Published in final edited form as:

Proc AAAI Conf Artif Intell. 2018 February ; 2018: 4406–4413.

Multi-Layer Multi-View Classification for Alzheimer's Disease Diagnosis

Changqing Zhang[a,b], **Ehsan Adeli**[c], **Tao Zhou**[a], **Xiaobo Chen**[a], **Dinggang Shen**[a]

[a]Department of Radiology and BRIC, University of North Carolina at Chapel Hill, North Carolina, USA

[b]School of Computer Science and Technology, Tianjin University, Tianjin, China

[c]Department of Psychiatry and Behavioral Sciences & Stanford AI Lab (SAIL), Stanford University, California, USA

Abstract

In this paper, we propose a novel multi-view learning method for Alzheimer's Disease (AD) diagnosis, using neuroimaging and genetics data. Generally, there are several major challenges associated with traditional classification methods on multi-source imaging and genetics data. First, the correlation between the extracted imaging features and class labels is generally complex, which often makes the traditional linear models ineffective. Second, medical data may be collected from different sources (i.e., multiple modalities of neuroimaging data, clinical scores or genetics measurements), therefore, how to effectively exploit the complementarity among multiple views is of great importance. In this paper, we propose a *Multi-Layer Multi-View Classification (ML-MVC)* approach, which regards the multi-view input as the first layer, and constructs a latent representation to explore the complex correlation between the features and class labels. This captures the high-order complementarity among different views, as we exploit the underlying information with a low-rank tensor regularization. Intrinsically, our formulation elegantly explores the nonlinear correlation together with complementarity among different views, and thus improves the accuracy of classification. Finally, the minimization problem is solved by the Alternating Direction Method of Multipliers (ADMM). Experimental results on Alzheimers Disease Neuroimaging Initiative (ADNI) data sets validate the effectiveness of our proposed method.

Introduction

Alzheimer's Disease (AD) is a severe irreversible neurodegenerative disease, devastating lives of millions in the world (Cuingnet *et al.* 2011). Its early diagnosis, and treatment can improve the quality of life dramatically for both patients and their caregivers. There have been several studies (Weiner *et al.* 2017) in the recent years exploiting different aspects of the disease, and hence there are multiple modalities of data (e.g., Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET)) or multiple types of features available for this task (May *et al.* 1999). Generally, an important aspect of such works is

Correspondence to: Dinggang Shen.

that these features are often complementary, since they are from different measurements representing the same subject(s). On the other hand, it is evident that each individual modality alone cannot characterize the categories comprehensively, as each of them encodes different but interrelated properties of the data (Chaudhuri *et al.* 2009; Xu *et al.* 2013; Gong *et al.* 2016; Luo *et al.* 2013a; 2013b). Considering each modality (or type of features) as one view of the data, we propose to model the problem as a multi-view learning framework. Specifically, in this paper, we introduce a novel model for multi-view learning applied to the vital task of AD diagnosis.

Owing to the usefulness of exploiting the complementarity among multiple modalities or multiple types of features, multi-view learning has been the focus of intense investigation. Earlier methods usually tried to minimize the disagreement between two views based on co-training (Kumar and Daumé 2011). There are various theoretical analyses (Blum and Mitchell 1998; Chaudhuri *et al.* 2009; Wang and Zhou 2007) supporting the success and appropriateness of such approaches. Besides, multiple kernel learning (MKL) (Zien and Ong 2007; Liu *et al.* 2017) is another way of handling multiple views, which uses a predefined set of kernels for multiple views and learns an optimal combination of kernels to integrate these views. Recently, some methods are proposed to advocate for the learning of a latent common subspace across different views, typically, based on canonical correlation analysis (CCA) (Chaudhuri *et al.* 2009; Kakade and Foster 2007). For AD diagnosis, the recent works (Zhu *et al.* 2014, 2016) propose to transform the original features from different modalities to a common space by canonical correlation analysis. Although great progress has been achieved, some main limitations still exist: (1) Most existing methods usually explore linear correlation between multi-view input data and class labels, thus, they are not applicable to uncover complex correlations, compared to nonlinear methods; (2) MKL based methods map the features into a kernel space to explore the nonlinearity among the features and labels, however, simply weighting different views will not be enough for exploiting the complex correlation within each view and among different views, e.g., high-order correlations.

In this paper, we propose a novel multi-view learning approach termed as *Multi-Layer Multi-View Classification (ML-MVC)*, which focuses on addressing the above limitations in a unified framework. As shown in Fig. 1, given the data with multiple views (taking multiple modalities as example), our method aims to simultaneously explore the complex correlation between input and output, as well as the complementarity among multiple views. Based on the multiple modalities or multiple types of features of data, referred to as multi-view input, we introduce a middle layer for feature extraction with kernel technique to account for nonlinearity. Accordingly, the classification model is learned based on the mapped and refined middle-layer features (or latent representation) instead of the original ones. Furthermore, to exploit the correlation among multiple views, the kernel matrices are jointly stacked and regarded as a tensor, which is low-rank constrained to capture the complementary information from multiple views. As shown in Fig. 1, the dashed box indicates the middle layer for the latent representation corresponding to the nonlinear feature mapping and high-order correlation of multiple views. Empirical results on real data demonstrate the effectiveness of the proposed method. The optimization of our model is conducted by the Alternating Direction Method of Multipliers (ADMM) (Boyd *et al.* 2011).

The highlights of the proposed ML-MVC method and this paper are summarized as follows: (1) We simultaneously explore the complex correlation between features and classes, while exploiting the high-order correlation among multiple kernel matrices of different views. (2) The method can be regarded as a multi-layer model, where the middle layer is equipped with kernel trick to account for nonlinearity, corresponding to the latent representation. (3) Instead of performing prediction based on the kernel mapping features, our method learns the prediction model based on the refined kernel mapping features, which thoroughly explores the correlation of multiple views. (4) Based on the Alternating Direction Method of Multipliers (ADMM) (Boyd *et al.* 2011), our method is optimized efficiently and the convergence can be practically reached. (5) The experiments on multi-modalilty and multi-feature Alzheimers Disease Neuroimaging Initiative (ADNI) dataset validate the effectiveness of our method for classification on multi-view data.

