frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 09 November 2022
DOI 10.3389/fnins.2022.1036244

Multi-modal feature selection with anchor graph for Alzheimer's disease

Jiaye Li[1], Hang Xu[1], Hao Yu[1]*, Zhihao Jiang[2] and Lei Zhu[3] for the Alzheimer's Disease Neuroimaging Initiative

[1]School of Computer Science and Engineering, Central South University, Changsha, China, [2]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, [3]College of Information and Intelligence, Hunan Agricultural University, Changsha, China

OPEN ACCESS

EDITED BY
Yuan Yang,
University of Oklahoma, United States

REVIEWED BY
Yanzhi Bi,
Institute of Psychology (CAS), China
Na Zhang,
Shenzhen Institutes of Advanced Technology (CAS), China
Pedro Miguel Rodrigues,
Escola Superior de Biotecnologia-Universidade Católica Portuguesa, Portugal

*CORRESPONDENCE
Hao Yu
yhgxnu@gmail.com

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 04 September 2022
ACCEPTED 21 October 2022
PUBLISHED 09 November 2022

CITATION
Li J, Xu H, Yu H, Jiang Z and Zhu L (2022) Multi-modal feature selection with anchor graph for Alzheimer's disease. *Front. Neurosci.* 16:1036244. doi: 10.3389/fnins.2022.1036244

In Alzheimer's disease, the researchers found that if the patients were treated at the early stage of the disease, it could effectively delay the development of the disease. At present, multi-modal feature selection is widely used in the early diagnosis of Alzheimer's disease. However, existing multi-modal feature selection algorithms focus on learning the internal information of multiple modalities. They ignore the relationship between modalities, the importance of each modality and the local structure in the multi-modal data. In this paper, we propose a multi-modal feature selection algorithm with anchor graph for Alzheimer's disease. Specifically, we first use the least square loss and $l_{2,1}-$norm to obtain the weight of the feature under each modality. Then we embed a modal weight factor into the objective function to obtain the importance of each modality. Finally, we use anchor graph to quickly learn the local structure information in multi-modal data. In addition, we also verify the validity of the proposed algorithm on the published ADNI dataset.

KEYWORDS

Alzheimer's disease, anchor graph, local structure, multi-modal, feature selection

1. Introduction

Alzheimer's disease is a degenerative disease of the central nervous system in the elderly. It is one of the most common chronic diseases in human aging. Its clinical manifestations are memory impairment, aphasia, impairment of abstract thinking and computing power, personality, and behavior changes, etc.. At present, it can not be cured, only through comprehensive treatment to delay the development. However, the current study shows that if effective treatment is carried out in the early stage of Alzheimer's disease (i.e., mild cognitive impairment), further deterioration of the disease can be prevented. Therefore, how to accurately judge which stage the patient is in is very important.

On the one hand, in the information data about patients with Alzheimer's disease, due to personal privacy and other reasons, the data volume is relatively small, but the data dimension is relatively high. For example, in references Jie et al. (2013) and Liu et al. (2014), the amount of data is small, but their dimensions are indeed hundreds of thousands. On the other hand, the data on patients with Alzheimer's disease are often multi-modal, i.e., it includes three different modalities: magnetic resonance imaging

(MRI), positron emission tomography (PET), and cerebrospinal fluid (CSF) biomarkers. Different modalities have different characteristics and functions. Therefore, we not only need to reduce the dimension of the data, but also need to conduct multi-modal analysis of the data (i.e., the relationship between different modalities).

In addition, the current dimension reduction algorithms have high time complexity, i.e., the graph needs to calculate the similarity relationship between all sample points, which leads to a large amount of calculation. For example, Zhang et al. (2017) proposed a multi-modal feature selection algorithm. It uses the traditional graph laplace method to learn the local structure information in the data, and its calculation amount is large. Therefore, for the diagnosis of Alzheimer's disease, we need to carry out effective multi-modal feature selection with low time complexity.

In view of the above problems, this paper proposes a multi-modal feature selection algorithm for Alzheimer's disease diagnosis. Specifically, we first use the data under three modalities (i.e., MRI, PET, and CSF) to perform linear fitting with the labels, respectively, so as to obtain the weight relationship matrix for each modality. Then, we introduce anchor graph to quickly construct the relationship between samples, which can not only reduce the time complexity of the algorithm, but also learn the graph structure information in the data. Finally, we introduce $l_{2,1}$ sparse regularization term to obtain the weight of each feature and perform multi-modal feature selection. In addition, the proposed method also considers the relationship between modalities (i.e., the importance of each modality).

The main contributions of this paper are as follows:

- We propose a multi-modal feature selection framework for Alzheimer's disease, which can select important feature subsets to help the early diagnosis and prediction of Alzheimer's disease.
- The anchor graph is embedded in the proposed algorithm, which can reduce the time complexity of the algorithm.
- We apply a new alternative iterative optimization strategy to optimize proposed multi-modal feature selection algorithm. It can make the proposed objective function monotonically decrease until convergence in each iteration.
- For the proposed algorithm, we have carried out a series of experiments on ADNI datasets to verify the validity of the proposed method.

2. Related work

In this section, we will introduce some work on Alzheimer's disease from two aspects. That is, 1. Research on Alzheimer's disease with feature selection Algorithm 2. Research on Alzheimer's disease with multi-modal learning technology.

2.1. Feature selection for Alzheimer's disease

Alzheimer's disease is the most common dementia disease in the elderly. In 2016, a survey showed that more than 40 million people worldwide suffer from Alzheimer's disease, and this number is expected to double every 20 years. At the same time, the researchers found that if the patients were treated at the early stage of the disease, it could effectively delay the development of the disease. On the other hand, with the development of machine learning and deep learning, researchers use artificial intelligence algorithms to explore and understand the pathogenesis of Alzheimer's disease, thus providing a fast and effective way to explore the disease.

At present, most researchers use classification (Yu et al., 2022), regression, and clustering techniques to predict Alzheimer's disease data (Zhang et al., 2022b). But this ignores the problems caused by the high-dimensional features and redundant features in the data. Therefore, some researchers use feature selection algorithm to preprocess the data. For example, Mahendran and Vincent (2022) proposed an embedded feature selection method for early detection of Alzheimer's disease. Specifically, it first uses quality control, downstream analysis, and normalization to preprocess the data. Then it uses four feature selection algorithms to reduce the dimension of the data, so as to select the most suitable feature selection algorithm. Finally, it uses the deep learning model to classify and predict the reduced dimension data. Gallego-Jutglà et al. (2015) proposed a hybrid feature selection algorithm for early diagnosis of Alzheimer's disease. It classifies each feature by selecting the value range. Rani Kaka and Prasad (2021) used integrated feature selection and multiple support vector machines to predict Alzheimer's disease. Specifically, it first uses adaptive histogram equalization to improve the contrast. Then, it uses fuzzy c-means clustering algorithm to distinguish proteins, cerebrospinal fluid and gray matter. Finally, it uses the feature selection algorithm based on integration to reduce the dimension of the data, so as to classify them by the support vector machine. Chaves et al. (2012) proposed a feature selection algorithm based on association rules for Alzheimer's disease. On the one hand, this method uses principal components analysis (PCA) and partial least squares to reduce the dimension of data. On the other hand, it uses support vector machine to classify data. Thapa et al. (2020) proposed a data-driven technology based on feature selection for early diagnosis of Alzheimer's disease. They pointed out that the combination of neuropsychological scores and MRI features could be helpful for the early diagnosis of Alzheimer's disease. Liu et al. (2019) proposed a deep feature selection algorithm for Alzheimer's disease. This method combines deep learning, feature selection, causal reasoning, and genetic imaging analysis. Chyzhyk et al. (2014) proposed a wrapped feature selection to analyze MRI. This method uses extreme learning machines to train algorithms, thereby extracting original features from brain MRI. Niyas and Thiyagarajan (2022) used fisher scores and greedy search to select features in Alzheimer's disease data. Specifically, it first preprocesses the data. Then it uses fisher's score to rank all features and select the best feature subset. Finally, it uses greedy search to select the sub optimal minimum feature subset.

2.2. Multi-modal learning for Alzheimer's disease

The above feature selection algorithms are all for the single-mode Alzheimer's disease dataset. In addition to single-mode, there are multi-modal. Multi-modal learning means that there are more than one source and form of data, and the process of learning in these forms is called multi-modal learning. Multi-modal learning can be divided into five categories: multi-modal representation learning (Zhang C. et al., 2021), modal transformation, alignment (Zhu et al., 2022), multi-modal fusion, and collaborative learning (Li et al., 2019). In this paper, because we use multi-modal feature selection algorithm, we focus on multi-modal feature selection in multi-modal representation learning.

