Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 1583969, 9 pages
https://doi.org/10.1155/2018/1583969

*Research Article*

# Feature Selection and Classification for High-Dimensional Incomplete Multimodal Data

**Wan-Yu Deng, Dan Liu, and Ying-Ying Dong**

*School of Computer, Xi'an University of Post & Telecommunications, Shaanxi, China*

Correspondence should be addressed to Dan Liu; isliudan@163.com

Received 4 May 2018; Accepted 26 July 2018; Published 12 August 2018

Academic Editor: Arkadiusz Zak



Due to missing values, incomplete dataset is ubiquitous in multimodal scene. Complete data is a prerequisite of the most existing multimodality data fusion methods. For incomplete multimodal high-dimensional data, we propose a feature selection and classification method. Our method mainly focuses on extracting the most relevant features from the high-dimensional features and then improving the classification accuracy. The experimental results show that our method produces considerably better performance on incomplete multimodal data such as ADNI dataset and Office dataset, compared to the case of complete data.

## 1. Introduction

In the era of Internet, there are many different modalities, such as images, video, and text. Different modalities can provide complementary information; therefore, multimodal classification can generally produce better performance than individual modality in accuracy and reliability. The diagnoses of Alzheimer's Disease (AD) by multimodal classification are a great example and have achieved remarkable success compared to single modal methods in multiple experiments. Pang et al. [1] explored the possibility of improving emotion prediction by highly nonlinear relationships between low-level features in different modalities. Zhang et al. [2] incorporated three modalities of biomarkers (structural MR imaging (MRI), Positron-Emission Tomography (PET), and cerebrospinal fluid (CSF)) to discriminate AD (or mild cognitive impairment (MCI)) from healthy controls. Pang et al. [3] recommended using multilabel multiple-kernel learning with visual and textual features for multilabel image classification. Hu et al. [4] utilized multimodality data including both tag feature and visual feature for popularity prediction on social media. Ballard [5] suggested a multimodal learning interface which could learn words from natural interactions with users. Liu et al. [6] mentioned a multihypergraph learning (MHL) method to deal with multimodality data. This method achieved promising results in AD/MCI classification. Zhang et al. [7] proposed multimodal multitask learning to jointly predict multiple variables from multimodal data. Liu et al. [8] proposed a linearized and kernelized sparse multitask learning for predicting cognitive outcomes in Alzheimer's Disease. Li et al. [9, 10] proposed a multitask deep learning method for diagnosing Alzheimer's Disease by combining MRI, PET, and Assessment Scale-Cognitive subscale (ADAS-Cog) with the restricted Boltzmann machine. Wang et al. [11] explained a novel multimodality multicenter classification method for autism spectrum disorder diagnosis; they regarded the classification of each imaging center as one task and solved the classification for all imaging centers by introducing the task-task and modality-modality regularizations. Liu et al. [12] proposed a view-aligned hypergraph learning (VAHL) method and utilized incomplete multimodality data for AD/MCI diagnosis.

Complete data is a prerequisite of the most existing multimodality data fusion methods. Since complete data requires the modality type and the modality number to be consistent, it is rare in reality. In Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, for example, only about 1/3 of its total samples contain complete MRI, PET, and CSF data at baseline. In view of incomplete multimodality data, it usually explores imputing the missing values [13, 14] and discarding