Problem Formulation

Notations

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathbb{R}^D$ denote N feature vectors of N samples in the D-dimensional space, and $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ is the $D \times N$ feature matrix whose columns are the samples. For the v^{th} view, we use $\mathbf{x}_i^{(v)}$ and $\mathbf{X}^{(v)}$ to denote one sample and the feature matrix, respectively.$\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]$ is the corresponding label matrix with $\mathbf{y}_i = [y_{i1}, \cdots, y_{iC}]^\top$ being the label vector of the i^{th} sample, and $y_{ij} = 1$ if sample $\mathbf{x}_i$ belong to the j^{th} class, and $y_{ij} = 0$ otherwise, where C is the number of classes. We use the bold calligraphic font to denote a high-order tensor, e.g., $\mathcal{X}$. For clarity, the main notations used in this paper are listed in Table 1.

Background

Given the multi-view training data as $\{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(V)}; \mathbf{Y}\}$, where $\mathbf{X}^{(v)} \in \mathbb{R}^{D_v \times N}$ is the feature matrix for the v^{th} view and $\mathbf{Y} \in \mathbb{R}^{C \times N}$ is the class label matrix. Accordingly, a straightforward formulation for the multi-view learning is as follows:

$$\min_{\mathbf{W}, \mathbf{B}} \|\mathbf{Y} - \mathbf{W}\mathbf{X} - \mathbf{B}\|_F^2 + \lambda \|\mathbf{W}\|_F^2, \tag{1}$$

where $\mathbf{X} = [\mathbf{X}^{(1)}; \cdots; \mathbf{X}^{(V)}] \in \mathbb{R}^{\sum_{v=1}^{V} D_v \times N}$ concatenates different views directly, with D_V being the dimensionality of the v^{th} view. $\|\cdot\|_F$ is the Frobenius norm. $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_C]^\top \in \mathbb{R}^{C \times D}$ are learned models for C classes, where $D = \sum_{v=1}^{V} D_v$. $\mathbf{B} \in \mathbb{R}^{C \times N}$ corresponds to bias. This objective function directly extends the conventional ridge regression for multi-view data. Although simple in form and easy for optimization, there are two main issues: (1) The simple concatenation of multiple views may suffer from the curse of dimensionality and could not well explore the complementarity among different views. (2) This model focuses on linear correlation between multi-view input and class labels, which makes it improper for more complex problems. In this work, we focus on addressing these issues in a seamless framework.

Multi-Layer Multi-View Classification

To address the nonlinearity issues, we aim to design a multilayer objective function with the following general form

$$\min_{\mathbf{S},\mathbf{W}^{(v)}} \mathscr{L}(\mathbf{SZ},\mathbf{Y}) + \underbrace{\lambda_1 \mathscr{R}_1(\{\mathbf{W}^{(v)}\}_{v=1}^{V})}_{\text{Feature-mapping Regularization}} + \underbrace{\lambda_2 \mathscr{R}_2(\mathbf{S})}_{\text{Model Regularization}}, \tag{2}$$

where $\mathscr{L}(\cdot)$ is the loss function and $\lambda_1 > 0$, $\lambda_2 > 0$ are tradeoff factors for two regularization terms $\mathscr{R}_1(\cdot)$ and $\mathscr{R}_2(\cdot)$. $\mathbf{Z} = [\mathbf{Z}^{(1)}; \cdots; \mathbf{Z}^{(V)}]$ concatenates the latent representation of multiple views. Compared with the straightforward formulation in Eq. 1, rather than directly learning classification model based on the original features, we introduce a middle layer to learn the latent representation, i.e., $\mathbf{Z}^{(v)} = [\mathbf{z}_1^{(v)}, \ldots, \mathbf{z}_N^{(v)}]$, where $\mathbf{z}_i^{(v)} = \mathbf{W}^{(v)}\mathbf{x}_i^{(v)} + \mathbf{b}^{(v)}$ and the bias $\mathbf{b}^{(v)}$ can be omitted since it can be absorbed into the projection matrix$\mathbf{W}^{(v)}$ (Nie *et al.* 2010). Then, based on the latent representation, the classification model $\mathbf{S}$ is learned, forming our multi-layer model. For nonlinearity, according to the Representer Theorem (Dinuzzo and Schölkopf 2012), we have:

Theorem 1—Given any fixed matrix $\mathbf{S}$, the objective function in (2) w.r.t. $\mathbf{W}^{(v)}$ is defined over a Hilbert space $\mathscr{H}$. If (2) has a minimizer w.r.t. $\mathbf{W}^{(v)}$, it admits a linear representer theorem of the form $\mathbf{W}^{(v)} = \mathbf{P}^{(v)}\mathbf{X}^{(v)\top}$, where $\mathbf{P}^{(v)} \in \mathbb{R}^{K\times N}$ is the coefficient matrix.

According to (Dinuzzo and Schölkopf 2012), the proof of Theorem 1 is straightforward due to the decoupled property for each model $\mathbf{w}_c^{(v)}$. By introducing kernel mapping with Representer Theorem, we have $\Phi(\mathbf{X}^{(v)}) = [\phi(\mathbf{x}_1^{(v)}), \cdots, \phi(\mathbf{x}_N^{(v)})]$ with $\phi(\cdot)$ mapping the original feature $\mathbf{x}^{(v)}$ to $\phi(\mathbf{x}^{(v)})$, and accordingly, we have $\mathbf{W}^{(v)} = \mathbf{P}^{(v)}\Phi(\mathbf{X}^{(v)})^\top$. For simplicity, we use the same $\phi(\cdot)$ for different views. Therefore, based on (2) the objective function turns out to be

$$\min_{\mathbf{S},\mathbf{P}^{(v)},\mathbf{K}^{(v)}} \frac{1}{2}\|\mathbf{SZ}-\mathbf{Y}\|_F^2 + \underbrace{\frac{\gamma}{2}\sum_{v=1}^{V}\|\mathbf{P}^{(v)}\Phi(\mathbf{X}^{(v)})^\top\|_F^2}_{\mathscr{R}_1:\text{Feature-mapping Regularization}} + \underbrace{\frac{\eta}{2}\|\mathbf{S}\|_F^2}_{\mathscr{R}_2:\text{Model Regularization}}. \tag{3}$$