In the study of early diagnosis of Alzheimer's disease, datasets often include three different modalities: magnetic resonance imaging (MRI), positron emission tomography (PET), and cerebrospinal fluid (CSF). Therefore, it is necessary to select multi-modal features of datasets. For example, Zhang Y. et al. (2021) used neuroimaging embedding and feature selection to do early diagnosis of Alzheimer's disease. Specifically, it first uses the $l_{2,1}-$norm and multiple hinge losses to obtain the feature weights for each modality. Then it uses l_p-norm to fuse the complementary information of each modality. Finally, the convergence of the proposed method is proved theoretically. Shao et al. (2020) proposed a hypergraph based multi-task feature selection algorithm for Alzheimer's disease. Specifically, it first learns the feature subset for each modality separately. Then it selects the common feature subset of all modalities. Finally, it introduces the regularization term of hypergraph to establish the high-order structural relationship between samples. Jie et al. (2013) proposed a feature selection algorithm based on manifold learning for Alzheimer's disease. Specifically, it first performs single task learning in each modality. Then it uses a set of sparse regularization terms to learn the relationship between modalities. Finally, it introduces a laplace regularization term to maintain the geometric distribution in the data structure, so as to make more accurate feature selection. Bi et al. (2020) studied the multi-modal data of Alzheimer's disease by using evolutionary random forest algorithm. Specifically, it randomly selects samples and features to improve the generalization performance of random forest. In addition, it also uses hierarchical clustering to obtain the best decision tree. Jiao et al. (2022) proposed a multi-modal feature selection algorithm based on feature structure fusion for Alzheimer's disease. Specifically, it first calculates the similarity between features to construct the correlation regularization. Then, it uses manifold learning to obtain the local structure information of the data. Finally, it uses two regularization terms combined with low rank learning technology to obtain the feature subset of multi-modal data. Hao et al. (2020) proposed a multi-modal feature selection for Alzheimer's disease. Specifically, it first uses the random forest strategy to obtain the similarity of each mode. Then it uses a group sparse regularization terms and similarity regularization terms to constrain the objective function, so as to obtain the feature subsets for multiple modalities. Finally, it uses multi-kernel support vector machine to classify the reduced dimension data. Zhu et al. (2014) also proposed a multi-modal feature selection method for Alzheimer's disease. Specifically, it first uses canonical correlation analysis to consider the correlation between features. Then it uses the least square loss and $l_{2,1}-$norm to select features in multi-modal. Finally, according to the selected features, it performs multi-task learning.

From the above works, we can see that whether it is single-modal feature selection or multi-modal feature selection for Alzheimer's disease. Their core is to select the features that are most helpful for the early diagnosis of Alzheimer's disease, and then use these features to classify and predict the data.

3. Method

3.1. Notation

In this paper, we use capital bold letters, lowercase bold letters and ordinary letters to represent matrices, vectors, and scalars, respectively. Given a data matrix $\mathbf{X}$. $\mathbf{X}_v$ represents the data in the v-th modality. The l_f- norm of the matrix $\mathbf{X}$ is expressed as $\|\mathbf{X}\|_F = \left(\sum_j \|\mathbf{x}^j\|_2^2\right)^{1/2}$. The $l_{2,1}-$ norm of $\mathbf{X}$ is expressed as $\|\mathbf{X}\|_{2,1} = \sum_i \sqrt{\mathbf{x}_i^T\mathbf{x}_i + \varepsilon}$. In addition, we use $\mathbf{X}^T$, $\mathbf{X}^{-1}$ and $tr(\mathbf{X})$ to represent the transposition, inverse and trace of matrix $\mathbf{X}$, respectively.

3.2. Multi-modal learning

In practical applications, datasets often describe the same sample in many forms. For example, when we describe an animal, we can describe it in text, audio, or video. At this time, text, audio, and video can be considered as three modalities. Similarly, in the actual dataset, there are often some multi-modal datasets, and the traditional data mining algorithms can not be well-applied to these datasets. They can only mechanically learn a single modality. Multi-modal learning refers to using a function to model a specific view, and using redundant views of the same input to jointly optimize all functions, ultimately improving the learning effect. The traditional multi-modal feature selection function is as follows:

$$\min_{\mathbf{W}_v} \left\| \mathbf{W}_v^T \mathbf{X}_v^T - \mathbf{Y} \right\|_F^2 + R \left\| \mathbf{W}_v \right\| \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n \times d_v}$ represents the data in the v-th modality. $\mathbf{W}_v \in \mathbb{R}^{d_v \times k}$ represents the feature selection matrix in the v-th modality. $\mathbf{Y} \in \mathbb{R}^{k \times n}$ represents the label of all data. $R \left\| \mathbf{W}_v \right\|$ represents the regular term of $\mathbf{W}_v$, which can be the l_1-norm, $l_{2,1}-$norm, and $l_{2,p}-$norm that can realize feature selection.

3.3. Anchor graph construction

In graph learning, we often construct graphs according to the similarity calculation between samples (Zhang and Li, 2021). Specifically, we first regard each sample as a node of the graph, and then use the metric function to calculate the similarity or relationship between the samples (Zhang et al., 2022a). Finally, we use the obtained relations or weights in the previous step to construct the edges between nodes (samples), so as to construct the graph structure in the data, so as to learn the local structure or global structure information in the data. The traditional graph learning method is as follows:

$$\min_{\mathbf{s}_i^T 1 = 1, 0 \le s_{ij} \le 1} \sum\nolimits_{i,j} \left(\left\| \mathbf{x}_i - \mathbf{x}_j \right\|_2^2 s_{ij} + \eta s_{ij}^2 \right) \tag{2}$$

In Equation (2), it uses the square of Euclidean distance to calculate the distance between the i-th sample $\mathbf{x}_i$ and the j-th sample $\mathbf{x}_j$. $\mathbf{S}$ is a similarity matrix. After $\mathbf{S}$ is obtained by solving Equation (2), We can obtain the laplace matrix by $\mathbf{L} = \mathbf{D} - \mathbf{S}$, so as to learn the local structure information in the data.

Although the above method can learn the graph structure in the data, its time complexity is relatively high, because it needs to calculate the similarity between each sample and all other samples. Therefore, some researchers have proposed anchor point graph construction. Specifically, it first generates anchor points from all the data, and then establishes the similarity matrix between the anchor points and the sample points. If the anchor point is selected by random sampling method, its time complexity is $O(1)$. Suppose there are m anchor points generated, and the total number of samples is n, and each sample has d features. The time complexity of generating the similarity matrix is $O[nd \log(m)]$. The formula for constructing anchor point graph is as follows:

$$\min_{\mathbf{z}_i^T 1 = 1, z_i \ge 0} \sum_{j=1}^{m} \left\| \mathbf{x}_i - \mathbf{a}_j \right\|_2^2 z_{ij} + \eta \sum_{j=1}^{m} z_{ij}^2 \tag{3}$$

where $\mathbf{a}_j$ is the generated j-th anchor point and $\mathbf{Z} \in \mathbb{R}^{n \times m}$ is the similarity matrix. η is an adjustable parameter. According to references Nie et al. (2014) and Nie et al. (2021), we can get the solution of $\mathbf{Z}$ as:

$$z_{ij} = \frac{d_{i,k+1} - d_{ij}}{k d_{i,k+1} - \sum_{j=1}^{k} d_{ij}} \tag{4}$$

where $d_{ij} = \left\| \mathbf{x}_i - \mathbf{a}_j \right\|_2^2$ and k is a non-negative parameter. After obtaining the matrix $\mathbf{Z}$, we can obtain the similarity matrix through the following formula:

$$\mathbf{S} = \mathbf{Z} \Delta^{-1} \mathbf{Z}^T \tag{5}$$

where Δ is a diagonal matrix whose diagonal elements are $\Delta_{jj} = \sum_{i=1}^{n} z_{ij}$. The function of anchor point graph is to reduce the time complexity. After the similarity matrix $\mathbf{S}$ is obtained by constructing anchor points, it is still necessary to obtain the laplace matrix with $\mathbf{L} = \mathbf{D} - \mathbf{S}$.