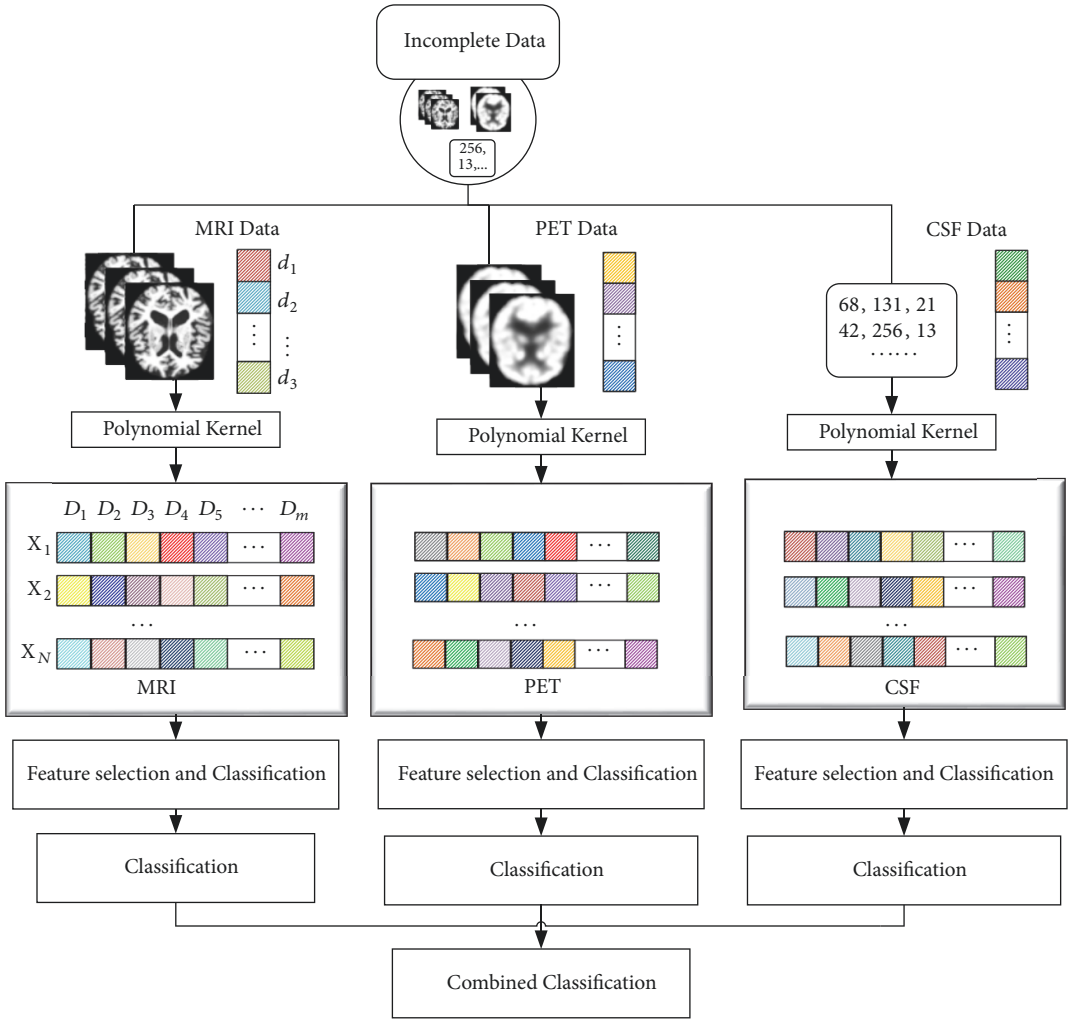


Figure 1: Multimodal classification framework based on high-dimensional feature selection.

samples, and doing these will lead to waste of or bring unpredictable noise. To address the incomplete multimodality data, Zhao et al. [15] proposed an unsupervised method which processes the incomplete multimodality data by transforming the original and incomplete data to a new and complete representation in a latent space. Thung et al. [16] used incomplete multimodal dataset via matrix shrinkage and completion to identify AD patients. Li et al. [17] proposed a pioneer work to handle two-modal incomplete data case by projecting the partial data into a common latent subspace via nonnegative matrix factorization (NMF) and $l_1$ sparse regularizer. Following this line, Shao et al. [18] proposed a similar idea of weighted NMF and $l_{2,1}$ regularizer.

Most existing incomplete multimodality methods have low efficiency with high-dimensional data. Inspired by this, we propose a feature selection and classification for incomplete multimodal high-dimensional data. Our method has the following features: (1) It focuses on incomplete data and makes full use of the data from different modalities without data wasting. (2) It selects the most relevant features in high-dimensional space and facilitates the discovery of the inherent relationship between features. (3) It achieves better classification accuracy when compared with the other methods.

The rest of the paper will demonstrate the details of the proposed approach; experiments on various datasets and comparison between our method and the currently most advanced methods.

## 2. Multimodal High-Dimensional Feature Selection and Classification

*2.1. The Framework.* Figure 1 is the schematic illustration of our feature selection and classification framework for incomplete multimodal high-dimensional data. It contains three parts, polynomial kernel explicit expansion, feature selection, and classification. We start with the exploitation of the polynomial kernel explicit method to expand the original low-order feature to high-order feature. Then the identification of the high-order feature subsets by feature selection for incomplete multimodality data is achieved. Next, we classify AD patients and healthy controls. Finally, we integrate classifiers and export the final classification accuracy.

*2.2. The Explicit Mapping.* In a set of data samples $\mathrm{X} = [\mathrm{x}_1, \ldots, \mathrm{x}_i, \ldots, \mathrm{x}_N]$, $i = 1, \ldots, N$, $N$ is the total number of samples. Each sample has $m$ modalities; i.e., $\mathrm{x}_i = [\mathrm{x}_i^{(1)}, \mathrm{x}_i^{(2)}, \ldots, \mathrm{x}_i^{(m)}]$. For incomplete multimodality data, we define $\widehat{\mathrm{X}} = \{\{\mathrm{X}_1^{(1)}, \cdots, \mathrm{X}_{N-k_1}^{(1)}\}, \{\mathrm{X}_1^{(2)}, \cdots, \mathrm{X}_{N-k_2}^{(2)}\}, \cdots, \{\mathrm{X}_1^{(m)}, \cdots, \mathrm{X}_{N-k_m}^{(m)}\}\}$ as an incomplete modal sample set. In $\mathrm{X}^{(\mathrm{m})} \in \mathbb{R}^{(N-k_\mathrm{m})\times d_m}$, $k_m$, $d_m$ represent the number of missing samples and the feature dimension in the $m$th modalities, respectively. $\mathrm{Y} = [y_1, \ldots, y_i, \ldots, y_N]$ is the corresponding label of X, and label $y_i \in \mathrm{L} = \{1, \ldots, c\}$, and $c$ is the number of data category. Since complete data discard a lot of precious data, the size of complete data is relatively small. As previously indicated, incomplete data is a collection of the various modal data subsets.