Since we aim to explore the correlations among different views using a tensor structure instead of directly concatenating each type of features, there is one important issue that we need to take care of. Specifically, we should note that the data from different views (e.g., different modalities of medical imaging data) often have different dimensionalities, while we have to arrange them into a single tensor with fixed dimensionality for all of them. Thanks to the advantages of the kernel technique, our objective function could naturally resolve the mentioned issues and explore the high-order correlations among multiple views as follows:

$$\begin{aligned}
&\min_{\mathbf{S},\mathbf{P}^{(v)},\widetilde{\mathbf{K}^{(v)}}} \frac{1}{2}\|\mathscr{P}_o(\mathbf{SZ}-\mathbf{Y})\|_F^2 \\
&+\underbrace{\alpha\|\widetilde{\mathscr{K}}\|_* + \frac{\beta}{2}\|\mathscr{K}-\widetilde{\mathscr{K}}\|_F^2}_{\mathscr{R}_3:\text{High-order correlation}} \\
&+\underbrace{\frac{\gamma}{2}\sum_{v=1}^{V}\|\mathbf{P}^{(v)}\widetilde{\Phi(\mathbf{X}^{(v)})^\top}\|_F^2}_{\mathscr{R}_1:\text{Feature-mapping Regularization}} + \underbrace{\frac{\eta}{2}\|\mathbf{S}\|_F^2}_{\mathscr{R}_2:\text{Model Regularization}} \\
&s.t.\ \mathscr{K}=\mathscr{T}(\mathbf{K}^{(1)},\ldots,\mathbf{K}^{(V)}),\ \widetilde{\mathscr{K}}=\mathscr{T}(\widetilde{\mathbf{K}^{(1)}},\ldots,\widetilde{\mathbf{K}^{(V)}}), \\
&\mathbf{Z}=[\mathbf{Z}^{(1)};\cdots;\mathbf{Z}^{(V)}] \text{ and } \mathbf{Z}^{(v)}=\mathbf{P}^{(v)}\widetilde{\mathbf{K}^{(v)}},
\end{aligned} \tag{4}$$

where the operator $\mathscr{T}(\cdot)$ constructs a tensor by combining multiple kernel matrices (naturally with equal dimensionality) as shown in Fig. 1. We have the kernel matrix corresponding to the v^{th} view $\mathbf{K}^{(v)} = \Phi(\mathbf{X}^{(v)})^\top\Phi(\mathbf{X}^{(v)})$ and try to seek the more reasonable $\widetilde{\mathbf{K}^{(v)}}$ to exploit the high-order correlation, i.e., $\mathbf{K}^{(v)} = \widetilde{\mathbf{K}^{(v)}} + \mathbf{E}^{(v)}$. $\mathscr{P}_O$ acts as a filter function, which forces the loss to only account for the labeled samples. Specifically, let o_i be an indicator variable showing the existence of label for sample i, i.e., $o_i = 1$ if we have the label, and a very small scalar $\varepsilon > 0$ otherwise. $\mathbf{o}$ will then be defined as the indicator vector from all indicator variables of training samples. Accordingly, we can define a diagonal matrix $\mathbf{O} = \text{diag}(\mathbf{o})$, denoted as the filter matrix, and hence $\mathscr{P}_\mathbf{O}(\mathbf{A}) = \mathbf{AO}$. Note that $\varepsilon > 0$ is a small value to strictly guarantee the unique solution of the optimization problem (see $\mathbf{P}^{(v)}$ and $\widetilde{\mathbf{K}^{(v)}}$-subproblems in the next section).

We introduce a low-rank tensor constraint to jointly explore the intrinsic correlations across multiple kernel matrices of these multiple views. Note that tensor can be seen as a generalization of the matrix concept, and hence we define the tensor nuclear norm similar to (Liu *et al.* 2013b; Tomioka *et al.* 2011), which generalizes the matrix (i.e., 2- mode or 2-order tensor) case (e.g., (Liu *et al.* 2013a)) to higher-order tensor as

$$\|\mathscr{K}\|_* = \sum_{m=1}^{M}\xi_m\|\mathbf{K}_{(m)}\|_*, \tag{5}$$

where ξ_m's are constants satisfying $\xi_m > 0$ and $\sum_{m=1}^{M}\xi_m = 1$. Without prior, we set $\xi_1 = \ldots = \xi_M = 1/M$. $\mathscr{K} \in \mathbb{R}^{I_1\times I_2\times\ldots\times I_M}$ is a M-order tensor, and $\mathbf{K}_{(m)}$ is the matrix by unfolding the tensor $\mathscr{K}$ along the m^{th} mode defined as $\text{unfold}_m(\mathscr{K}) = \mathbf{K}_{(m)} \in \mathbb{R}^{I_m\times(I_1\times\ldots\times I_{m-1}\times I_{m+1}\times\ldots\times I_M)}$ (De Lathauwer *et al.* 2000; Zhang *et al.* 2015). The nuclear norm $\|\cdot\|_*$ controls the tensor under a low-rank constraint. In essence, the nuclear norm of a tensor is a convex combination of the nuclear norms of all matrices unfolded along each mode.

Remarks—1) The model regularizer $\mathscr{R}_2(\cdot)$ for $\mathbf{S}$ can be customized for different tasks. For example, we employ Frobenius norm for AD/PD diagnosis which belongs to multiclass classification, while for multi-label classification, we can use other techniques (e.g., low-rank) to explore the correlation among different labels. 2) The matrices $\widetilde{\mathbf{K}^{(v)}}$

s are approximations of the kernel matrices $\mathbf{K}^{(v)}$s, and it is difficult to ensure $\widetilde{\mathbf{K}^{(v)}}$s to be strict kernel matrices. 3) For our model, the kernels themselves can be regarded as the entries of feature vectors within a generalized linear model (Roth 2004), i.e., $\phi(\mathbf{x}^{(v)}) = [k(\mathbf{x}^{(v)}, \mathbf{x}_1^{(v)}), \ldots, k(\mathbf{x}^{(v)}, \mathbf{x}_N^{(v)}), 1]^\top$.

To summarize, our model has the following merits: (1) Our model focuses on exploring complex correlations among the features and the class labels by introducing a middle layer equipped with kernel technique; (2) Benefiting from the kernel technique, the high-order correlation of different views is thoroughly exploited by learning the latent representation approximate to the kernel matrices of different views equipped with a low-rank tensor; (3) Both the complex input-output correlation and the high-order multi-view correlation are addressed seamlessly in a unified framework.