3.4. Proposed multi-modal feature selection with anchor graph

Equation (1) enables feature selection unless select the appropriate regular term $R \left\| \mathbf{W}_v \right\|$. Due to the wide applicability of the $l_{2,1}-$norm, in this paper, we choose the $l_{2,1}-$norm to limit the weight matrix $\mathbf{W}_v$. i.e., the following formula can be further obtained:

$$\min_{\mathbf{W}_v} \left\| \mathbf{W}_v^T \mathbf{X}_v^T - \mathbf{Y} \right\|_F^2 + \alpha \left\| \mathbf{W}_v \right\|_{2,1} \tag{6}$$

Although Equation (6) can be used for multi-modal feature selection, it ignores the deviation problem in the process of data fitting. In addition, it does not learn the graph structure information existing in the data. Therefore, we further introduce the deviation term and the graph regularization term, as shown in the following formula:

$$\begin{aligned} \min_{\mathbf{W}_v, \mathbf{b}, \boldsymbol{\theta}_v} & \left\| \mathbf{W}_v^T \mathbf{X}_v^T + \mathbf{b} \mathbf{1}_n^T - \mathbf{Y} \right\|_F^2 + \alpha \left\| \mathbf{W}_v \right\|_{2,1} \\ & + \beta tr(\mathbf{W}_v^T \mathbf{X}_v^T \mathbf{L}_v \mathbf{X}_v \mathbf{W}_v) \end{aligned} \tag{7}$$

where $\mathbf{L}_v$ is the laplace matrix, $\mathbf{b}$ is the deviation term, and α and β are adjustable hyperparameters. Since what we are doing is multi-modal feature selection, Equation (7) cannot consider the weight of each modality. Different modalities should have different importance. Therefore, we need learn the weight of each modality in Equation (7) and further obtain the final objective function, as shown below:

$$\begin{aligned} \min_{\mathbf{W}_v, \mathbf{b}, \boldsymbol{\theta}_v} & \left\| \mathbf{W}_v^T \boldsymbol{\Theta}_v \mathbf{X}_v^T + \mathbf{b} \mathbf{1}_n^T - \mathbf{Y} \right\|_F^2 + \alpha \left\| \mathbf{W}_v \right\|_{2,1} \\ & + \beta tr(\mathbf{W}_v^T \mathbf{X}_v^T \mathbf{L}_v \mathbf{X}_v \mathbf{W}_v) \\ s.t. & [\boldsymbol{\theta}_1; \boldsymbol{\theta}_2; \cdots; \boldsymbol{\theta}_v]^T 1_d = 1, \theta \ge 0 \end{aligned} \tag{8}$$

To sum up, Equation (8) improves multi-modal feature selection from three aspects: 1. The weight of each modality, i.e., $\boldsymbol{\Theta}_v$, is considered. 2. The deviation problem in the process of data fitting is considered. 3. Using anchor graph construction to improve the slow learning speed of traditional graph structure.

3.5. Optimization

In this section, we optimize the proposed objective function, i.e., Equation (8), by alternating iterations.

Update b by Fixing W_v and $\boldsymbol{\theta}_v$.

When $\mathbf{W}_v$ and $\boldsymbol{\theta}_v$ are fixed, we can obtain the following formula:

$$\min_{\mathbf{b}} \left\| \mathbf{W}_v^T \boldsymbol{\Theta}_v \mathbf{X}_v^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y} \right\|_F^2 \tag{9}$$

Next, we take the derivative of $\mathbf{b}$ with Equation (9) and make the derivative zero, as follows:

$$\frac{\partial \left\| \mathbf{W}_v^T \boldsymbol{\Theta}_v \mathbf{X}_v^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y} \right\|_F^2}{\partial \mathbf{b}} = 0 \tag{10}$$

Further, Equation (10) is equivalent to the following equation:

$$2\mathbf{W}_v^T \boldsymbol{\Theta}_v \mathbf{X}_v^T \mathbf{1}_n + 2\mathbf{b}\mathbf{1}_n^T \mathbf{1}_n - 2\mathbf{Y}\mathbf{1}_n = 0 \tag{11}$$

Through Equation (11), we can obtain the solution of $\mathbf{b}$ as follows:

$$\mathbf{b} = \tfrac{1}{n}(\mathbf{Y}\mathbf{1}_n - \mathbf{W}_v^T \boldsymbol{\Theta}_v \mathbf{X}_v^T \mathbf{1}_n) \tag{12}$$

Update W_v by Fixing b and $\boldsymbol{\theta}_v$.

When $\mathbf{b}$ and $\boldsymbol{\theta}_v$ are fixed, Equation (8) can be converted into the following equation:

$$\begin{aligned} \min_{\mathbf{W}_v} & \left\| \mathbf{W}_v^T \boldsymbol{\Theta}_v \mathbf{X}_v^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y} \right\|_F^2 + \alpha \|\mathbf{W}_v\|_{2,1} \\ & + \beta tr(\mathbf{W}_v^T \mathbf{X}_v^T \mathbf{L}_v \mathbf{X}_v \mathbf{W}_v) \end{aligned} \tag{13}$$

On the one hand, since $\|\mathbf{W}_v\|_{2,1} = \sum_{i=1}^{d_v} \|\mathbf{w}_{vi}\|_2$ and $\|\mathbf{w}_{vi}\|_2$ are likely to be 0, this will cause Equation (13) to be non differentiable. Therefore, we introduce a sufficiently small constant ε to solve this problem, i.e., replace $\|\mathbf{w}_{vi}\|_2$ with $\sqrt{\mathbf{w}_{vi}^T \mathbf{w}_{vi} + \varepsilon}$. On the other hand, $\mathbf{L}_v = \mathbf{D}_v - \mathbf{S}_v$, where $\mathbf{L}_v$ is a laplace matrix, $\mathbf{D}_v$ is a degree matrix, and $\mathbf{S}_v$ is a similarity matrix. Since we use anchor point graph construction, $\mathbf{S}_v = (\mathbf{B}\mathbf{B}^T)_v$, where $\mathbf{B} = \mathbf{Z}\Delta^{-\frac{1}{2}}$. For the degree matrix $\mathbf{D}_v$, its diagonal element value is:

$$D_{ii} = \textstyle\sum_{sj} Z_{is}(\Delta_{ss})^{-1} Z_{js} = \sum_s Z_{is} = 1 \tag{14}$$

Therefore, we can get the degree matrix $\mathbf{D}_v = \mathbf{I}_v$. Further, Equation (13) may be written as the following equation:

$$\begin{aligned} \min_{\mathbf{W}_v} & \left\| \mathbf{W}_v^T \boldsymbol{\Theta}_v \mathbf{X}_v^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y} \right\|_F^2 \\ & + \alpha \sum_{i=1}^{d_v} \sqrt{\mathbf{w}_{vi}^T \mathbf{w}_{vi} + \varepsilon} + \beta tr(\mathbf{W}_v^T \mathbf{X}_v^T (\mathbf{I}_v - (\mathbf{B}\mathbf{B}^T)_v)\mathbf{X}_v \mathbf{W}_v) \end{aligned} \tag{15}$$

Further, we use Equation (15) to find the derivative of $\mathbf{W}_v$ and let the derivative be 0 to obtain the following formula:

$$\frac{\partial \begin{pmatrix} \left\| \mathbf{W}_v^T \boldsymbol{\Theta}_v \mathbf{X}_v^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y} \right\|_F^2 + \alpha \sum_{i=1}^{d_v} \sqrt{\mathbf{w}_{vi}^T \mathbf{w}_{vi} + \varepsilon} \\ + \beta tr(\mathbf{W}_v^T \mathbf{X}_v^T (\mathbf{I}_v - (\mathbf{B}\mathbf{B}^T)_v)\mathbf{X}_v \mathbf{W}_v) \end{pmatrix}}{\partial \mathbf{W}_v} = 0 \tag{16}$$

Equation (16) is equivalent to the following equation:

$$\begin{aligned} & 2\boldsymbol{\Theta}_v \mathbf{X}_v^T \mathbf{X}_v \boldsymbol{\Theta}_v^T \mathbf{W}_v + 2\boldsymbol{\Theta}_v \mathbf{X}_v^T \mathbf{1}_n \mathbf{b}^T - 2\boldsymbol{\Theta}_v \mathbf{X}_v^T \mathbf{Y}^T \\ & + 2\alpha \mathbf{N}_v \mathbf{W}_v + 2\beta \mathbf{X}_v^T (\mathbf{I}_v - (\mathbf{B}\mathbf{B}^T)_v)\mathbf{X}_v \mathbf{W}_v = 0 \end{aligned} \tag{17}$$

where the value of each element in $\mathbf{N}_v$ is:

$$N_{vii} = \frac{1}{2\sqrt{\mathbf{w}_{vi}^T \mathbf{w}_{vi} + \varepsilon}} \tag{18}$$

According to Equation (17), we can obtain the closed form solution of $\mathbf{W}_v$ as:

$$\begin{aligned} \mathbf{W}_v = & \begin{pmatrix} \boldsymbol{\Theta}_v \mathbf{X}_v^T \mathbf{X}_v \boldsymbol{\Theta}_v^T + \alpha \mathbf{N}_v \\ + \beta \mathbf{X}_v^T (\mathbf{I}_v - (\mathbf{B}\mathbf{B}^T)_v)\mathbf{X}_v \end{pmatrix}^{-1} \\ & (\boldsymbol{\Theta}_v \mathbf{X}_v^T \mathbf{Y}^T - \boldsymbol{\Theta}_v \mathbf{X}_v^T \mathbf{1}_n \mathbf{b}^T) \end{aligned} \tag{19}$$

Update $\boldsymbol{\theta}_v$ by Fixing b and W_v.

