Deep Correlated Predictive Subspace Learning for Incomplete Multi-View Semi-Supervised Classification

Zhe Xue[1], **Junping Du**[1*], **Dawei Du**[2], **Wenqi Ren**[3] and **Siwei Lyu**[2]
[1] Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] Computer Science Department, University at Albany, SUNY, Albany, NY, 12222, USA
[3] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
xuezhe@bupt.edu.cn, junpingdu@126.com, {ddw, slyu}@albany.edu, renwenqi@iie.ac.cn

Abstract

Incomplete view information often results in failure cases of the conventional multi-view methods. To address this problem, we propose a Deep Correlated Predictive Subspace Learning (DCPSL) method for incomplete multi-view semi-supervised classification. Specifically, we integrate semi-supervised deep matrix factorization, correlated subspace learning, and multi-view label prediction into a unified framework to jointly learn the deep correlated predictive subspace and multi-view shared and private label predictors. DCPSL is able to learn proper subspace representation that is suitable for class label prediction, which can further improve the performance of classification. Extensive experimental results on various practical datasets demonstrate that the proposed method performs favorably against the state-of-the-art methods.

1 Introduction

Data can be represented by multiple views in many real-world applications. For instance, an image can be described by different types of features such as LBP, SIFT, and color histogram. Different views usually provide complementary information and leveraging multi-view data is beneficial to improving the overall learning performance. However, some views may not contain complete information in real tasks (*e.g.*, some Web pages contain both image and text information while some may only contain text information). The incomplete data may lead to performance degradation or even failure of the conventional multi-view methods.

Up to now, several approaches have been developed for the incomplete multi-view unsupervised or semi-supervised learning. For two-view incomplete data clustering, some methods [Li *et al.*, 2014; Zhao *et al.*, 2016; Xu *et al.*, 2018] adopt matrix factorization model to seek a common latent subspace, in which the samples of different views are constrained to have the same representation. To conduct clustering with more than two views, MIC [Shao *et al.*, 2015] adopts a co-regularized method to push the learned subspace of all the views into a common consensus. DAIMC [Hu and Chen, 2018] constructs a consistent data representation with $l_{2,1}$-norm to reduce the influence of missing views. A spectral clustering based method is proposed [Wen *et al.*, 2019] to learn the consensus representation and the similarity graphs for all views. In addition to unsupervised learning, semi-supervised learning methods for incomplete multi-view data are developed recently. SLIM [Yang *et al.*, 2018] learns the classifiers by leveraging the intrinsic view consistency and extrinsic unlabeled information, which is further used to predict the class of unlabeled samples. To predict multiple labels from incomplete views, iMVWL [Tan *et al.*, 2018] learns a consistent subspace by considering the label correlations which can reinforce the prediction results.

While some successful incomplete multi-view learning methods have been proposed, there remains room for practical improvements and additional theoretical understanding. First, most of the incomplete multi-view learning methods are based on shallow models, which cannot learn robust representation for data with complex distributions. Second, since some samples may lack of feature descriptions in some views, the data correlation is an important clue to make the data complement each other and improve the discriminating power of data representation. Third, different views have shared and independent information, omitting such shared and private nature of multi-view data would limit the performance of classification.

To solve the aforementioned problems, we propose a *Deep Correlated Predictive Subspace Learning* (DCPSL) method for incomplete multi-view semi-supervised classification. We integrate semi-supervised deep matrix factorization, correlated subspace learning and multi-view class label prediction into a unified objective function to jointly learn the *deep correlated predictive subspace* and the *shared and private label predictors*. The proposed DCPSL is able to learn the proper subspace representation that is suitable for class label prediction, which can further improve the classification performance. The discriminating power of the learned subspace representation is guaranteed from three aspects. First, the label information is incorporated into deep matrix factorization model which makes the learned representation achieve obvi-

*Corresponding Author

ous category structure. Second, data correlation is utilized to make the incomplete multi-view data complement each other, which further enhances the effectiveness of the subspace. Third, we introduce the shared and private label predictors for classification so that the complementarity information of the learned multi-view subspace can be leveraged and more accurate class labels can be predicted. The experiments conducted on several multi-view datasets verify the effectiveness of our method. The main contributions of this work are summarized as follows:

- We extract abstract and high-level multi-view representation by semi-supervised deep matrix factorization model. By encoding the label information into the representation, we can significantly improve the discriminating power of the subspace representation.
- We learn low-rank subspaces from deep matrix factorization model, where data correlation can be effectively extracted. Then the data correlation is further used to enhance the effectiveness of the learned subspace representation.
- We propose to divide the label predictor into shared and private parts, which can effectively utilize multi-view complementary information and improve the performance of class label prediction.

