lee, and Dinggang Shen. Strength and similarity guided group-level brain functional network construction for MCI diagnosis. *Pattern Recognition*, 88:421–430, 2019.
12. H Zou and T Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67(5):301–320, 2005.