When $\mathbf{b}$ and $\mathbf{W}_v$ are fixed, we solve $\boldsymbol{\theta}_v$. Since $\boldsymbol{\theta}_v$ is the weight for each modality, we can solve the weight under all modalities at once, i.e., $\boldsymbol{\theta} = [\boldsymbol{\theta}_1; \boldsymbol{\theta}_2; \cdots; \boldsymbol{\theta}_v]$. When $\mathbf{W}_v$ is solved, i.e., after $\mathbf{W} = [\mathbf{W}_1; \mathbf{W}_2; \cdots; \mathbf{W}_v]$ is obtained. At this time, Equation (8) can be written as follows:

$$\begin{aligned} & \min_{\boldsymbol{\theta}} \left\| \mathbf{W}^T \boldsymbol{\Theta} \mathbf{X}^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y} \right\|_F^2 \\ & s.t. \boldsymbol{\theta}^T \mathbf{1}_d = 1, \theta \geq 0 \end{aligned} \tag{20}$$

We bring Equation (12) into Equation (20) to further obtain the following equation:

$$\begin{aligned} & \min_{\boldsymbol{\theta}} \left\| \mathbf{W}^T \boldsymbol{\Theta} \mathbf{X}^T \mathbf{H} - \mathbf{Y}\mathbf{H} \right\|_F^2 \\ & s.t. \boldsymbol{\theta}^T \mathbf{1}_d = 1, \theta \geq 0 \end{aligned} \tag{21}$$

where $\mathbf{H} = \mathbf{I}_n - (1/n)\mathbf{1}_n \mathbf{1}_n^T$. Through simple transformation, we write Equation (21) in the form of trace, and the following formula can be obtained:

$$\begin{aligned} & \min_{\boldsymbol{\theta}} \left(tr(\boldsymbol{\Theta}\mathbf{X}^T \mathbf{H}\mathbf{H}^T \mathbf{X} \boldsymbol{\Theta}^T \mathbf{W}\mathbf{W}^T) - tr(2\boldsymbol{\Theta}\mathbf{X}^T \mathbf{H}\mathbf{H}^T \mathbf{Y}^T \mathbf{W}^T) \right) \\ & s.t. \boldsymbol{\theta}^T \mathbf{1}_d = 1, \theta \geq 0 \end{aligned} \tag{22}$$

Because $\mathbf{H}\mathbf{H}^T = \mathbf{H}$ and $\mathbf{\Theta}^T = \mathbf{\Theta}$. Therefore, Equation (22) may be further written as follows:

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \left(tr(\mathbf{\Theta}\mathbf{X}^T\mathbf{H}\mathbf{X}\mathbf{\Theta}\mathbf{W}\mathbf{W}^T) - tr(2\mathbf{\Theta}\mathbf{X}^T\mathbf{H}\mathbf{Y}^T\mathbf{W}^T) \right) \\ &s.t.\boldsymbol{\theta}^T\mathbf{1}_d = 1, \theta \geq 0 \end{aligned} \tag{23}$$

Next, we introduce the following lemma to solve Equation (23), and lemma 3.5 is as follows:

Lemma 1. *If a is diagonal, then* $tr(\mathbf{ABAC}) = \mathbf{a}^T(\mathbf{B}^T \circ \mathbf{C})\mathbf{a}$.

Proof.

$$\begin{aligned} tr(\mathbf{ABAC}) &= \mathbf{a}^T diag(\mathbf{BAC}) \\ &= \mathbf{a}^T vec\{\mathbf{b}_i^T\mathbf{A}\mathbf{c}_i\} \\ &= \mathbf{a}^T vec\{(\mathbf{b}_i \circ \mathbf{c}_i)^T\mathbf{a}\} \\ &= \mathbf{a}^T(\mathbf{B}^T \circ \mathbf{C})^T\mathbf{a} = \mathbf{a}^T(\mathbf{B}^T \circ \mathbf{C})\mathbf{a} \end{aligned} \tag{24}$$

By lemma 3.5, Equation (23) can be written as follows:

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \boldsymbol{\theta}^T \left((\mathbf{X}^T\mathbf{H}\mathbf{X})^T \circ (\mathbf{W}\mathbf{W}^T) \right) \boldsymbol{\theta} - \boldsymbol{\theta}^T diag(2\mathbf{X}^T\mathbf{H}\mathbf{Y}^T\mathbf{W}^T) \\ &s.t.\boldsymbol{\theta}^T\mathbf{1}_d = 1, \theta \geq 0 \end{aligned} \tag{25}$$

Further, Equation (25) may be written as the following equation:

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \boldsymbol{\theta}^T\mathbf{Q}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{s} \\ &s.t.\boldsymbol{\theta}^T 1_d = 1, \theta \geq 0 \end{aligned} \tag{26}$$

where

$$\begin{cases} \mathbf{Q} = (\mathbf{X}^T\mathbf{H}^T\mathbf{X}) \circ (\mathbf{W}\mathbf{W}^T) \\ \mathbf{s} = diag(2\mathbf{X}^T\mathbf{H}\mathbf{Y}^T\mathbf{W}^T) \end{cases} \tag{27}$$

Next, we use the augmented lagrange multiplier method to solve Equation (26). We first introduce the variable $\mathbf{u}$ to rewrite Equation (26), as follows:

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \boldsymbol{\theta}^T\mathbf{Q}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{s} \\ &s.t.\theta^T\mathbf{1}_d = 1, \theta \geq 0, \mathbf{u} = \boldsymbol{\theta} \end{aligned} \tag{28}$$

Further, we construct the augmented lagrangian function as follows:

$$\begin{aligned} f(\boldsymbol{\theta}, \mathbf{u}, \mu, \boldsymbol{\lambda}_1, \lambda_2) = {} & \boldsymbol{\theta}^T\mathbf{Q}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{s} + \tfrac{\mu}{2}\left\| \boldsymbol{\theta} - \mathbf{u} + \tfrac{1}{\mu}\boldsymbol{\lambda}_1 \right\|_F^2 \\ & + \tfrac{\mu}{2}(\boldsymbol{\theta}^T\mathbf{1}_d - 1 + \tfrac{1}{\mu}\lambda_2)^2 \\ s.t.u \geq 0 & \end{aligned} \tag{29}$$

where μ is a lagrange multiplier. Since the variable $\mathbf{u}$ is introduced, we still solve Equation (29) by alternating iterative optimization. i.e., when the variable $\mathbf{u}$ is fixed, Equation (29) is equivalent to the following equation:

$$\min_{\boldsymbol{\theta}} \tfrac{1}{2}\boldsymbol{\theta}^T\mathbf{E}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{g} \tag{30}$$

where

$$\begin{cases} \mathbf{E} = 2\mathbf{Q} + \mu\mathbf{I}_d + \mu\mathbf{1}_d\mathbf{1}_d^T \\ \mathbf{g} = \mu\mathbf{u} + \mu\mathbf{1}_d - \lambda_2\mathbf{1}_d - \boldsymbol{\lambda}_1 + \mathbf{s} \end{cases} \tag{31}$$

Obviously, we can get the solution of $\boldsymbol{\theta}$ as follows:

$$\boldsymbol{\theta} = \mathbf{E}^{-1}\mathbf{g} \tag{32}$$

When $\boldsymbol{\theta}$ is fixed, Equation (29) can be written as follows:

$$\min_{u \geq 0} \left\| \mathbf{u} - (\boldsymbol{\theta} + \tfrac{1}{\mu}\boldsymbol{\lambda}_1) \right\|^2 \tag{33}$$