2 Proposed Method

2.1 Preliminaries

For complete multi-view data with V views, sample x_i is composed of V features $\{x_i^{(1)}, x_i^{(2)}, ..., x_i^{(V)}\}$, where $x_i^{(v)} \in R^{d_v}$ and d_v is the feature dimension of the v-th view. Due to the incomplete data problem, the features of some views may be missing for x_i. Let τ denote the index of views that x_i appears in, then x_i is compose of $\{x_i^{(j)}\}_{j\in\tau}$. Given n samples, incomplete multi-view data can be denoted by a set of matrices $\{X^{(v)} \in \mathbb{R}^{d_v\times n_v}\}_{v=1}^{V}$, where $X^{(v)}$ consists of n_v samples appear in the v-th view. We have $n_v \leq n$ due to the multi-view data missing problem. In semi-supervised setting, some samples have labels while others do not. We use $Y = [y_1, y_2, ..., y_l] \in \{0, 1\}^{c\times l}$ to represent the label matrix, where l is the number of labeled samples and c is the class number. For a labeled sample x_j, if it belongs to the i-th class, $Y_{ij} = 1$; otherwise $Y_{ij} = 0$. Our objective is to predict the class labels of the unlabeled samples based on the incomplete multi-view feature matrices $\{X^{(v)}\}_{v=1}^{V}$ and the label matrix Y.

To capture complex data distributions, deep Semi-NMF [Trigeorgis *et al.*, 2014] is proposed to learn the inherent attributes and higher-level feature representation. It decomposes data matrix X into m layers as

$$\begin{aligned} X &\approx Z_1 H_1^+, \\ X &\approx Z_1 Z_2 H_2^+, \\ &\vdots \\ X &\approx Z_1 Z_2 ... Z_m H_m^+, \end{aligned} \tag{1}$$

where $Z_i \in \mathbb{R}^{k_{i-1}\times k_i}$ and $H_m \in \mathbb{R}^{k_m\times n}$ are the loading matrix and coefficient matrix, respectively. k_i is the dimension of the i-th layer, and $(a)^+ = \max(0, a)$ corresponds to the hinge operation.

Subspace clustering [Vidal, 2011; Elhamifar and Vidal, 2013] assumes that data are drawn from different subspaces, and its goal is to cluster data according to their underlying subspaces. The low-rank representation (LRR) method [Liu *et al.*, 2013] learns a low-rank subspace representation and achieves promising clustering performance. Given a data matrix $X \in \mathbb{R}^{d\times n}$, LRR solves the self-representation problem by finding the lowest rank representation of all data as

$$\min_S ||S||_*, \quad s.t.\ X = XS + E, \tag{2}$$

where $S \in \mathbb{R}^{n\times n}$ is the learned low-rank subspace representation of data X, E is the error matrix, $||\cdot||_*$ denotes the nuclear norm of a matrix, which equals to the sum of its singular values. The learned subspace S is capable of capturing the correlations of data samples.

2.2 Problem Formulation

There are two main problems in semi-supervised incomplete multi-view learning: i) how to obtain the proper multi-view data representation for classification, and ii) how to predict the class label of unlabeled samples. To this end, we first introduce deep correlated subspace learning to properly represent multi-view data, and then introduce multi-view shared and private label prediction. Finally, the ultimate objective function integrates the two subproblems to obtain both proper subspace representation and accurate classification results.