First-order spatial selection is difficult to reveal the high-order dependency relationship between features. Since the incomplete data has limitations about the number of samples and the feature dimensions, we need more features to discover the correlation between them. We can consider using different kernel functions to extend the data, for example, a linear kernel function or Gaussian kernel function. The linear kernel function directly linearizes the data, which makes it difficult to reflect the correlation between the features and may result in the loss of data. The Gaussian kernel function is too expensive to calculate. Therefore, we transform low-dimensional features into high-dimensional features by nonlinear kernel explicit expression and then reveal high-order correlation between features.

For degree-d polynomials, the polynomial kernel is defined as

$$k(x, y) = \left(\gamma x^{\mathrm{T}} y + \rho\right)^d \tag{1}$$

where $x$ and $y$ are vectors in the input space, and $\rho \geq 0$ is a free parameter trading off the influence of higher-order versus lower-order terms in the polynomial. As a kernel, $k$ corresponds to an inner product in a feature space based on some mapping $\varphi$:

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle \tag{2}$$

Let $d = 2$, and we get the special case of the quadratic kernel. After using the multinomial theorem and regrouping,

$$\begin{aligned} k(x, y) &= \left(\sum_{i=1}^{m} x_i y_i + \rho\right)^2 \\ &= \sum_{i=1}^{m} \left(x_i^2\right)\left(y_i^2\right) + \sum_{i=2}^{m}\sum_{j=1}^{i-1} \left(\sqrt{2}x_i x_j\right)\left(\sqrt{2}y_i y_j\right) \\ &\quad + \sum_{i=1}^{m} \left(\sqrt{2\rho}x_i\right)\left(\sqrt{2\rho}y_i\right) + \rho^2 \end{aligned} \tag{3}$$

From this, the explicit feature mapping of polynomial kernel is

$$\begin{aligned} \varphi(x) = \Big(& x_m^2, \ldots, x_1^2, \sqrt{2}x_m x_{m-1}, \ldots, \sqrt{2}x_m x_1, \\ & \sqrt{2}x_{m-1}x_{m-2}, \ldots, \sqrt{2}x_{m-1}x_1, \ldots, \sqrt{2}x_2 x_1, \sqrt{2\rho}x_m, \ldots, \\ & \sqrt{2\rho}x_1, \rho\Big) \end{aligned} \tag{4}$$

Compared with the linear case, the second-order feature map contains dependency of the feature pair. The key problem of this explicit feature mapping is that its features are high dimensional in the extended feature space. For polynomial kernel expansion, the dimension of the feature map increases exponentially. When $d = 2$, $m$ is the original feature dimension, and extended dimension is $(m+2)(m+1)/2$. Generally, when $m = 10^6$, the extended dimension is approximately $10^{12}$.

*2.3. Feature Selection and Classification.* At first, we introduce a feature selection vector $\mathrm{d} \in \{0, 1\}$, whose entries are 1 for selected features and 0 otherwise. Let $D = \{\mathrm{d} \mid \mathrm{d} \in \{0, 1\}\}$ be the domain of d. We use $\|\mathrm{d}\|_1 \leq B$ to control the sparsity of the feature selection, where $B$ controls the number of selected features; then the proposed problem can be written as

$$\begin{aligned} \min_{\mathrm{d}} \min_{w,\xi} \quad & \frac{1}{2}\|\mathrm{w}\|_2^2 + \frac{C}{2}\sum_{i=1}^{N}\xi_i^2 \\ \text{s.t.} \quad & \mathrm{w}\left(\mathrm{x}_i \odot \mathrm{d}\right) - y_i = \xi_i, \quad i = 1, \cdots, N \end{aligned} \tag{5}$$

in which constant $C$ is a regularization parameter that makes a trade-off between the model complexity and the fitness of the feature selection.

By introducing the dual variable $\alpha$, $\alpha \in A = \{\alpha \mid \alpha_i \geq 0,\ i = 1, \cdots, n\}$, the Lagrangian function of (5) can be written as

$$\begin{aligned} L(\mathrm{w}, \xi, \alpha) = & \frac{1}{2}\|\mathrm{w}\|_2^2 + \frac{C}{2}\sum_{i=1}^{N}\xi_i^2 \\ & + \left\langle \alpha, \xi_i - \mathrm{w}'\left(\mathrm{x}_i \odot \mathrm{d}\right) + y_i \right\rangle \end{aligned} \tag{6}$$

in which $\langle \cdot, \cdot \rangle$ denotes the inner product. By setting the derivatives of $L(\mathrm{w}, \xi, \alpha)$ with respect to w and $\xi$ to zero, we can obtain the Karush-Kuhn-Tucker (KKT) conditions, $w = \alpha(x_i \odot d)$ and $\xi_i = -\alpha/C$. By substituting the above results into the Lagrangian function, problem (5) can be transformed into the following dual formulation:

$$\min_{\mathrm{d}\in D} \max_{\alpha \in A} f(\alpha, \mathrm{d}) \tag{7}$$

where $f(\alpha, \mathrm{d}) = (1/2)\sum_{j=1}^{m} d_j^2 w_j^2 + (1/2C)\alpha'\alpha$ and $w = \sum_{i=1}^{n} \alpha_i y_i \mathrm{x}_i$.