According to Equation (33), we can obtain the solution of $\mathbf{u}$ as follows:

$$\mathbf{u} = pos(\boldsymbol{\theta} + \tfrac{1}{\mu}\boldsymbol{\lambda}_1) \tag{34}$$

The function of $pos(x)$ is to assign the negative element of x to 0. For the reader's understanding, we summarize the pseudo code of the algorithm as shown in Algorithm 1.

```
Input: Training set $[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_v] \in \mathbb{R}^{n\times(d_1+d_2+,\ldots,d_v)}$,
       adjustable parameters $\alpha$ and $\beta$;
Output: $\mathbf{W}^{(t)} \in \mathbb{R}^{d\times k}$;
Initialize $t$ = 0;
Randomly initialize $\mathbf{W}_v^{(0)}$;
repeat
    Updata $\mathbf{b}^{(t+1)}$ via Equation (12);
    Compute $\mathbf{N}_v$ via $N_{vii} = \frac{1}{2\sqrt{\mathbf{w}_{vi}^T\mathbf{w}_{vi}+\varepsilon}}$ ;
    Compute $\mathbf{W}^{(t+1)}$ via Equation (19) ;
    Updata $\boldsymbol{\theta}^{(t+1)}$ via Equation (32);
    $t$ = $t$+1 ;
until converge;
After getting the $\mathbf{W}_v$ on each mode, we put them
together, i.e., $\mathbf{W}^{(t)} = [\mathbf{W}_1; \mathbf{W}_2; \ldots; \mathbf{W}_v]$;
```

Algorithm 1. Pseudo code for proposed method.

3.6. Convergence analysis

In this section, we prove the convergence of the algorithm. We first introduce the following lemma:

Lemma 2. *For any non-zero vector* $\mathbf{x}$ *and* $\mathbf{y}$*, the following formula holds:*

$$\|\mathbf{x}\|_2 - \frac{\|\mathbf{x}\|_2^2}{2\|\mathbf{y}\|_2} \leq \|\mathbf{y}\|_2 - \frac{\|\mathbf{y}\|_2^2}{2\|\mathbf{y}\|_2} \tag{35}$$

Theorem 1. *The value of the proposed objective function monotonically decreases in each iteration until it converges.*

Proof. When $\mathbf{b}$ and $\boldsymbol{\theta}_v$ are fixed, we use $\mathbf{W}_v^{(t)}$ and $\mathbf{W}_v^{(t+1)}$ to represent the values of $\mathbf{W}_v$ at the t-th and $(t+1)$-th iterations, respectively. According to Equation (19), we can obtain the following formula:

$$\begin{aligned} \mathbf{W}_v^{(t+1)} = \arg\min_{\mathbf{W}_v} tr\Big(&(\mathbf{X}_v\boldsymbol{\Theta}_v^T\mathbf{W}_v^{(t)} + \mathbf{1}_n\mathbf{b}^T - \mathbf{Y}^T) \\ &(\mathbf{W}_v^{(t)^T}\boldsymbol{\Theta}_v\mathbf{X}_v^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y})\Big) \\ &+\alpha tr(\mathbf{W}_v^{(t)^T}\mathbf{N}_v^{(t)}\mathbf{W}_v^{(t)}) + \beta tr(\mathbf{W}_v^{(t)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t)}) \end{aligned} \tag{36}$$

Further, we can obtain the following formula:

$$\begin{aligned} & tr\Big((\mathbf{X}_v\boldsymbol{\Theta}_v^T\mathbf{W}_v^{(t+1)} + \mathbf{1}_n\mathbf{b}^T - \mathbf{Y}^T)(\mathbf{W}_v^{(t+1)^T}\boldsymbol{\Theta}_v\mathbf{X}_v^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y})\Big) \\ &+\alpha tr(\mathbf{W}_v^{(t+1)^T}\mathbf{N}_v^{(t)}\mathbf{W}_v^{(t+1)}) + \beta tr(\mathbf{W}_v^{(t+1)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t+1)}) \\ &\leq tr\Big((\mathbf{X}_v\boldsymbol{\Theta}_v^T\mathbf{W}_v^{(t)} + \mathbf{1}_n\mathbf{b}^T - \mathbf{Y}^T)(\mathbf{W}_v^{(t)^T}\boldsymbol{\Theta}_v\mathbf{X}_v^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y})\Big) \\ &+\alpha tr(\mathbf{W}_v^{(t+1)^T}\mathbf{N}_v^{(t)}\mathbf{W}_v^{(t)}) + \beta tr(\mathbf{W}_v^{(t)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t)}) \end{aligned} \tag{37}$$

Equation (37) may be rewritten as follows:

$$\begin{aligned} & tr\Big((\mathbf{X}_v\boldsymbol{\Theta}_v^T\mathbf{W}_v^{(t+1)} + \mathbf{1}_n\mathbf{b}^T - \mathbf{Y}^T)(\mathbf{W}_v^{(t+1)^T}\boldsymbol{\Theta}_v\mathbf{X}_v^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y})\Big) \\ &+\alpha\sum_{i=1}^{d_v}(\left\|(\mathbf{W}_v^{(t+1)})_i\right\|_2 + \frac{\left\|(\mathbf{W}_v^{(t+1)})_i\right\|_2^2}{2\left\|(\mathbf{W}_v^{(t)})_i\right\|_2} - \left\|(\mathbf{W}_v^{(t+1)})_i\right\|_2) + \beta tr(\mathbf{W}_v^{(t+1)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t+1)}) \\ &\leq tr\Big((\mathbf{X}_v\boldsymbol{\Theta}_v^T\mathbf{W}_v^{(t)} + \mathbf{1}_n\mathbf{b}^T - \mathbf{Y}^T)(\mathbf{W}_v^{(t)^T}\boldsymbol{\Theta}_v\mathbf{X}_v^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y})\Big) \\ &+\alpha\sum_{i=1}^{d_v}(\left\|(\mathbf{W}_v^{(t)})_i\right\|_2 + \frac{\left\|(\mathbf{W}_v^{(t)})_i\right\|_2^2}{2\left\|(\mathbf{W}_v^{(t)})_i\right\|_2} - \left\|(\mathbf{W}_v^{(t)})_i\right\|_2) + \beta tr(\mathbf{W}_v^{(t)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t)}) \end{aligned} \tag{38}$$

According to lemma 2, we can get:

$$\begin{aligned} & tr\Big((\mathbf{X}_v\boldsymbol{\Theta}_v^T\mathbf{W}_v^{(t+1)} + \mathbf{1}_n\mathbf{b}^T - \mathbf{Y}^T)(\mathbf{W}_v^{(t+1)^T}\boldsymbol{\Theta}_v\mathbf{X}_v^T + \mathbf{b}\mathbf{1}_n^T - \mathbf{Y})\Big) \\ &+\alpha\left\|\mathbf{W}_v^{(t+1)}\right\|_{2,1} + \beta tr(\mathbf{W}_v^{(t+1)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t+1)}) \\ &\leq tr\Big((\mathbf{X}_v\boldsymbol{\Theta}_v^T\mathbf{W}_v^{(t)} + \mathbf{1}_n\mathbf{b}^T - \mathbf{Y}^T)(\mathbf{W}_v^{(t)^T}\boldsymbol{\Theta}_v\mathbf{X}_v^T + \mathbf{b}\mathbf{1}_n^T - Y)\Big) \\ &+\alpha\left\|\mathbf{W}_v^{(t)}\right\|_{2,1} + \beta tr(\mathbf{W}_v^{(t)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t)}) \end{aligned} \tag{39}$$