Deep Correlated Subspace Learning

To learn the proper multi-view representation for semi-supervised classification, three factors should be considered. First, the class label information should be used to guide the representation learning. Second, data correlations contain abundant descriptions of relations between data, which can be used to enhance the effectiveness of representation. Third, data samples may be produced by complex data distributions. Compared to shallow models, deep models can better learn the inherent attributes and higher-level data representation. In the light of the above points, we introduce the following deep correlated subspace learning objective to obtain data representation

$$\begin{aligned} &\min J_1(Z_i^{(v)}, H_m^{(v)}, S^{(v)}) \\ &= \sum_{v=1}^{V} \{||X^{(v)} - Z_1^{(v)} Z_2^{(v)} ... Z_m^{(v)} H_m^{(v)} P^{(v)}||_F^2 \\ &+ ||H_m^{(v)} P^{(v)} - H_m^{(v)} P^{(v)} S^{(v)}||_F^2 + \alpha ||S^{(v)}||_* \\ &s.t.\ H_m^{(v)} \geq 0, S^{(v)} \geq 0 \end{aligned} \tag{3}$$

where m is the number of layers. $Z_i^{(v)}$ and $H_m^{(v)}$ are the learned loading matrix and coefficient matrix for the v-th view, respectively. Inspired by [Liu *et al.*, 2012], the label constraint matrix $P^{(v)} \in R^{(n_v - l_v + c)\times n_v}$ is used to guide the learning of $H_m^{(v)}$, where c is the class number and l_v is number of labeled samples in the v-th view. For example, given samples $\{x_1^{(v)}, x_2^{(v)}, ..., x_{n_v}^{(v)}\}$ from the v-th view, $x_1^{(v)}, x_2^{(v)}$

belong to class 1, $x_3^{(v)}, x_4^{(v)}$ belong to class 2, and the other samples are unlabeled, then $P^{(v)}$ is defined as

$$P^{(v)} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & I_{n_v-4} \end{pmatrix} \tag{4}$$

where $I_{n_v-4} \in R^{(n_v-4)\times(n_v-4)}$ is an identity matrix. The new data representation becomes $H_m^{(v)}P^{(v)}$ where the label information can be well preserved. It is easy to check that if $x_i^{(v)}$ and $x_j^{(v)}$ have the same label, then $(H_m^{(v)}P^{(v)})_{(:,i)} = (H_m^{(v)}P^{(v)})_{(:,j)}$.

Although $H_m^{(v)}P^{(v)}$ achieves high-level feature representation and preserves the label information, it does not take advantage of the data correlations, which may affect its discriminating power. We further learn the low-rank subspace $S^{(v)} \in \mathbb{R}^{n_v\times n_v}$ from $H_m^{(v)}P^{(v)}$ to reveal the intrinsic data correlations. Through self-representation learning, each element in $S_i^{(v)}$ well captures the relations between two samples. Then we can use $H_m^{(v)}P^{(v)}S_i^{(v)}$ as the enhanced data subspace representation to achieve better prediction.

Multi-view Shared and Private Label Prediction

Given the enhanced data subspace presentation $H_m^{(v)}P^{(v)}S_i^{(v)} \in R^{k_m\times n_v}$ and label matrix $Y \in R^{c\times l}$, the class labels of unlabeled samples should be predicted. We adopt linear regression model for label prediction due to its convenience and effectiveness. It should be noted that different views share some common information while retaining some independent information. Only by considering the shared and private nature of multi-view data can we make accurate predictions of data labels. Hence, we propose multi-view shared and private label prediction as follows

$$\begin{aligned} &\min J_2(W_s, W_p^{(v)}, F) \\ &= \sum_{v=1}^{V} \{||(W_s + W_p^{(v)})H_m^{(v)}P^{(v)}S^{(v)} - FQ^{(v)}||_F^2 \\ &\quad +\beta_1||W_p^{(v)}||_1 + \beta_2||W_s||_F^2\} \\ &s.t.\ F_{\pi_l} = Y \end{aligned} \tag{5}$$

where $F \in R^{c\times n}$ is the predicted label matrix, π_l is the index set of the labeled samples. $F_{\pi_l} = Y$ restricts the prediction on labeled data as same as the ground truth. The label predictor for the v-th view can be denoted by $W^{(v)} = W_s + W_p^{(v)}$, where $W_s \in R^{c\times k_m}$ is the shared part that used for all the views and $W_p^{(v)} \in R^{c\times k_m}$ is the private part that used only for the v-th view. We impose l_1-norm regularization on $W_p^{(v)}$ to make it adaptively capture the private components of the v-th view. Different views adopt different label predictors so that multi-view complementary information can be leveraged to generate more accurate results.