Since the feature selection vector is zero-one vector, this is still a nonconvex problem. Following the convex relaxation in [19], we have

$$\min_{\mathrm{d}\in D} \max_{\alpha \in A} f(\alpha, \mathrm{d}) \geq \max_{\alpha \in A} \min_{\mathrm{d}\in D} f(\alpha, \mathrm{d}) \tag{8}$$

(1) Initialize $\alpha = 1$ and the constraint subset $C = \phi$.
Let iteration index $T = 1$.
(2) Find the most active constraint $\mathrm{d}_T$, update set $C$ by $C = C \cup \{\mathrm{d}_T\}$.
(3) Update $\alpha$ by solving the problem (10).
(4) Let $T = T + 1$. Repeat step (2)-(3) until convergence.

Algorithm 1: The cutting plane algorithm.

By introducing an additional variable $\theta \in \mathbb{R}$, the above problem can be converted into the following convex quadratically constrained quadratic program (QCQP) problem:

$$\begin{aligned} \max_{\alpha \in A, \theta \in \mathbb{R}} \quad & \theta \\ \text{s.t.} \quad & \theta \leq f(\alpha, \mathrm{d}), \quad \forall \mathrm{d} \in D \end{aligned} \tag{9}$$

It is very hard to solve as there are infinite number of quadratic inequality constraints in (9), and we solve this problem by the cutting plane algorithm [20]. We generate an active constraint and add it to an active constraint set $C$ which is initialized to empty set $\phi$. The active constraint set $C$ is a subset of $D$; i.e., $C \subseteq D$. Based on a new active constraint set $C$, we need to solve QCQP problem to update $\alpha$. Specifically, we need to solve the following problem:

$$\begin{aligned} \max_{\alpha \in A, \theta \in \mathbb{R}} \quad & \theta \\ \text{s.t.} \quad & \theta \leq f(\alpha, \mathrm{d}_T), \quad \forall \mathrm{d}_T \in C \end{aligned} \tag{10}$$

The cutting plane algorithm can be presented in Algorithm 1.

*2.3.1. Learning d.* The cutting plane algorithm mainly deals with how to find the most active constraint $\mathrm{d}_T$ of problem (10) at the $T$th iteration. Let $c_j = w_j^2$, and the optimization problem becomes

$$\frac{1}{2}\sum_{j=1}^{m} w_j^2 d_j^2 = \frac{1}{2}\sum_{j=1}^{m} c_j d_j^2 \tag{11}$$

Due to $d_j \in \{0, 1\}$, problem (11) can be solved by sorting $c_j$, and then find the largest $c_j$.

*2.3.2. The Optimization of α.* After updating the active constraint set $C$, we solve the problem in (10) with constraints which are defined by $C$. Because the number of constraints in $C$ is no longer large, we can use subgradient method to solve this problem. However, it is very expensive to get the dual variables $\alpha$ when $n$ is very large.

In view of the above problems, problem (10) can be solved in the following primal form:

$$\min_{\mathrm{w}} \frac{1}{2}\left(\sum_{t=1}^{T} \|w_t\|\right)^2 + g(w) \tag{12}$$

where $g(w) = (C/2)\sum_{i=1}^{N} \xi_i^2$ and $\xi_i = \mathrm{w}'(\mathrm{x}_i \odot \mathrm{d}) - y_i$.

For convenience, we define $p(w) = (1/2)(\sum_{t=1}^{T} \|w_t\|)^2$ and $F(w) = p(w) + g(w)$. We solve the primal problem by using the accelerated proximal gradient (APG) [21], which minimizes the following quadratic approximation of (12):

$$\begin{aligned} Q_\tau(w, v) &= g(v) + \langle \nabla g(v), w - v \rangle + \frac{\tau}{2}\|w - v\|^2 \\ &\quad + p(w) \\ &= \frac{\tau}{2}\|w - z\|_F^2 + p(w) + g(v) \\ &\quad - \frac{1}{2\tau}\|\nabla g(v)\|_F^2 \end{aligned} \tag{13}$$

in which $\nabla g$ denotes the gradient of $g$ at point $v, \tau > 0$ denotes the Lipschitz constant, and $z = v - (1/\tau)\nabla g(v)$. We need to solve the following Moreau Projection problem:

$$\min_{w} \quad \frac{\tau}{2}\|w - z\|_F^2 + p(w) \tag{14}$$

Problem (14) has a unique closed-form solution, it can be solved in the following manner via Moreau Projection [22].

Suppose $S_\tau(z)$ be the optimal solution to problem (14) and $o = [o_1, o_2, \cdots, o_k] \in \mathbb{R}^k$ be an intermediate variable. Then $S_\tau(z)$ is unique and can be calculated as

$$S_\tau(z_t) = \begin{cases} \dfrac{o_t}{\|z_t\|} z_t, & if\ o_t \geq o, \\ 0, & otherwise. \end{cases} \tag{15}$$

in which $t = \{1, 2, \cdots, k\}$. The intermediate vector $o_t$ can be calculated via a soft-threshold operator $soft(u, \varsigma)$ in [22, 23]

$$o_t = [soft(u, \varsigma)]_t = \begin{cases} u_t - \varsigma, & if\ u_t \geq \varsigma \\ 0, & Otherwise. \end{cases} \tag{16}$$

and the threshold value $\varsigma$ can be calculated as in step (4) of Algorithm 2.

The overall APG algorithm for solving problem (14) is summarized in Algorithm 3. We can obtain $w$ from APG and then predict the results of each modality by our method. It can be expressed as $y = w(\mathrm{x}_i \odot \mathrm{d})$. Eventually, the integration of the prediction result will enable us to do classification.