In the $(t+1)$-iteration, when $\mathbf{W}_v^{(t)}$ and $\boldsymbol{\theta}_v^{(t)}$ are fixed, we can obtain the closed form solution of $\mathbf{b}^{(t+1)}$ according to Equation (12). Therefore, it is easy to obtain the following formula:

$$\begin{aligned} & tr\Big((\mathbf{X}_v\boldsymbol{\Theta}_v^{(t)^T}\mathbf{W}_v^{(t)} + \mathbf{1}_n\mathbf{b}^{(t+1)^T} - \mathbf{Y}^T)(\mathbf{W}_v^{(t)^T}\boldsymbol{\Theta}_v^{(t)}\mathbf{X}_v^T + \mathbf{b}^{(t+1)}\mathbf{1}_n^T - \mathbf{Y})\Big) \\ &+\alpha\left\|\mathbf{W}_v^{(t)}\right\|_{2,1} + \beta tr(\mathbf{W}_v^{(t)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t)}) \\ &\leq tr\Big((\mathbf{X}_v\boldsymbol{\Theta}_v^{(t)^T}\mathbf{W}_v^{(t)} + \mathbf{1}_n\mathbf{b}^{(t)^T} - \mathbf{Y}^T)(\mathbf{W}_v^{(t)^T}\boldsymbol{\Theta}_v^{(t)}\mathbf{X}_v^T + \mathbf{b}^{(t)}\mathbf{1}_n^T - \mathbf{Y})\Big) \\ &+\alpha\left\|\mathbf{W}_v^{(t)}\right\|_{2,1} + \beta tr(\mathbf{W}_v^{(t)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t)}) \end{aligned} \tag{40}$$

When $\mathbf{W}_v^{(t+1)}$ and $\mathbf{b}^{(t+1)}$ are fixed, we can get the following according to Equation (32):

$$\begin{aligned} & tr\Big((\mathbf{X}_v\boldsymbol{\Theta}_v^{(t+1)^T}\mathbf{W}_v^{(t+1)} + \mathbf{1}_n\mathbf{b}^{(t+1)^T} - \mathbf{Y}^T)(\mathbf{W}_v^{(t+1)^T}\boldsymbol{\Theta}_v^{(t+1)}\mathbf{X}_v^T + \mathbf{b}^{(t+1)}\mathbf{1}_n^T - \mathbf{Y})\Big) \\ &+\alpha\left\|\mathbf{W}_v^{(t+1)}\right\|_{2,1} + \beta tr(\mathbf{W}_v^{(t+1)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t+1)}) \\ &\leq tr\Big((\mathbf{X}_v\boldsymbol{\Theta}_v^{(t)^T}\mathbf{W}_v^{(t+1)} + \mathbf{1}_n\mathbf{b}^{(t+1)^T} - \mathbf{Y}^T)(\mathbf{W}_v^{(t+1)^T}\boldsymbol{\Theta}_v^{(t)}\mathbf{X}_v^T + \mathbf{b}^{(t+1)}\mathbf{1}_n^T - \mathbf{Y})\Big) \\ &+\alpha\left\|\mathbf{W}_v^{(t+1)}\right\|_{2,1} + \beta tr(\mathbf{W}_v^{(t+1)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t+1)}) \end{aligned} \tag{41}$$

According to Equations (39)–(41), we can finally obtain:

$$\begin{aligned} & tr\Big((\mathbf{X}_v\boldsymbol{\Theta}_v^{(t+1)^T}\mathbf{W}_v^{(t+1)} + \mathbf{1}_n\mathbf{b}^{(t+1)^T} - \mathbf{Y}^T)(\mathbf{W}_v^{(t+1)^T}\boldsymbol{\Theta}_v^{(t+1)}\mathbf{X}_v^T + \mathbf{b}^{(t+1)}\mathbf{1}_n^T - \mathbf{Y})\Big) \\ &+\alpha\left\|\mathbf{W}_v^{(t+1)}\right\|_{2,1} + \beta tr(\mathbf{W}_v^{(t+1)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t+1)}) \\ &\leq tr\Big((\mathbf{X}_v\boldsymbol{\Theta}_v^{(t)^T}\mathbf{W}_v^{(t)} + \mathbf{1}_n\mathbf{b}^{(t)^T} - \mathbf{Y}^T)(\mathbf{W}_v^{(t)^T}\boldsymbol{\Theta}_v^{(t)}\mathbf{X}_v^T + \mathbf{b}^{(t)}\mathbf{1}_n^T - \mathbf{Y})\Big) \\ &+\alpha\left\|\mathbf{W}_v^{(t)}\right\|_{2,1} + \beta tr(\mathbf{W}_v^{(t)^T}\mathbf{X}_v^T\mathbf{L}_v\mathbf{X}_v\mathbf{W}_v^{(t)}) \end{aligned} \tag{42}$$

From Equation (42), we can see that the proposed algorithm is monotonically decreasing and convergent. Thus, theorem 1 is proved.

4. Experiment

In this section, we compare the proposed algorithm with six comparison algorithms on three ADNI sub-datasets.

4.1. Dataset

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[1]. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

We downloaded three sub-datasets from the ADNI website, namely AD vs. NC (sick state vs. normal control), MCI vs. NC (moderate cognitive impairment vs. normal control), and pMCI vs. sMCI (Progress MCI vs. stable MCI). ADNI is the authoritative data center for studying Alzheimer's disease. It was jointly funded by the national institutes of health and the national institute of aging in 2004. It is dedicated to collecting and sorting out the data of Alzheimer's disease patients, tracking the onset process of patients, exploring the changes and causes of the onset process, so as to reveal the pathogenesis of Alzheimer's disease and find a cure. It includes clinical data, magnetic resonance imaging data, positron emission computed tomography data, genetic data, and biological sample data.

In this paper, we obtained basic MRI, PET, and CSF data from 202 experimental subjects (including 51 AD subjects, 52

[1] adni.loni.usc.edu

TABLE 1 Demographic information of the subjects.

	AD (51)	NC (52)	MCI-C (43)	MCI-NC (56)
Female/male	18/33	18/34	15/28	17/39
Age	75.2 ± 7.4	75.3 ± 5.2	75.8 ± 6.8	74.8 ± 7.1
Education	14.7 ± 3.6	15.8 ± 3.2	16.1 ± 2.6	15.8 ± 3.2
MMSE	23.8 ± 2.0	29.0 ± 1.2	26.6 ± 1.7	28.4 ± 1.7
adas-Cog	18.3 ± 6.0	12.1 ± 3.8	12.9 ± 3.9	8.03 ± 3.8

The numbers in parentheses denote the number of subjects in each clinical category. (MCI-C, MCI converters; MCI-NC, MCI non-converters; Mean ± SD).

NC subjects, and 99 MCI subjects). Specifically, we retained 93 features from MRI as the first modality, 93 features from PET as the second modality and three features from CSF as the third modality. Detailed object information is shown in Table 1.

4.2. Comparison algorithm

OMVFS (Online unsupervised Multi-View Feature Selection; Shao et al., 2016): this method is a multi-modal feature selection algorithm. It does not store all data, but performs data processing step by step to compress the required data into a matrix. In addition, it also combines graph regularization term, sparse learning and non negative matrix decomposition technology to select features.

K-OFSD (Online Feature Selection based on the Dependency in K nearest neighbors; Zhou et al., 2017): it is a feature selection algorithm for class imbalance data. Specifically, according to the neighborhood rough set theory, it uses the information of k-nearest neighbors to select features, so as to improve the separability between the majority class and the minority class. In addition, it also uses the relationship between labels and features to obtain the importance of each feature.

RLSR (Rescaled Linear Square Regression; Chen et al., 2017): this method is a semi-supervised feature selection algorithm. It scales the regression sparsity of the least squares loss function again by using the scaling factor, so as to obtain the weight of each feature. In addition, it also explains that this method can learn the global structure of data and get sparse solution.

PMFS (Pareto-based feature selection algorithm for multi-label classification; Hashemi et al., 2021): this method is a pareto based feature selection algorithm. Specifically, it first establishes a dual objective optimization model of feature redundancy and feature correlation by using multiple labels. Then, it uses pareto to solve the established model in the previous step. Finally, it verifies the performance of the proposed method in experiments.

MDFS (Embedded feature selection method *via* manifold regularization; Zhang et al., 2019): this method is a multi-label feature selection algorithm. It uses the original features to construct a low dimensional embedding to learn the local structure information of the data. In addition, it embeds $l_{2,1}-$norm into the proposed objective function to select the feature subset.

SDFS (Sparsity Discriminant Feature Selection; Wang et al., 2020): this method is a feature selection algorithm based on $l_{2,0}-$norm. It uses structured sparse subspace constraints to overcome the problem of parameter adjustment. In addition, it also uses the objective function to improve the resolution of the model.

4.3. Experimental setup

In this section, we conducted 10-fold cross validation experiments. Specifically, we first carried out comparative experiments on classification accuracy (acc), sensitivity (sen), specificity (spe), and area under curve (auc) of all algorithms on three datasets. Then, we carry out the parameter sensitivity experiment of the proposed algorithm. Finally, we verify the convergence of the proposed algorithm on three datasets. For all algorithms, after obtaining the selected feature subset, we use support vector machine (SVM) to classify them, so as to compare the performance of all algorithms. For parameters α and β. We set their value range as $\alpha, \beta \in \{10^{-3}, 10^{-2}, 10, 1, 10, 10^{2}, 10^{3}\}$. In addition, we also set the convergence condition of the proposed algorithm as $\frac{|obj(t+1)-obj(t)|}{obj(t)} \leq 10^{-5}$, where $obj(t)$ and $obj(t+1)$ represent the values of the objective function in the t-th iteration and the $(t+1)$-th iteration, respectively.