Optimization

Our objective function in Eq. (4) simultaneously seeks to optimize multiple projections $\mathbf{P}^{(v)}$s, matrices $\widetilde{\mathbf{K}^{(v)}}$s and model $\mathbf{S}$. Since it is not jointly convex with respect to all the variables $\mathbf{P}^{(v)}$s, $\widetilde{\mathbf{K}^{(v)}}$s and $\mathbf{S}$, we employ Alternating Direction Method of Multipliers (ADMM) (Boyd *et al.* 2011). To adopt the alternating direction minimization strategy to our problem, we need to make our objective function separable. Therefore, we introduce auxiliary variables $\mathcal{G}$, and induce the following equivalent problem to be minimized

$$\begin{aligned}
&\mathcal{L}(\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(V)}; \widetilde{\mathbf{K}^{(1)}}, \ldots, \widetilde{\mathbf{K}^{(V)}}; \mathbf{S}; \mathcal{G}; \mathcal{W}) \\
&= \frac{1}{2}\|\mathcal{P}_O(\mathbf{SZ} - \mathbf{Y})\|_F^2 \\
&+ \frac{\gamma}{2}\sum_{v=1}^{V} \|\mathbf{P}^{(v)}\widetilde{\Phi(\mathbf{X}^{(v)})^\top}\|_F^2 \\
&+ \frac{\beta}{2}\|\mathcal{K} - \widetilde{\mathcal{K}}\|_F^2 + \langle \mathcal{W}, \widetilde{\mathcal{K}} - \mathcal{G}\rangle + \frac{\mu}{2}\|\widetilde{\mathcal{K}} - \mathcal{G}\|_F^2 \\
&+ \frac{\eta}{2}\|\mathbf{S}\|_F^2 + \alpha\|\mathcal{G}\|_*,
\end{aligned} \tag{6}$$

where $\mathcal{W}$ is the Lagrange multiplier in tensor form. The operator $\langle \cdot, \cdot \rangle$ defines the tensor inner product and μ is a positive penalty scalar. For the above objective function, the sub-problems can be solved as follows:

- **Update $\mathbf{P}^{(v)}$.** The objective function with respect to updating $\mathbf{P}^{(v)}$ is

$$\min_{\mathbf{P}^{(v)}} \frac{1}{2}\|(\sum_{u \neq v} \mathbf{S}^{(u)}\mathbf{Z}^{(u)} + \mathbf{S}^{(v)}\mathbf{P}^{(V)}\widetilde{\mathbf{K}^{(v)}} - \mathbf{Y})\mathbf{O}\|_F^2 + \frac{\gamma}{2}\|\mathbf{P}^{(v)}\widetilde{\Phi(\mathbf{X}^{(v)})^\top}\|_F^2.$$

Taking the derivative with respect to $\mathbf{P}^{(v)}$ and setting it to zero, we get

$$\begin{aligned}&\mathbf{A}\mathbf{P}^{(v)}+\mathbf{P}^{(v)}\mathbf{B}=\mathbf{C}\\&\text{with }\mathbf{A}=\gamma(\mathbf{S}^{\top}\mathbf{S})^{-1},\mathbf{B}=\widetilde{\mathbf{K}^{(v)}}\mathbf{O}^{\top}\mathbf{O}\text{ and}\\&\mathbf{C}=(\mathbf{S}^{\top}\mathbf{S})^{-1}(\mathbf{S}^{(v)^{\top}}(\mathbf{Y}-\sum_{u\neq v}\mathbf{S}^{(u)}\mathbf{Z}^{(u)}))\mathbf{O}^{\top}\mathbf{O}.\end{aligned}\tag{7}$$

The above equation is a Sylvester equation (Bartels and Stewart 1972), and we have the follow proposition:

Proposition 1

The Sylvester equation (7) has a unique solution.

Proof—The Sylvester equation $\mathbf{A}\mathbf{P}^{(v)}+\mathbf{P}^{(v)}\mathbf{B}=\mathbf{C}$ has a unique solution for $\mathbf{P}^{(v)}$ exactly when there are no common eigenvalues of **A** and -**B** (Bartels and Stewart 1972). Since **B** is a positive definite matrix, all of its eigenvalues are positive: $b_i>0$. While since **A** is a positive semi-definite matrix, all of its eigenvalues are nonnegative: $a_i\geq 0$. Hence, for any eigenvalues of **A** and **B**, $a_i+b_j>0$. Accordingly, the Sylvester equation (7) has a unique solution.

- **Update $\widetilde{\mathbf{K}^{(v)}}$**. To update $\widetilde{\mathbf{K}^{(v)}}$, we should optimize the following objective function

$$\begin{aligned}&\min_{\widetilde{\mathbf{K}^{(v)}}}\frac{1}{2}\|(\sum_{u\neq v}\mathbf{S}^{(u)}\mathbf{Z}^{(u)}+\mathbf{S}^{(v)}\mathbf{P}^{(v)}\widetilde{\mathbf{K}^{(v)}}-\mathbf{Y})\mathbf{O}\|_F^2\\&+\frac{\gamma}{2}\|\mathbf{P}^{(v)}\widetilde{\Phi(\mathbf{X}^{(v)})}\|_F^2+\frac{\beta}{2}\sum_{v=1}^{V}\|\mathbf{K}^{(v)}-\widetilde{\mathbf{K}^{(v)}}\|_F^2\\&+\langle\mathbf{W}^{(v)},\widetilde{\mathbf{K}^{(v)}}-\mathbf{G}^{(v)}\rangle+\frac{\mu}{2}\|\widetilde{\mathbf{K}^{(v)}}-\mathbf{G}^{(v)}\|_F^2,\end{aligned}$$

where $\mathcal{G}=\mathcal{T}(\mathbf{G}^{(1)},\ldots,\mathbf{G}^{(V)})$, $\mathcal{W}=\mathcal{T}(\mathbf{W}^{(1)},\ldots,\mathbf{W}^{(V)})$ with $\mathbf{G}^{(v)}$ and$\mathbf{W}^{(v)}$ corresponding to the v^{th} view. Taking the derivative with respect to $\widetilde{\mathbf{K}^{(v)}}$ and setting it to zero, we get the following equation

$$\begin{aligned}&\mathbf{A}\widetilde{\mathbf{K}^{(v)}}+\widetilde{\mathbf{K}^{(v)}}\mathbf{B}=\mathbf{C}\text{with }\mathbf{A}=(\beta+\mu)(\mathbf{P}^{\top}\mathbf{S}^{\top}\mathbf{S}\mathbf{P})^{-1},\mathbf{B}=\mathbf{O}^{\top}\mathbf{O},\mathbf{C}=(\mathbf{P}^{\top}\mathbf{S}^{\top}\mathbf{S}\mathbf{P})^{-1}\\&(\mathbf{P}^{\top}\mathbf{S}^{T}\mathbf{Y}\mathbf{O}^{\top}\mathbf{O}+\beta\mathbf{K}+\mu\mathbf{G}-\mathbf{P}^{\top}\mathbf{S}^{\top}\sum_{u\neq v}\mathbf{S}^{(u)}\mathbf{Z}^{(u)}\mathbf{O}^{\top}\mathbf{O}-\gamma\mathbf{P}^{\top}\mathbf{P}-\mathbf{W}).\end{aligned}\tag{8}$$

Similar to (7), the above equation is also a Sylvester equation (Bartels and Stewart 1972) and has a unique equation.