## 3. Performance Evaluation

*3.1. Datasets.* We evaluate the performance of our method by employing the ADNI and Office dataset, respectively. The ADNI dataset was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and nonprofit organizations. The primary purpose of ADNI project was to study the effects of combining multiple biomarkers, such as MRI, PET, and CSF data accompanied with neuropsychological assessments, to predict the progression of MCI and early AD. We employ a 3-modality (MRI, PET, and CSF) dataset with 103 subjects which include 51 AD patients and 52 healthy controls. The

Table 1: Subject information.

| | AD (n=51;18F/33M) | | | HC (n=52; 18F/34M) | | |
|---|---|---|---|---|---|---|
| | Mean | SD | Range | Mean | SD | Range |
| Age | 75.2 | 7.4 | 59-88 | 75.3 | 5.2 | 62-85 |
| Education | 14.7 | 3.6 | 4-20 | 15.8 | 3.2 | 8-20 |
| MMSE | 23.8 | 2.0 | 20-26 | 29 | 1.2 | 25-30 |
| CDR | 0.7 | 0.3 | 0.5-1 | 0 | 0.0 | 0-0 |

The numbers refer to baseline data. AD=Alzheimer's Disease, HC=Healthy Control, MMSE= Mini-Mental State Examination, and CDR= Clinical Dementia Rating.

Table 2: Summarization of Office dataset.

| Data | Type of feature | Original the dimension of feature | Extended the dimension of feature | The number of categories |
|---|---|---|---|---|
| amazon | Decaf-LeNet/ | 4096 | | |
| dslr | Surf/ | 800 | 180902 | 10 |
| webcam | Decaf-AlexNet | 4096 | | |

Mini-Mental State Examination (MMSE) is a standardized cognitive impairment examination method for screening Alzheimer's Disease. MMSE scores between 24 and 30 and a Clinical Dementia Rating (CDR) of 0 are designated as healthy controls; MMSE scores between 20 and 26 and CDR of 0.5 or 1.0 are considered as AD. Table 1 lists the demographics of all these subjects.

The multimodality data has 189 dimensionality features; for each subject, we obtain 93 features from MRI image, another 93 features from PET image, and 3 features from the CSF biomarkers. The size of feature dimension is relatively small. The nonlinear explicit expression is used to expand the dimension of data. After each item becomes one dimension, the 189 features are expanded into 8940 features. Now we can obtain the feature of a combination high-order disease.

Office dataset is as follows: amazon (e.g., images downloaded from the Internet), webcam (e.g., low-resolution images captured by web cameras), and dslr (e.g., high-resolution images taken from digital SLR cameras). Each dataset has 10 object classes. Specifically, Surf and Decaf features are extracted for all the images, and Decaf-LeNet and Decaf-AlexNet represent different Decaf features by training LeNet and AlexNet model, respectively. The feature dimension of Surf is 800 and the feature dimension of Decaf by training LeNet and AlexNet model is 4096, respectively. Table 2 lists the summarization of Office dataset. We expand these features into 180902 dimensional features by applying nonlinear explicit polynomial expression method.

*3.2. Results on ADNI.* We first use a 10-fold cross-validation strategy to classify AD and healthy controls in the single modality. We select 29 samples as training data and 10 samples as testing data from the ADNI dataset. For the purpose of the robustness and repeatability, this process is repeated 10 times to calculate the average of the classification accuracy as the final classification accuracy. The results are demonstrated in Table 3. For complete data, the classification accuracy on individual modalities MRI, PET, and CSF are 83.50%,

Given input $z = [z^1, z^2, \ldots, z^k]$ and $s = 1/\tau$.
(1) Calculate $\widehat{u} = \|z_t\|_F$ for all $t = 1, \ldots, k$.
(2) Sort $\widehat{u}$ to obtain $u$ such that $u_{(1)} \geq \cdots \geq u_{(k)}$.
(3) Find
$$\rho = \max\left\{t \mid u_t - \frac{s}{1+t_s}\sum_{i=1}^{t} u_i \geq 0,\ t = 1, \cdots, k\right\}.$$
(4) Compute the threshold value $\varsigma = (s/(1+\rho^s))\sum_{i=1}^{\rho} u_i$.
(5) Calculate $o = soft(\widehat{u}, \varsigma)$.
(6) Compute and output $S_\tau(z)$.

Algorithm 2: Moreau Projection $S_\tau(z)$.

Initialization: Initialize the Lipschitz constant $L_t = L_{t-1}$ and set $w^{-1} = w^0$ by warm start, $\tau_0 = L_t$, $\eta \in (0, 1)$, parameter $\rho^{-1} = \rho^0 = 1$, and $k = 0$.
(1) Set $v^k = w^k + ((\rho^{k-1} - 1)/\rho^k)(w^k - w^{k-1})$.
(2) Set $\tau = \eta\tau_k$.
For $j = 0, 1, \ldots,$
Set $z = v^k - (1/\tau)\nabla g(v^k)$, compute $S_\tau(z)$.
if $F(S_\tau(z)) \leq Q(S_\tau(z), v^k)$,
set $\tau_k = \tau$, stop;
else
$\tau = \min\{\eta^{-1}\tau, L_t\}$.
End
end
(3) Set $w^{k+1} = S_{\tau_k}(z)$.
(4) Compute $\rho^{k+1} = [1 + \sqrt{(1 + 4(\rho^k)^2)}]/2$. Let $k = k + 1$.
(5) Quit if stopping condition achieves. Otherwise, go to step (1).