4.4. Analysis of experimental results

Figure 1 shows the classification accuracy of each fold of all algorithms on three datasets. From Figure 1, we can see that the classification accuracy of all algorithms is not very stable due to the randomness of 10 fold cross validation. From the first and third subgraphs, we can see that PMFS and SDFS have the worst performance. Therefore, we also carried out experiments on the average classification accuracy, average sen, average spe, and average auc of all algorithms on three datasets, as shown in Table 2. From Table 2, we can see that the proposed algorithm achieves the best classification accuracy. Specifically, compared with the worst comparison algorithm PMFS and the best comparison algorithm K-OFSD, the proposed algorithm improves by 8.27 and 0.93%, respectively on the AD vs. NC dataset. On the MCI vs. NC dataset, compared with the poor comparison algorithms OMVFS, K-OFSD, RLSR, and the best comparison algorithm MDFS, the proposed algorithm has improved by 0.71 and 0.34%, respectively. On the pMCI vs. sMCI dataset, compared with the worst comparison algorithm SDFS and the best comparison algorithm MDFS, the proposed algorithm improves by 5.95 and 1.2%, respectively. The reason

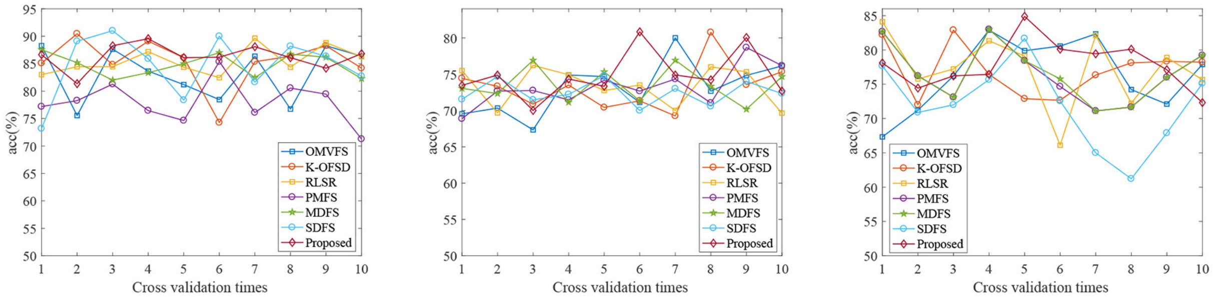

FIGURE 1
Classification accuracy results of all algorithms on three datasets (From left to right, the results on datasets AD vs. NC, MCI vs. NC, and pMCI vs. sMCI are in turn).

TABLE 2 Classification results of all algorithms on the three datasets (%).

Datasets	AD vs. NC				MCI vs. NC				pMCI vs. sMCI			
	acc	sen	spe	auc	acc	sen	spe	auc	acc	sen	spe	auc
OMVFS	81.70	81.27	81.70	90.76	73.15	9.08	99.21	38.73	76.32	71.18	82.13	81.76
K-OFSD	85.40	85.12	85.80	92.25	73.15	9.08	99.21	**38.73**	76.42	71.18	82.30	81.71
RLSR	84.24	85.46	83.21	93.08	73.15	9.08	99.21	38.68	76.32	71.18	82.13	81.75
PMFS	78.06	76.95	77.94	86.94	73.28	8.80	99.26	37.24	76.61	**73.18**	80.92	**83.30**
MDFS	84.77	84.99	84.14	92.64	73.52	8.87	99.30	37.84	76.72	**73.18**	81.06	83.13
SDFS	84.65	85.89	85.44	93.49	72.46	**36.08**	87.20	33.21	71.97	67.69	78.99	82.84
Proposed	**86.33**	**87.22**	**86.90**	**93.76**	**73.86**	10.15	**99.52**	37.02	**77.92**	72.34	**82.38**	81.98

The bold values indicate the best experimental results.

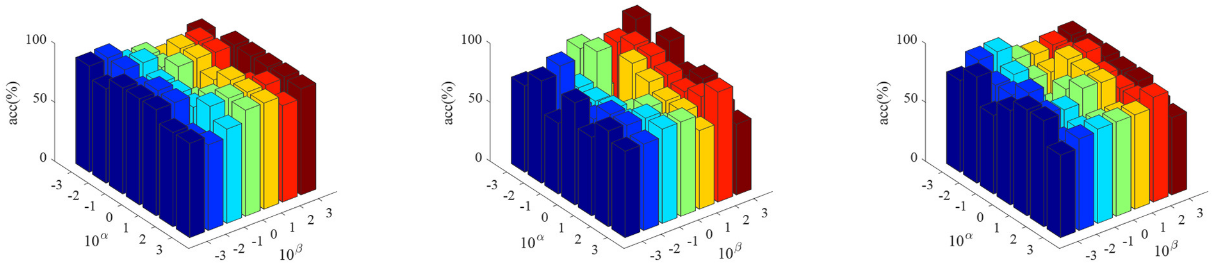

FIGURE 2
The classification accuracy of the proposed algorithm varies with different parameter values. (From left to right, the results on datasets AD vs. NC, MCI vs. NC, and pMCI vs. sMCI are in turn.)

for this phenomenon is that the proposed algorithm not only considers the relationship between different modes, but also considers the graph structure information in multi-modal data.

Figure 2 shows the parameter sensitivity of the proposed algorithm, i.e., the classification accuracy of the proposed algorithm changes with the change of the values of parameters α and β. From Figure 2, we can see that the performance of the proposed algorithm will be affected by the parameter values. Therefore, we need to carefully adjust the values of parameters α and β. In addition, we also conducted the convergence experiment of the algorithm, as shown in Figure 3. From Figure 3, we can see that the proposed algorithm has good convergence. On the three datasets, the convergence was achieved within 10 iterations of the objective function. This shows that the proposed algorithm has fast convergence effect.

5. Conclusion

In this paper, we have proposed a multi-modal feature selection algorithm with anchor graph for Alzheimer's disease.

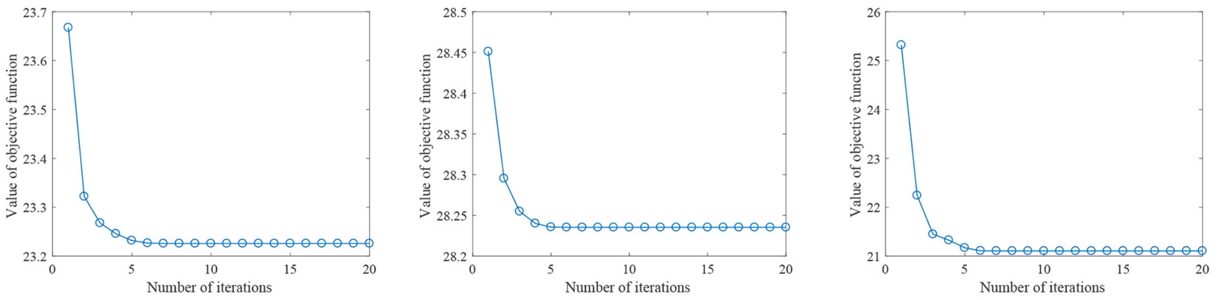

FIGURE 3
The value of the proposed objective function changes with the number of iterations (From left to right, the results on datasets AD vs. NC, MCI vs. NC, and pMCI vs. sMCI are in turn.)