- **Update S**. To update the model **S**, we should optimize the following objective function

$$\min_{\mathbf{S}}\frac{1}{2}\|(\mathbf{S}\mathbf{Z}-\mathbf{Y})\mathbf{O}\|_F^2+\frac{\eta}{2}\|\mathbf{S}\|_F^2.$$

Taking the derivative with respect to **S** and setting it to zero, we get the updating rule as

$$\mathbf{S} = (\mathbf{YOO}^{\top}\mathbf{Z}^{\top})(\mathbf{ZOO}^{\top}\mathbf{Z}^{\top} + \eta\mathbf{I})^{-1}. \tag{9}$$

- **Update** $\mathscr{G}$. To update the tensor auxiliary variable $\mathscr{G}$, we should optimize the following objective function

$$\min_{\mathscr{G}} \alpha\|\mathscr{G}\|_{*} + \frac{\mu}{2}\|\mathscr{G} - (\widetilde{\mathscr{K}} + \frac{1}{\mu}\mathscr{W})\|_{F}^{2}.$$

According to the tensor rank definition in Eq. (5), we have the equivalent formulation as

$$\min_{\mathbf{G}_{(m)}} \alpha \sum_{m=1}^{M} \|\mathbf{G}_{(m)}\|_{*} + \frac{\mu}{2}\sum_{m=1}^{M} \|\mathbf{G}_{(m)} - (\widetilde{\mathbf{K}_{(m)}} + \frac{1}{\mu}\mathbf{W}_{(m)})\|_{F}^{2}. \tag{10}$$

Accordingly,$\mathbf{G}_{(m)}$ could be efficiently updated with $\mathbf{G}_{(m)}^{*} = \text{prox}_{\lambda_m}^{tr}(\widetilde{\mathbf{K}_{(m)}} + \frac{1}{\mu}\mathbf{W}_{(m)})$. $\lambda_m = \alpha/\mu$ denotes the thresholds of the spectral soft-threshold operation $\text{prox}_{\lambda_m}^{tr}(\mathbf{L}) = \mathbf{U}\max(\mathbf{S} - \lambda_m, 0)\mathbf{V}^{T}$ with $\mathbf{L} = \mathbf{USV}^{T}$ being the Singular Value Decomposition (SVD) of the matrix **L**, and the max operation being taken element-wise. Intuitively, the solution is truncated according to the matrix $\widetilde{\mathbf{K}_{(m)}}$. We update all $\mathbf{G}_{(m)}$s and thus the tensor $\mathscr{G}$ is updated accordingly.

Additionally, the Lagrange multipliers can be updated as follows:

$$\mathscr{W} \leftarrow \mathscr{W} + \mu(\widetilde{\mathscr{K}} - \mathscr{G}). \tag{11}$$

For clarification, the optimization procedure is summarized in **Algorithm ??**.

Remarks

Note that, simply initializing all the block variables with zero will mislead the optimizations to trivial solutions. Based on this, we randomly initialize **S** and can obtain rather stable performance in practice.

Algorithm 1

Optimization for our ML-MVC model.

Input: Multi-view data $\{\mathbf{X}^{(1)}, ..., \mathbf{X}^{(V)}\}$, class label matrix $\mathbf{Y}$ and parameters α, β, γ, η.
Initialize: $\mathbf{P}^{(1)} = \cdots = \mathbf{P}^{(V)} = \mathbf{0}$,
$\mathcal{G} = \widetilde{\mathcal{K}} = \mathcal{W} = \mathbf{0}$, $\rho = 1.2$, $\epsilon = 10^{-6}$, $\max_\mu = 10^6$;
Initialize $\mathbf{S}$ with random values.
while *not converged* **do**
  for *each of* V *views* **do**
    Update $\mathbf{P}^{(v)}$ according to (7);
    Update $\widetilde{\mathbf{K}^{(v)}}$ according to (8);
    Update $\mathbf{Z}^{(v)}$ with $\mathbf{Z}^{(v)} = \mathbf{P}^{(v)}\widetilde{\mathbf{K}^{(v)}}$;
  end
  Update $\mathbf{S}$ according to (9);
  Update $\mathcal{G}$ according to (10);
  Update multipliers $\mathcal{W}$ according to (11);
  Update the parameter μ by $\mu = \min(\rho\mu, \max_\mu)$;
  Check the convergence conditions: $||\widetilde{\mathcal{K}} - \mathcal{G}||_\infty < \epsilon$;
end
Output: $\mathbf{S}$, $\{\widetilde{\mathbf{K}^{(v)}}\}_{v=1}^{V}$ and $\{\mathbf{P}^{(v)}\}_{v=1}^{V}$.

Complexity and Convergence

Our method is composed of four main sub-problems. For updating $\mathbf{P}^{(v)}$ and $\widetilde{\mathbf{K}^{(v)}}$, the classical algorithm for the Sylvester equation is the Bartels Stewart algorithm (Bartels and Stewart 1972), whose complexity is $O(N^3)$. The complexity of updating $\mathbf{S}$ is $O(N^2C + CNK + K^3)$, where C, K and N are the size of label set, the dimension of latent representation, and the number of samples, respectively. For updating $\mathcal{G}$ (the nuclear norm proximal operator), the complexity is $O(N^3)$. Overall, the total complexity is $O(N^2C + CNK + K^3 + N^3)$ for each iteration. Under the condition $C \ll K$ and $C \ll N$, the total complexity is basically $O(K^3 + N^3)$. It is difficult to generally prove the convergence for our algorithm. Fortunately, empirical evidence on the real data presented suggests that the proposed algorithm has very strong and stable convergence behavior even with randomly initialized $\mathbf{S}$.

Experiments

Experiment setup

In all experiments, the data are split into 10 non-overlapping folds with 9/10 and 1/10 as training and testing data, and reporting the average results and standard deviation. We conduct standard 10-fold cross-validation for each split with the hyperparameters selected from {0.01, 0.1, 1, 10, 100} for α, and {0.1, 1, 10,

100} for the other hyperparameters. Gaussian kernel is employed for each type of features, i.e., $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = \exp\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$ where $\sigma = \text{median}(\{\|x_i - x_j\|\}_{i \neq j})$. For hyperparameters of other methods, they are tuned for the best performance according to their respective published papers. We conducted experiments on two different sets of data with multiple modalities and multiple types of features. We evaluate the performance of all methods in terms of accuracy.