Algorithm 3: Accelerated proximal gradient for solving problem (10).

Table 3: Comparison of classification accuracy of incomplete and complete multimodal data.

| The type of data | Modalities | The size of data | ACC (%) |
|---|---|---|---|
| Individual modality | MRI | 39 | 83.50 |
| | PET | 39 | 77.50 |
| | CSF | 39 | 78.70 |
| Two modalities | MRI+PET | 32 | 83.30 |
| | MRI+CSF | 32 | 82.00 |
| | PET+CSF | 32 | 81.00 |
| Complete data | MRI+CSF+PET | 32 | 81.40 |
| Incomplete data | {MRI, CSF+PET, MRI+CSF+PET} | 103 | 89.30 |
| | {CSF, MRI+PET, MRI+CSF+PET} | 103 | **91.10** |
| | {PET, MRI+CSF, MRI+CSF+PET} | 103 | 87.80 |

Table 4: Comparison of performance of different multimodal classification methods.

| Method | The type of data | The size of data | ACC (%) | SEN (%) | SPE (%) |
|---|---|---|---|---|---|
| DTSVM (Cheng et al., 2015) | Incomplete data | 103 | 62.50 | 53.82 | 61.17 |
| MKL (Zhang et al., 2011) | Incomplete data | 103 | 83.09 | 81.85 | 83.63 |
| FSC (Proposed method) | Incomplete data | 103 | **91.10** | **90.00** | **91.38** |

ACC=classification accuracy, SEN=sensitivity, and SPE=specificity.

77.50%, and 78.70%, respectively. When using MRI and PET combination, the accuracy is 83.30%. When using PET and CSF combination, the accuracy is only 81.00%. The combined measurements of all three biomarkers of MRI, PET, and CSF achieves a classification accuracy of 81.40%.

Furthermore, due to the limitation of complete data, the size of incomplete data is larger than complete data. Specifically, our multimodal classification method for incomplete data achieves a classification accuracy of 91.10%, while the classification accuracy for complete data is only 81.40%. As we see from Table 3, incomplete data demonstrates much better performance than complete data in AD and healthy controls classification. The flexibility of incomplete data is better than complete data, because it takes advantage of valuable data samples and does not lead to waste data.

In Figure 2, we plot classification accuracy of complete and incomplete data corresponding to different iterations. The classification accuracies of incomplete data are better than complete data.

As mentioned in Section 2.3, B controls the sparsity of feature selection and has an important effect in the process of feature selection. In Figure 3, since different $B$ values produce different classification accuracies when MRI is used, the classification accuracy is greatly impacted by the choice of appropriate B value. In Figure 3, when $B = 30$, the mean of classification accuracy is higher than others. Therefore, we choose $B = 30$. So far, our method demonstrates much better performance on incomplete data.

In Table 4, we use incomplete multimodality data to compare the proposed method with other methods, including domain transfer support vector machine (denoted as DTSVM) [24] and multiple-kernel learning method (denoted as MKL) proposed in [2] using Lasso as feature selection.

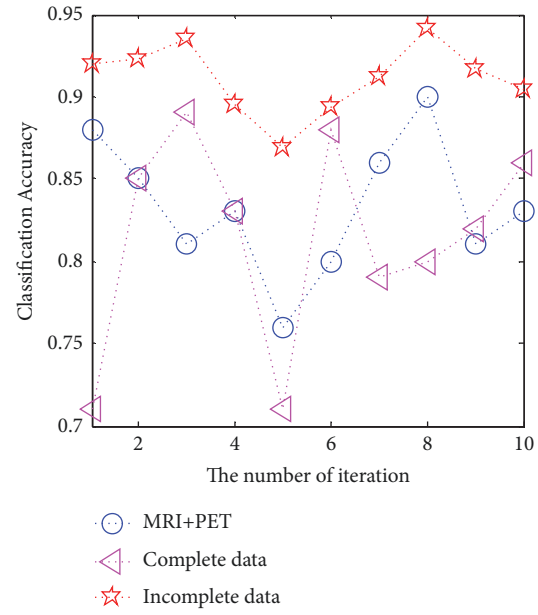


Figure 2: Classification accuracy of complete and incomplete data with respect to different iterations.

Table 4 lists the comparison of different methods for AD and HC classification.