To cope with the incomplete multi-view data, the alignment matrix $Q^{(v)} \in R^{n\times n_v}$ is introduced which represents the correspondence between n_v samples appear in the v-th view and all the n samples. For example, if the three samples from the v-th view, *i.e.*, $H_m^{(v)}P^{(v)}S_i^{(v)} \in R^{k_m\times 3}$, correspond to the 2nd, 3rd and 5-th sample in F, respectively, then $Q^{(v)} \in R^{n\times 3}$ is constructed as

$$Q^{(v)} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ & \cdots & \end{pmatrix} \tag{6}$$

Our method can flexibly deal with the incomplete multi-view data, so that all the samples from different views can be leveraged for label prediction effectively. After obtaining F, any unlabeled sample can be assigned to a class, *e.g.*, the class of the unlabeled sample x_i is determined by $\arg\max_j F_{ji}$.

Ultimate Objective Function

The ultimate objective function is formed by combining the above sub-problems. By jointly conduct deep correlated subspace learning and multi-view shared and private label predictor learning, the ultimate objective function is proposed, which is

$$\begin{aligned} &\min J(Z_i^{(v)}, H_m^{(v)}, S^{(v)}, W_s, W_p^{(v)}, F) \\ &= \sum_{v=1}^{V} \{||X^{(v)} - Z_1^{(v)}Z_2^{(v)}...Z_m^{(v)}H_m^{(v)}P^{(v)}||_F^2 \\ &\quad +||H_m^{(v)}P^{(v)} - H_m^{(v)}P^{(v)}S^{(v)}||_F^2 \\ &\quad +\lambda||(W_s + W_p^{(v)})H_m^{(v)}P^{(v)}S^{(v)} - FQ^{(v)}||_F^2 \\ &\quad +\Phi(S^{(v)}, W_p^{(v)}, W_s)\} \\ &s.t.\ H_m^{(v)} \geq 0, S^{(v)} \geq 0, F_{\pi_l} = Y, \end{aligned} \tag{7}$$

where,

$$\Phi(S^{(v)}, W_p^{(v)}, W_s) = \alpha||S^{(v)}||_* + \beta_1||W_p^{(v)}||_1 + \beta_2||W_s||_F^2. \tag{8}$$

We introduce λ to control the weight of label prediction term. By solving the ultimate objective function J, our method can jointly learn subspace representation, label predictors and the predicted label matrix F, so that more effective data representation and accurate classification results can be obtained.

3 Optimization

In this section, we provide more details on solving the ultimate objective function (7) by the iterative block coordinate descent algorithm, where only one variable is updated while the others are fixed in each iteration. Specifically, we pre-train each layer and obtain initial $Z_i^{(v)}$ and $H_m^{(v)}$. Then, each variable such as $Z_i^{(v)}$, $H_m^{(v)}$, $S^{(v)}$, W_s, $W_p^{(v)}$ and F is updated. The details of pre-training and derivations of update rules are presented as follows. Algorithm 2 presents the algorithm for solving problem (7).

3.1 Pre-training of Deep Semi-NMF

The latent factors $Z_i^{(v)}$ and $H_i^{(v)}$ in Deep Semi-NMF should be pre-trained before solving the other variables. For each view $v = 1, ..., n_v$, the first layer is learned by $X^{(v)} \approx Z_1^{(v)}H_1^{(v)}$, where $Z_1^{(v)} \in \mathbb{R}^{d_v\times k_1}$ and $H_1^{(v)} \in \mathbb{R}^{k_1\times n}$. After that, coefficient matrix $H_1^{(v)}$ is further decomposed by $H_1^{(v)} \approx Z_2^{(v)}H_2^{(v)}$, where $Z_2^{(v)} \in \mathbb{R}^{k_1\times k_2}$ and $H_2^{(v)} \in \mathbb{R}^{k_2\times n}$. This process is continued until all the layers are pre-trained, *i.e.*, $H_2^{(v)} \approx Z_3^{(v)}H_3^{(v)}, \cdots, H_{m-1}^{(v)} \approx Z_m^{(v)}H_m^{(v)}$.

3.2 Updating $Z_i^{(v)}$

To solve $Z_i^{(v)}$, let the partial derivative $\partial(J)/\partial(Z_i^{(v)}) = 0$, and then the update rule can be obtained by $Z_i^{(v)} \leftarrow \Psi^{\dagger} X^{(v)} \Omega^{\dagger}$, where $\Psi = Z_1^{(v)} \cdots Z_{i-1}^{(v)}$ and $\Omega = Z_{i+1}^{(v)} \cdots Z_m^{(v)} H_m^{(v)} P^{(v)}$. $(\cdot)^{\dagger}$ is the Moore-Penrose pseudo-inverse operator and $A^{\dagger} = (A^T A)^{-1} A^T$.