Compared methods

To comprehensively evaluate the proposed method, we divide the compared methods into 3 groups, i.e., methods using one, two and all three types of modalities/features. We employ a support vector classification model as the basic classifier which is from the LIBSVM toolbox [1] publicly available for the compared methods. The comparison methods include: • Single view and two-view concatenation using SVM (with Gaussian kernel); • Multiview CCA (Rupnik and Shawe-Taylor 2010) which can obtain one common space for multiple views. • Matrix Completion (Cabral *et al.* 2011) which predicts the class label with matrix completion based on a Rank Minimization criterion, with all views concatenated. • Multiclass Multiple Kernel Learning (Zien and Ong 2007) which provides a convenient and principled way based on MKL for multiclass problems. • Vector-valued Manifold Regularization based Multi-View Learning (VMR-MVL) (Minh *et al.* 2013), which is a semi-supervised multi-view classification method.

The intuitions for comparing with these methods are: (1) Single-view methods operate on each view independently using SVM, thus, they provide the evaluation of the quality for each view. Moreover, it can clarify if the multi-view treatment is essential for the overall performance or not. (2) Multiple-view methods can integrate multiple views, and here several of them are employed as comparisons to evaluate the effectiveness of our method in integrating multiple views. (3) Since nonlinearity (using kernel technique) is involved in our method, we employed kernel SVM as the basic classifier. (4) VMR-MVL is a very related method to ours, which also uses both training and testing multi-view data in the formulation.

Results on data with multiple modalities

First, we test our method on the multi-modality data set with 3 modalities, i.e., MRI, PET and Single Nucleotide Polymorphisms (SNP) genetics data. There are 360 subjects in this study, including 85 AD, 185 mild cognitive impairment (MCI), and 90 normal controls (NC) subjects, where MCI is the early stage of AD and these subjects have their MRIs scanned at first screening time.

For this study, we download ADNI 1.5T MR and PET images from the ADNI website [2]. The MR images are collected by using a variety of scanners with protocols individualized for each scanner. To ensure the quality, these MR images are corrected for spatial distortion caused by B1 field inhomogeneity and gradient nonlinearity. The PET images are collected

[1] https://www.csie.ntu.edu.tw/cjlin/libsvm/
[2] http://adni.loni.usc.edu/

by 30–60 min post Fluoro-Deoxy Glucose (FDG) injection. The operations, i.e., averaging, spatially alignment, interpolation to standard voxel size, intensity normalization, and common resolution smoothing are performed for these images. In our experiments, we extract 93 ROI-based neuroimaging features for each neuroimage (i.e., MRI or PET). In addition, for SNP data, according to the AlzGene database [3], only SNPs that belong to the top AD gene candidates are selected. Accordingly, there are 3123 SNP features used.

Results and Analysis—The performance of our method along with the compared methods are reported in Table 2, where View1, View2 and View3 correspond to MRI, SNP and PET data, respectively. The values in red, green and blue indicate the top three performers, and several observations are drawn as follows: (1) The methods using multiple views are generally superior to the methods with one single view. For example, compared with SVM using View1, SVM with both View1 and View3 achieves an improvement of about 6%, and the performance of SVM with two views are usually much better than those of SVM with single view. This confirms the necessity and effectiveness of integrating multiple views. (2) Compared with other multi-view methods, ours outperforms all, which demonstrates the effectiveness of our method for classification with multi-view data. (3) Though competitive result is achieved, with low-rank tensor constraint, the performance improvement of 4.6% is further obtained. This validates the effectiveness of exploring multiple views with low-rank high-order tensor. It is very important to note that we are classifying the the data into three classes simultaneously, as opposed to binary methods that are widely and conventionally used in neuroimaging fields. Hence, it is not fair to directly compare our results with theirs, as our method exploits a more realistic and practical case.

Results on data with multiple types of features

Here, we also conduct experiments on the resting-state functional MRI (RS-fMRI) data set with multiple types of features. In this study, there are 195 subjects, including 32 AD, 95 MCI, and 68 NC subjects. The RS-fMRI data are acquired from ADNI and parcellated into 116 regions according to the Automated Anatomical Labeling (AAL) template. The mean RS-fMRI time series of each brain region is band-pass filtered (0.015–0.15 Hz). Head motion parameters (Friston24), mean BOLD signal of white matter, and mean BOLD signal of cerebrospinal fluid are all regressed out from the RS-fMRI data to further reduce artifacts. Similar to fMRI analysis methods, we construct the functional connectivity network for each subjects, by calculating the Pearson's correlation of the mean signals from each pair of the ROIs. This constructs a full graph with correlation values and weights on the edges.

Three types of features are extracted from these graphs, and each is considered as a view in our multi-view method: (1) Nodal betweenness: The betweenness centrality is a measure of centrality in a graph, based on shortest paths. For each pair of nodes in a graph, there exists at least one shortest path between the nodes. The nodal betweenness centrality is the number of these shortest paths that pass through node *i*. (2) Nodal clustering coefficients: The coefficients are computed for each node to quantify the probability that the neighbors of node *i* are also connected to each other. (3) Nodal local efficiency: The efficiency of

[3] http://www.alzgene.org/

a network measures how efficiently information is exchanged within a network, which gives a precise quantitative analysis of the networks' information flow. The local efficiency represents the efficiency of a subgraph, which consists of all node i's neighbors.

Results and Analysis—The performance of all compared methods are listed in Table 3, where View 1, View 2 and View 3 denote nodal betweenness, nodal clustering coefficients and nodal local efficiency, respectively. According to the performance, several observations are drawn as follows: (1) Generally, SVM with multiple views is slightly superior to SVM for each single view. We note that these multiple types of features for this dataset are extracted from different aspects of one single modality, which generally leads to less complementarity among different views than that of multiple modalities. (2) Similar to the results reported in Table 2, our method outperforms all the other competitors, while much better performance is achieved when using the low-rank tensor constraint. (3) The kernelized methods are generally superior than linear ones, which demonstrates that exploring nonlinear correlation between features and class label is powerful. Overall, the results validate the effectiveness of simultaneously exploring nonlinear correlation between features and labels, and exploiting the complimentary information among multiple views as well.

Model Analysis—To well characterize our model, we provide several analytical curves for our method. Firstly, as shown in the top row of Fig. 2, in practice the convergence of our algorithm can be achieved within less than 50 iterations for both multi-modality and multi-feature cases. Secondly, according to the middle row of Fig. 2, it is observed that our low-rank tensor constraint is relatively effective and the performance is robust with respect to different tradeoff hyperparameter α in our objective function (4). Finally, the dimensionality of the latent representation is explored in the bottom row of Fig. 2, which demonstrates that our method can achieve promising results with a relatively low dimensionality.