Since our method uses nonlinear kernel explicit expansion and it maps features into high-dimensional features space, it is better in revealing high-order correlation between features. As we see in Table 4, our method outperforms the other methods for AD and HC classification. Our method achieves the classification accuracy of 91.10% with 90.00%

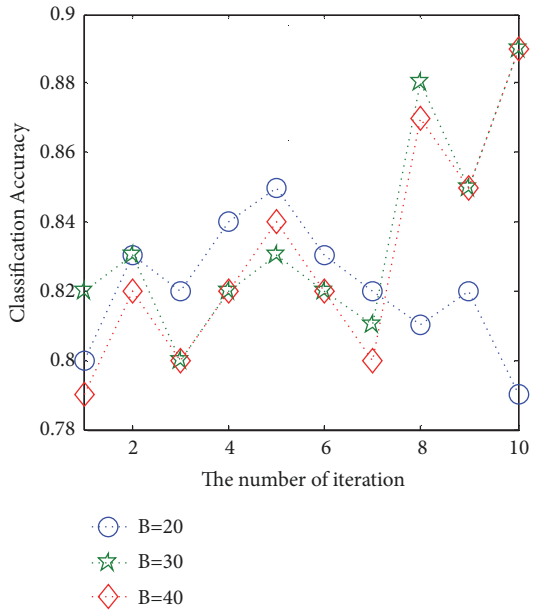


Figure 3: Performances of our method using different B parameters.

Table 5: Comparison of classification accuracy of incomplete and complete multimodality data on amazon.

| The type of data | Modalities | The size of data | ACC (%) |
|---|---|---|---|
| Individual modality | Decaf-LeNet | 300 | 80.00 |
| | Surf | 300 | 75.00 |
| | Decaf-AlexNet | 300 | 77.50 |
| Two modalities | Surf+Decaf-LeNet | 300 | 76.67 |
| | Surf+Decaf-AlexNet | 300 | 73.33 |
| | Decaf-LeNet+Decaf-AlexNet | 300 | 78.33 |
| Complete data | Surf+Decaf-LeNet+Decaf-AlexNet | 224 | 77.50 |
| Incomplete data | {Decaf-LeNet, Decaf-AlexNet+Surf, Sur+Decaf-LeNet+Decaf-AlexNet} | 824 | 94.34 |
| | {Surf, Surf+Decaf-LeNet, Sur+Decaf-LeNet+Decaf-AlexNet} | 824 | 93.40 |
| | {Decaf-LeNet, Surf+Decaf-LeNet Surf+Decaf-LeNet+Decaf-AlexNet} | 824 | 94.10 |

sensitivity and 91.38% specificity. These results further validate the efficiency of our multimodal classification method.

*3.3. Results on Office Dataset.* In this section, we evaluate our method on Office dataset which includes the following three modalities Surf, Decaf-LeNet, and Decaf-AlexNet. We start the evaluation of conducting image classification by using our method on different modalities. Then we compare classification accuracy of incomplete and complete multimodal data on amazon, dslr, and webcam, respectively. In the experiments, we expand the dimensions of feature to 180902.

We test the classification performance on different datasets. Tables 5–7 show comparison of classification accuracy of incomplete and complete multimodality data on amazon, dslr, and webcam, respectively. As we see in Tables 5–7, the classification performance on incomplete multimodality data is better compared to complete multimodality data. We want to emphasize that our method maps the low-order feature to the high-dimensional space, and this is helpful to discover the nonlinear related features. Incomplete data not only make the best use of the precious samples, but also utilize the inherent relation and knowledge of all modalities data.

## 4. Conclusion

Authors proposed a feature selection and classification method for incomplete multimodal high-dimensional data. Our algorithm produces considerably better classification performance. The flexibility of incomplete data is better than complete data. Our method takes advantage of valuable data samples and does not lead to waste data. In addition, our method focuses on extracting the relevant features from

Table 6: Comparison of classification accuracy of incomplete and complete multimodality data on dslr.

| The type of data | Modalities | The size of data | ACC (%) |
|---|---|---|---|
| Individual modality | Decaf-LeNet | 50 | 72.00 |
| | Surf | 50 | 70.00 |
| | Decaf-AlexNet | 50 | 70.00 |
| Two modalities | Surf+Decaf-LeNet | 54 | 61.90 |
| | Surf+Decaf-AlexNet | 54 | 61.90 |
| | Decaf-LeNet+Decaf-AlexNet | 54 | 71.43 |
| Complete data | Surf+Decaf-LeNet+Decaf-AlexNet | 50 | 80.00 |
| Incomplete data | {Surf, Surf+Decaf-AlexNet, Surf+Decaf-LeNet+Decaf-AlexNet} | 154 | 93.25 |
| | {Decaf-LeNet, Surf+Decaf-LeNet, Sur+Decaf-LeNet+Decaf-AlexNet} | 154 | 92.98 |
| | {Decaf-LeNet, Decaf-AlexNet+Decaf-LeNet, Sur+Decaf-LeNet+Decaf-AlexNet} | 154 | 91.23 |

Table 7: Comparison of classification accuracy of incomplete and complete multimodality data on webcam.

| The type of data | Modalities | The size of data | ACC(%) |
|---|---|---|---|
| Individual modality | Decaf-LeNet | 85 | 83.33 |
| | Surf | 85 | 83.33 |
| | Decaf-AlexNet | 85 | 76.67 |
| Two modalities | Surf+Decaf-LeNet | 70 | 88.00 |
| | Surf+Decaf-AlexNet | 70 | 85.00 |
| | Decaf-LeNet+Decaf-AlexNet | 70 | 85.00 |
| Complete data | Surf+Decaf-LeNet+Decaf-AlexNet | 80 | 93.33 |
| Incomplete data | {Surf, Surf+Decaf-AlexNet, Surf+Decaf-LeNet+Decaf-AlexNet} | 235 | 96.08 |
| | {Decaf-LeNet, Surf+Decaf-LeNet, Sur+Decaf-LeNet+Decaf-AlexNet} | 235 | 95.67 |
| | {Decaf-AlexNe, Surf+Decaf-LeNet, Sur+Decaf-LeNet+Decaf-AlexNet} | 235 | 95.93 |