3.3 Updating $H_m^{(v)}$

To solve $H_m^{(v)}$, we take the related parts from J and obtain

$$\begin{aligned} &J(H_m^{(v)}) = ||X^{(v)} - Z_1^{(v)} Z_2^{(v)} ... Z_m^{(v)} H_m^{(v)} P^{(v)}||_F^2 \\ &+||H_m^{(v)} P^{(v)} - H_m^{(v)} P^{(v)} S^{(v)}||_F^2 \\ &+\lambda||W^{(v)} H_m^{(v)} P^{(v)} S^{(v)} - F Q^{(v)}||_F^2 \end{aligned} \tag{9}$$

where we introduce $W^{(v)} = W_s + W_p^{(v)}$. Using the standard Semi-NMF optimization method, we can derive the following multiplicative update rule,

$$(H_m^{(v)})_{ij} \leftarrow (H_m^{(v)})_{ij} \sqrt{\frac{(\Gamma_1)_{ij}}{(\Gamma_2)_{ij}}} \tag{10}$$

where

$$\begin{aligned} &\Gamma_1 = [\Psi^T X^{(v)} P^{(v)^T}]^{pos} + [2 H_m^{(v)} P^{(v)} S^{(v)} P^{(v)^T}]^{pos} \\ &+[\lambda W^{(v)^T} F Q^{(v)} S^{(v)^T} P^{(v)^T}]^{pos} \\ &+[\Psi^T \Psi H_m^{(v)} P^{(v)} P^{(v)^T}]^{neg} + [H_m^{(v)} P^{(v)} P^{(v)^T}]^{neg} \\ &+[H_m^{(v)} P^{(v)} S^{(v)} S^{(v)^T} P^{(v)^T}]^{neg} \\ &+[\lambda W^{(v)^T} W^{(v)} H_m^{(v)} P^{(v)} S^{(v)} S^{(v)^T} P^{(v)^T}]^{neg}, \end{aligned} \tag{11}$$

and

$$\begin{aligned} &\Gamma_2 = [\Psi^T X^{(v)} P^{(v)^T}]^{neg} + [2 H_m^{(v)} P^{(v)} S^{(v)} P^{(v)^T}]^{neg} \\ &+[\lambda W^{(v)^T} F Q^{(v)} S^{(v)^T} P^{(v)^T}]^{neg} \\ &+[\Psi^T \Psi H_m^{(v)} P^{(v)} P^{(v)^T}]^{pos} + [H_m^{(v)} P^{(v)} P^{(v)^T}]^{pos} \\ &+[H_m^{(v)} P^{(v)} S^{(v)} S^{(v)^T} P^{(v)^T}]^{pos} \\ &+[\lambda W^{(v)^T} W^{(v)} H_m^{(v)} P^{(v)} S^{(v)} S^{(v)^T} P^{(v)^T}]^{pos}. \end{aligned} \tag{12}$$

The notation of $[\cdot]^{pos}$ and $[\cdot]^{neg}$ in (11) and (12) denote the operation that replaces all the negative and positive elements in the matrix by 0, respectively, and can be defined as

$$[A]_{ij}^{\text{pos}} = \frac{|A_{ij}| + A_{ij}}{2}, \ [A]_{ij}^{\text{neg}} = \frac{|A_{ij}| - A_{ij}}{2}.$$

3.4 Updating $S^{(v)}$

Keeping the parts that are related to $S_i^{(v)}$ from ultimate objective function J, we obtain the following problem

$$\begin{aligned} &\min_{S^{(v)}} ||H_m^{(v)} P^{(v)} - H_m^{(v)} P^{(v)} S^{(v)}||_F^2 + \alpha ||S^{(v)}||_* \\ &+\lambda ||W^{(v)} H_m^{(v)} P^{(v)} S^{(v)} - F Q^{(v)}||_F^2 \\ &s.t.\ S^{(v)} \geq 0, \end{aligned} \tag{13}$$

which can be solved by alternating direction method of multipliers (ADMM) [Liu *et al.*, 2013]. The augmented Lagrangian function of problem (13) is,