Conclusion

We have proposed a novel multi-view learning method to take advantage of multiple views of data. By introducing kernel technique, our model well explores the complex correlations among features and class labels. Furthermore, by constraining the kernel matrices of different views to be low-rank tensor, the high-order correlation among different views is thoroughly exploited. Experiments on both multi-modality and multi-feature data clearly validated the superiority of our method over the state-of-the-arts. Although effective, there are also several directions to improve our method in the future, including incorporating weights for different views and more efficient optimization algorithm for large-scale data.

Acknowledgments

This work was supported in part by NIH grants (EB022880, AG041721, AG049371, AG042599), and National Natural Science Foundation of China (Grand No: 61602337 and 61773184).

References

Bartels, Richard H; Stewart, George W. Solution of the matrix equation AX+ XB= C. Communications of the ACM. 15 (9) 820–826. 1972.

Blum, Avrim; Mitchell, Tom. Combining labeled and unlabeled data with co-training. COLT. 92–100. 1998.

Boyd, Stephen; Parikh, Neal; Chu, Eric; Peleato, Borja; Eckstein, Jonathan. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning. 3 (1) 1–122. 2011.

Cabral, Ricardo S; Torre, Fernando; Costeira, João P; Bernardino, Alexandre. Matrix completion for multi-label image classification. NIPS. 190–198. 2011.

Chaudhuri, Kamalika; Kakade, Sham M; Livescu, Karen; Sridharan, Karthik. Multi-view clustering via canonical correlation analysis. ICML. 129–136. 2009.

Cuingnet, Rémi; Gerardin, Emilie; Tessieras, Jérôme; Auzias, Guillaume; Lehéricy, Stéphane; Habert, Marie-Odile; Chupin, Marie; Benali, Habib; Colliot, Olivier; , et al. Alzheimer's Disease Neuroimaging Initiative. Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. NeuroImage. 56 (2) 766–781. 2011. [PubMed: 20542124]

De Lathauwer, Lieven; De Moor, Bart; Vandewalle, Joos. On the best rank-1 and rank-(r_1, r_2, …, r_n) approximation of higher-order tensors. SIAM journal on Matrix Analysis and Applications. 21 (4) 1324–1342. 2000.

Dinuzzo, Francesco; Schölkopf, Bernhard. The representer theorem for hilbert spaces: a necessary and sufficient condition. NIPS. 189–196. 2012.

Gong, Chen; Tao, Dacheng; Maybank, Stephen J; Liu, Wei; Kang, Guoliang; Yang, Jie. Multi-modal curriculum learning for semi-supervised image classification. IEEE T-IP. 25 (7) 3249–3260. 2016.

Kakade, Sham M; Foster, Dean P. Multi-view regression via canonical correlation analysis. COLT. 82–96. 2007.

Kumar, Abhishek; Daumé, Hal. A co-training approach for multi-view spectral clustering. ICML. 393–400. 2011.

Liu, Guangcan; Lin, Zhouchen; Yan, Shuicheng; Sun, Ju; Yu, Yong; Ma, Yi. Robust recovery of subspace structures by low-rank representation. IEEE T-PAMI. 35 (1) 171–184. 2013.

Liu, Ji; Musialski, Przemyslaw; Wonka, Peter; Ye, Jieping. Tensor completion for estimating missing values in visual data. IEEE T-PAMI. 35 (1) 208–220. 2013.

Liu, Xinwang; Li, Miaomiao; Wang, Lei; Dou, Yong; Yin, Jianping; Zhu, En. Multiple kernel k-means with incomplete kernels. AAAI. 2259–2265. 2017.

Luo, Yong; Tao, Dacheng; Xu, Chang; Li, Dongchen; Xu, Chao. Vector-valued multi-view semi-supervsed learning for multi-label image classification. AAAI. 647–653. 2013.

Luo, Yong; Tao, Dacheng; Xu, Chang; Xu, Chao. Multiview vector-valued manifold regularization for multilabel image classification. IEEE T-NNLS. 709–722. 2013.

May A, Ashburner J, Büchel C, McGonigle DJ, Friston KJ, Frackowiak RSJ, Goadsby PJ. Correlation between structural and functional changes in brain in an idiopathic headache syndrome. Nature medicine. 5 (7) 836–838. 1999.

Minh, Ha Quang; Bazzani, Loris; Murino, Vittorio. A unifying framework for vector-valued manifold regularization and multi-view learning. ICML. (2) 100–108. 2013.

Nie, Feiping; Huang, Heng; Cai, Xiao; Ding, Chris H. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. NIPS. 1813–1821. 2010.

Roth, Volker. The generalized lasso. IEEE T-NN. 15 (1) 16–28. 2004.

Rupnik, Jan; Shawe-Taylor, John. Multi-view canonical correlation analysis. SiKDD. 1–4. 2010.

Tomioka, Ryota; Suzuki, Taiji; Hayashi, Kohei; Kashima, Hisashi. Statistical performance of convex tensor decomposition. NIPS. 972–980. 2011.

Wang, Wei; Zhou, Zhi-Hua. Analyzing co-training style algorithms. ECML. 454–465. 2007.

Weiner, Michael W; Veitch, Dallas P; Aisen, Paul S; Beckett, Laurel A; Cairns, Nigel J; Green, Robert C; Harvey, Danielle; Jack, Clifford R; Jagust, William; Morris, John C; , et al. Recent publications

from the alzheimer's disease neuroimaging initiative: Reviewing progress toward improved ad clinical trials. Alzheimer's & Dementia. 2017.

Xu, Chang; Tao, Dacheng; Xu, Chao. A survey on multi-view learning. 2013.

Zhang, Changqing; Fu, Huazhu; Liu, Si; Liu, Guangcan; Cao, Xiaochun. Low-rank tensor constrained multiview subspace clustering. ICCV. 1582–1590. 2015.

Zhu, Xiaofeng; Suk, Heung-Il; Shen, Dinggang. MICCAI. Springer; 2014. Multi-modality canonical feature selection for alzheimer's disease diagnosis; 162–169.

Zhu, Xiaofeng; Suk, Heung-Il; Lee, Seong-Whan; Shen, Dinggang. Canonical feature selection for joint regression and multi-class identification in alzheimer's disease diagnosis. Brain imaging and behavior. 10 (3) 818–828. 2016. [PubMed: 26254746]

Zien, Alexander; Ong, Cheng Soon. Multiclass multiple kernel learning. ICML. 1191–1198. 2007.