It can be used in the early auxiliary diagnosis of Alzheimer's disease. Specifically, we use the least square, $l_{2,1}-$norm and anchor graph regular term to learn the importance of modes, the weight of features and the local structure information of data. In addition, we also prove the convergence of the proposed method. Finally, on the three datasets i.e., AD vs. NC, MCI vs. NC, and pMCI vs. sMCI, we verify the validity of the proposed method and compare it with other advanced comparison algorithms. In the future work, we plan to study new representation methods of multi-modal data, so as to carry out more efficient feature selection for Alzheimer's disease.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Alzheimer's Disease Neuroimaging Initiative

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Author contributions

JL, HX, HY, and ZJ contributed to conception and design of the study. HY organized the database. JL performed the statistical analysis and wrote the first draft of the manuscript. JL, HX, HY, ZJ, and LZ wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work has been supported in part by the Natural Science Foundation of China under grant 61836016 and National Natural Science Foundation of China under grant 62072166, Natural Science Foundation of Hunan Province under grant 2022JJ40190. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI was funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer's Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern

California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bi, X.-A., Hu, X., Wu, H., and Wang, Y. (2020). Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest. *IEEE J. Biomed. Health Informatics* 24, 2973–2983. doi: 10.1109/JBHI.2020.2973324

Chaves, R., Ramírez, J., Górriz, J., Puntonet, C. G., Initiative, A. D. N., et al. (2012). Association rule-based feature selection method for Alzheimer's disease diagnosis. *Expert Syst. Appl.* 39, 11766–11774. doi: 10.1016/j.eswa.2012.04.075

Chen, X., Yuan, G., Nie, F., and Huang, J. Z. (2017). "Semi-supervised feature selection *via* rescaled linear regression," in *IJCAI, Vol. 2017* (Melbourne, NSW: Springer), 1525–1531. doi: 10.24963/ijcai.2017/211

Chyzhyk, D., Savio, A., and Gra na, M. (2014). Evolutionary elm wrapper feature selection for Alzheimer's disease cad on anatomical brain MRI. *Neurocomputing* 128, 73–80. doi: 10.1016/j.neucom.2013.01.065

Gallego-Jutglà, E., Solé-Casals, J., Vialatte, F.-B., Elgendi, M., Cichocki, A., and Dauwels, J. (2015). A hybrid feature selection approach for the early diagnosis of Alzheimer's disease. *J. Neural Eng.* 12:016018. doi: 10.1088/1741-2560/12/1/016018

Hao, X., Bao, Y., Guo, Y., Yu, M., Zhang, D., Risacher, S. L., et al. (2020). Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer's disease. *Med. Image Anal.* 60:101625. doi: 10.1016/j.media.2019.101625

Hashemi, A., Dowlatshahi, M. B., and Nezamabadi-pour, H. (2021). An efficient pareto-based feature selection algorithm for multi-label classification. *Inform. Sci.* 581, 428–447. doi: 10.1016/j.ins.2021.09.052

Jiao, Z., Chen, S., Shi, H., and Xu, J. (2022). Multi-modal feature selection with feature correlation and feature structure fusion for MCI and AD classification. *Brain Sci.* 12:80. doi: 10.3390/brainsci12010080

Jie, B., Zhang, D., Cheng, B., and Shen, D. (2013). "Manifold regularized multi-task feature selection for multi-modality classification in Alzheimer's disease," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Berlin: Springer), 275–283. doi: 10.1007/978-3-642-40811-3_35

Li, J., Wu, L., Wen, G., and Li, Z. (2019). Exclusive feature selection and multi-view learning for Alzheimer's disease. *J. Vis. Commun. Image Represent.* 64:102605. doi: 10.1016/j.jvcir.2019.102605

Liu, F., Wee, C.-Y., Chen, H., and Shen, D. (2014). Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. *NeuroImage* 84, 466–475. doi: 10.1016/j.neuroimage.2013.09.015

Liu, Y., Li, Z., Ge, Q., Lin, N., and Xiong, M. (2019). Deep feature selection and causal analysis of Alzheimer's disease. *Front. Neurosci.* 13:1198. doi: 10.3389/fnins.2019.01198

Mahendran, N., and Vincent, D. R. (2022). A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease. *Comput. Biol. Med.* 141:105056. doi: 10.1016/j.compbiomed.2021.105056

Nie, F., Liu, C., Wang, R., Wang, Z., and Li, X. (2021). Fast fuzzy clustering based on anchor graph. *IEEE Trans. Fuzzy Syst.* 30, 2375–2387. doi: 10.1109/TFUZZ.2021.3081990

Nie, F., Wang, X., and Huang, H. (2014). "Clustering and projected clustering with adaptive neighbors," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Beijing: ACM), 977–986. doi: 10.1145/2623330.2623726

Niyas, M. K. P., and Thiyagarajan, P. (2022). Feature selection using efficient fusion of fisher score and greedy searching for Alzheimer's classification. *J. King Saud Univ. Comput. Inform. Sci.* 34, 4993–5006. doi: 10.1016/j.jksuci.2020.12.009

Rani Kaka, J., and Prasad, K. S. (2021). Alzheimer's disease detection using correlation based ensemble feature selection and multi support vector machine. *Int. J. Comput. Digit. Syst.* 9–20.

Shao, W., He, L., Lu, C.-T., Wei, X., and Philip, S. Y. (2016). "Online unsupervised multi-view feature selection," in *2016 IEEE 16th International Conference on Data Mining* (Barcelona: IEEE), 1203–1208. doi: 10.1109/ICDM.2016.0160

Shao, W., Peng, Y., Zu, C., Wang, M., Zhang, D., Alzheimer's Disease Neuroimaging Initiative (2020). Hypergraph based multi-task feature selection for multimodal classification of Alzheimer's disease. *Comput. Med. Imaging Graph.* 80:101663. doi: 10.1016/j.compmedimag.2019.101663

Thapa, S., Singh, P., Jain, D. K., Bharill, N., Gupta, A., and Prasad, M. (2020). "Data-driven approach based on feature selection technique for early diagnosis of Alzheimer's disease," in *2020 International Joint Conference on Neural Networks* (Glasgow: IJCNN), 1–8. doi: 10.1109/IJCNN48605.2020.9207359

Wang, Z., Nie, F., Tian, L., Wang, R., and Li, X. (2020). "Discriminative feature selection *via* a structured sparse subspace learning module," in *IJCAI* (Yokohama), 3009–3015. doi: 10.24963/ijcai.2020/416

Yu, H., Zhang, C., Li, J., and Zhang, S. (2022). Robust sparse weighted classification for crowdsourcing. *IEEE Trans. Knowl. Data Eng.* doi: 10.1109/TKDE.2022.3201955

Zhang, C., Song, J., Zhu, X., Zhu, L., and Zhang, S. (2021). HCMSL: hybrid cross-modal similarity learning for cross-modal retrieval. *ACM Trans. Multim. Comput. Commun. Appl.* 17, 1–22. doi: 10.1145/3412847

Zhang, J., Luo, Z., Li, C., Zhou, C., and Li, S. (2019). Manifold regularized discriminative feature selection for multi-label learning. *Pattern Recogn.* 95, 136–150. doi: 10.1016/j.patcog.2019.06.003

Zhang, M., Yang, Y., Shen, F., Zhang, H., and Wang, Y. (2017). Multi-view feature selection and classification for Alzheimer's disease diagnosis. *Multim. Tools Appl.* 76, 10761–10775. doi: 10.1007/s11042-015-3173-5

Zhang, S., and Li, J. (2021). KNN classification with one-step computation. *IEEE Trans. Knowl. Data Eng.* doi: 10.1109/TKDE.2021.3119140

Zhang, S., Li, J., and Li, Y. (2022a). Reachable distance function for KNN classification. *IEEE Trans. Knowl. Data Eng.* 1–15. doi: 10.1109/TKDE.2022.3185149

Zhang, S., Li, J., Zhang, W., and Qin, Y. (2022b). Hyper-class representation of data. *Neurocomputing* 503, 200–218. doi: 10.1016/j.neucom.2022.06.082

Zhang, Y., Wang, S., Xia, K., Jiang, Y., Qian, P., Alzheimer's Disease Neuroimaging Initiative. (2021). Alzheimer's disease multiclass diagnosis *via* multimodal neuroimaging embedding feature selection and fusion. *Inform. Fus.* 66, 170–183. doi: 10.1016/j.inffus.2020.09.002

Zhou, P., Hu, X., Li, P., and Wu, X. (2017). Online feature selection for high-dimensional class-imbalanced data. *Knowl. Based Syst.* 136, 187–199. doi: 10.1016/j.knosys.2017.09.006

Zhu, L., Zhang, C., Song, J., Zhang, S., Tian, C., and Zhu, X. (2022). Deep multi-graph hierarchical enhanced semantic representation for cross-modal retrieval. *IEEE MultiMedia.* 29, 17–26. doi: 10.1109/ICME51207.2021.9428194

Zhu, X., Suk, H.-I., and Shen, D. (2014). "Multi-modality canonical feature selection for Alzheimer's disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Interventioni* (Cham: Springer), 162–169. doi: 10.1007/978-3-319-10470-6_21