incomplete multimodal data feature space. We use ADNI and Office dataset to verify the performance of our method. The results show that our method is better than other state-of-the-art methods. The limitation of this study is that it only considers simple model, and we will extend our method in more complex model in the future.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work is supported by National Natural Science Foundation of China Grant nos. 61572399, 61721002, 61532015, and 61532004; National Key Research and Development Program of China with Grant no. 2016YFB1000903; Shaanxi New Star of Science & Technology Grant no. 2013KJXX-29; New Star Team of Xian University of Posts & Telecommunications; Provincial Key Disciplines Construction Fund of General Institutions of Higher Education in Shaanxi.

## References

[1] L. Pang, S. Zhu, and C.-W. Ngo, "Deep Multimodal Learning for Affective Analysis and Retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2008–2020, 2015.

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

[3] Y. Pang, Z. Ma, Y. Yuan, X. Li, and K. Wang, "Multimodal learning for multi-label image classification," in *Proceedings of the 2011 18th IEEE International Conference on Image Processing, ICIP 2011*, pp. 1797–1800, Belgium, September 2011.

[4] J. Hu, T. Yamasaki, and K. Aizawa, "Multimodal learning for image popularity prediction on social media," in *Proceedings of*

*the 3rd IEEE International Conference on Consumer Electronics-Taiwan, ICCE-TW 2016*, pp. 1-2, May 2016.

[5] D. H. Ballard and C. Yu, "A multimodal learning interface for word acquisition," in *Proceedings of the 2003 IEEE International Conference on Accoustics, Speech, and Signal Processing*, vol. 5, pp. 784–787, Hong Kong, April 2003.

[6] M. Liu, Y. Gao, P.-T. Yap, and D. Shen, "Multi-Hypergraph Learning for Incomplete Multi-Modality Data," *IEEE Journal of Biomedical and Health Informatics*, 2017.

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

[8] X. Liu, P. Cao, J. Yang, and D. Zhao, "Linearized and Kernelized Sparse Multitask Learning for Predicting Cognitive Outcomes in Alzheimer's Disease," *Computational and Mathematical Methods in Medicine*, vol. 2018, Article ID 7429782, 13 pages, 2018.

[9] F. Li, L. Tran, K.-H. Thung, S. Ji, D. Shen, and J. Li, "A Robust Deep Model for Improved Classification of AD/MCI Patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 5, pp. 1610–1616, 2015.

[10] E. E. Bron, M. Smits, W. M. van der Flier et al., "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge," *NeuroImage*, vol. 111, pp. 562–579, 2015.

[11] J. Wang, Q. Wang, J. Peng et al., "Multi-task diagnosis for autism spectrum disorders using multi-modality features: A multi-center study," *Human Brain Mapping*, vol. 38, no. 6, pp. 3081–3097, 2017.

[12] M. Liu, J. Zhang, P.-T. Yap, and D. Shen, "View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data," *Medical Image Analysis*, vol. 36, pp. 123–134, 2017.

[13] T. Schneider, "Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[14] D. Shen and C. Davatzikos, "HAMMER: hierarchical attribute matching mechanism for elastic registration," *IEEE Transactions on Medical Imaging*, vol. 21, no. 11, pp. 1421–1439, 2002.

[15] H. Zhao, H. Liu, and Y. Fu, "Incomplete multi-modal visual data grouping," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016*, pp. 2392–2398, AAAI Press, July 2016.

[16] K. H. Thung, C.-Y. Wee, P.-T. Yap, and D. Shen, "Identification of Alzheimer's disease using incomplete multimodal dataset via matrix shrinkage and completion," in *Proceedings of the International Workshop on Machine Learning in Medical Imaging*, vol. 8184, pp. 163–170, Springer, New York, NY, USA, 2013.

[17] S.-Y. Li, Y. Jiang, and Z.-H. Zhou, "Partial multi-view clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1968–1974, July 2014.

[18] W. Shao, L. He, and P. S. Yu, "Multiple Incomplete Views Clustering via Weighted Nonnegative Matrix Factorization with $L_{2,1}$ Regularization," in *Proceedings of the Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*, vol. 9284, pp. 318–334, 2015.

[19] M. Tan, L. Wang, and I. W. Tsang, "Learning sparse SVM for feature selection on very high dimensional datasets," in *Proceedings of the 27th International Conference on Machine Learning, ICML 2010*, pp. 1047–1054, June 2010.

[20] J. E. Kelley, "The cutting-plane method for solving convex programs," *Journal of the Society For Industrial & Applied Mathematics*, vol. 8, no. 4, pp. 703–712, 1960.

[21] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems," *Pacific Journal of Optimization*, vol. 6, no. 3, pp. 615–640, 2010.

[22] N. Parikh, "Proximal Algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[23] I. Daubechies, M. Defrise, and C. de Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[24] B. Cheng, M. Liu, D. Zhang, B. C. Munsell, and D. Shen, "Domain Transfer Learning for MCI Conversion Prediction," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 10, pp. 1805–1817, 2015.