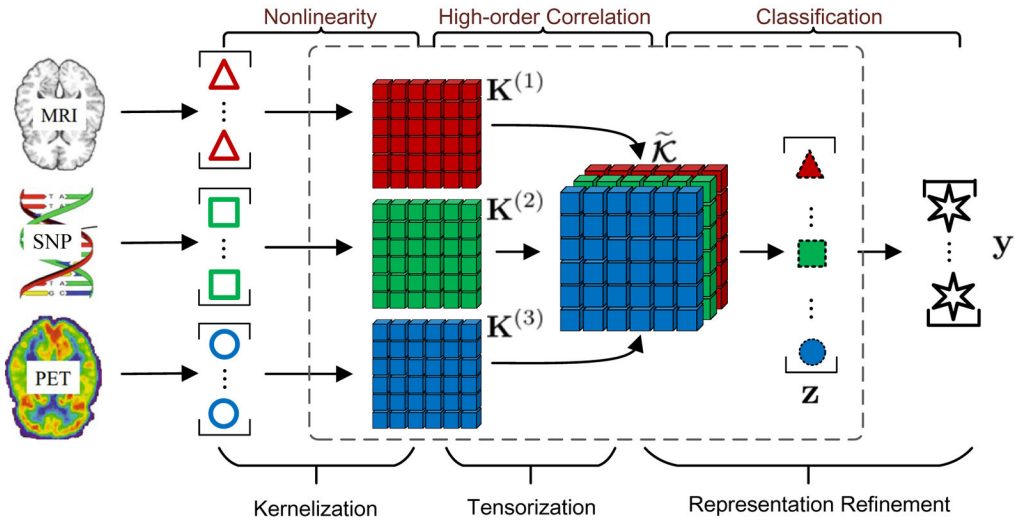

Figure 1.

Illustration of the multi-layer multi-view learning framework for AD prediction. Our model jointly exploits the nonlinear feature mapping, explores high-order correlation of multiple views and learns classification model.

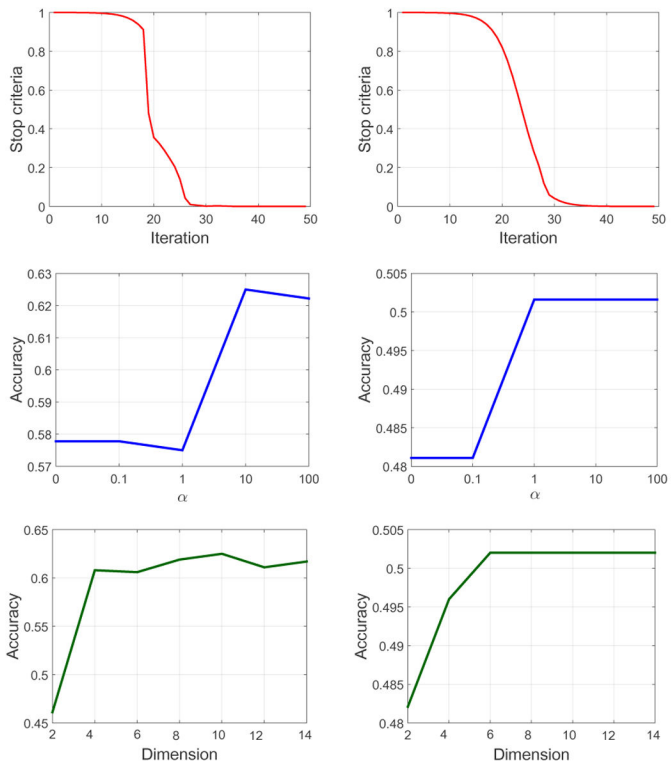

Figure 2.

Model analysis on multi-modality (left column) and multi-feature (right column) data. The rows from top to bottom correspond to convergence curves, performance with respect to α and K, respectively.

Table 1

Table of main notations used in the paper.

Model Specification	
Notation	**Meaning**
$\mathbf{X}^{(v)} \in \mathbb{R}^{D_v \times N}$	feature matrix of the v^{th} view
$\mathbf{Y} \in \mathbb{R}^{C \times N}$	label matrix
$\mathbf{Z}^{(v)} \in \mathbb{R}^{K \times N}$	latent representation for the v^{th} view
$\mathbf{P}^{(v)} \in \mathbb{R}^{K \times N}$	projection corresponding to the v^{th} view
$\mathbf{S} \in \mathbb{R}^{C \times VK}$	classification model
$\mathscr{K} \in \mathbb{R}^{K \times N \times V}$	tensor of kernel matrices
$\mathscr{G} \in \mathbb{R}^{K \times N \times V}$	auxiliary variables in tensor form
$\mathscr{W} \in \mathbb{R}^{K \times N \times V}$	Lagrange multiplier in tensor form
$\mathbf{K}_{(m)}/\mathbf{G}_{(m)}$	unfolded matrix of tensor $\mathscr{K}/\mathscr{G}$
$\mu > 0$	penalty hyperparameter for constraints

Table 2

Accuracy on multi-modality data.

No.	Configuration	Accuracy
1	View1	0.517 ± 0.064
2	View2	0.508 ± 0.107
3	View3	0.542 ± 0.101
4	View1+View2	0.531 ± 0.061
5	View1+View3	0.575 ± 0.073
6	View2+View3	0.556 ± 0.109
7	AllViewConcatenate	0.608 ± 0.075
8	Multiview CCA	0.581 ± 0.072
9	Matrix Completion	0.514 ± 0.092
10	MultiClass MKL	0.582 ± 0.091
11	VMR-MVL	0.579 ± 0.081
12	Ours ($\alpha = 0$)	0.579 ± 0.050
13	Ours + Tensor ($\alpha = 10$)	0.625 ± 0.069

Table 3

Accuracy on multi-feature data.

No.	Configuration	Accuracy
1	View1	0.461 ± 0.128
2	View2	0.466 ± 0.125
3	View3	0.471 ± 0.115
4	View1+View2	0.477 ± 0.117
5	View1+View3	0.462 ± 0.096
6	View2+View3	0.477 ± 0.122
7	AllViewConcatenate	0.467 ± 0.119
8	Multiview CCA	0.482 ± 0.106
9	Matrix Completion	0.410 ± 0.115
10	MultiClass MKL	0.451± 0.113
11	VMR-MVL	0.481 ± 0.131
12	Ours ($\alpha = 0$)	0.481 ± 0.108
13	Ours + Tensor ($\alpha = 100$)	0.502 ± 0.122