$$\begin{aligned} &\mathcal{L}(S^{(v)}, B_1, B_2, B_3, B_4) = ||H_m^{(v)} P^{(v)} - B_1||_F^2 + \alpha ||B_2||_* \\ &+\lambda ||B_3 - F Q^{(v)}||_F^2 + l_{R^+}(B_4) + \tfrac{\mu}{2} ||B_4 - S^{(v)} - R_4||_F^2 \\ &+\tfrac{\mu}{2} ||B_1 - H_m^{(v)} P^{(v)} S^{(v)} - R_1||_F^2 + \tfrac{\mu}{2} ||B_2 - S^{(v)} - R_2||_F^2 \\ &+\tfrac{\mu}{2} ||B_3 - W^{(v)} H_m^{(v)} P^{(v)} S^{(v)} - R_3||_F^2 \end{aligned} \tag{14}$$

where $\{B_i\}_{i=1}^{n_v}$ are the auxiliary variables, $\{R_i\}_{i=1}^{n_v}$ are the Lagrange multipliers. $l_{R^+}(\cdot)$ is defined as

$$l_{R^+}(A) = \begin{cases} 0 & \text{if } A \geq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

We set the partial derivative $\partial(\mathcal{L}(S^{(v)}, B_1, B_2, B_3, B_4))/\partial(S^{(v)}) = 0$ to solve $S^{(v)}$ as

$$\begin{aligned} S^{(v)} \leftarrow &(\xi^{(v)^T} \xi^{(v)} + \xi^{(v)^T} {W^{(v)}}^T W^{(v)} \xi^{(v)} + 2I)^{-1} \\ &(\xi^{(v)^T} \eta_1 + \eta_2 + \xi^{(v)^T} {W^{(v)}}^T \eta_3 + \eta_4) \end{aligned} \tag{15}$$

where $\eta_i = B_i - R_i$, $\xi^{(v)} = H_m^{(v)} P^{(v)}$ and I is the identity matrix. By using the similar updating method, we solve B_1 and B_3 as

$$B_1 \leftarrow \frac{1}{2+\mu} \left(2 H_m^{(v)} P^{(v)} + \mu (H_m^{(v)} P^{(v)} S^{(v)} + R_1) \right), \tag{16}$$

$$B_3 \leftarrow \frac{1}{2\lambda + \mu} (2\lambda F Q^{(v)} + \mu W^{(v)} H_m^{(v)} P^{(v)} S^{(v)} + \mu R_3) \tag{17}$$

We use Singular Value Thresholding operator [Cai *et al.*, 2010] to solve B_2. Let $\Theta_\tau(A) = U \Lambda_\tau V^{\mathrm{T}}$, where $A = U \Lambda_\tau V^{\mathrm{T}}$ is the singular value decomposition, and $\Lambda_\tau(a) = \text{sgn}(a)\max(|a| - \tau, 0)$ is the shrinkage operator. Then the update rule for B_2 is $B_2 \leftarrow \Theta_{\alpha/\mu}(S^{(v)} + R_2)$. Considering the non-negative constraint, B_4 can be solved by $B_4 \leftarrow \max(S^{(v)} + R_4, 0)$. Furthermore, Lagrangian multipliers R_1, R_2, R_3 and R_4 should also be updated. Algorithm 1 summarizes the ADMM optimization procedure for solving $S^{(v)}$.

3.5 Updating W_s and $W_p^{(v)}$

In this subproblem, we fix the related parts of W_s and $W_p^{(v)}$ from J and optimize the W_s and $W_p^{(v)}$. The optimization problem becomes

$$\sum_{v=1}^{V} (||(W_s + W_p^{(v)}) H_m^{(v)} P^{(v)} S^{(v)} - F Q^{(v)}||_F^2 + \beta_1 ||W_p^{(v)}||_1 + \beta_2 ||W_s||_F^2). \tag{18}$$

We can update W_s and $W_p^{(v)}$ by solving ridge regression and lasso regression problems. W_s can be updated by the closed form solution, and $W_p^{(v)}$ can be updated by the standard coordinate descent method [Lange, 2008].

3.6 Updating F

Considering the equality constraint $F_{\pi_l} = Y$ imposed on F, we introduce a penalty to arrive at the following equivalent objective function to solve F,

$$J(F) = \sum_{v=1}^{V} ||W^{(v)} H_m^{(v)} P^{(v)} S^{(v)} - F Q^{(v)}||_F^2 + \eta ||FU - Y||_F^2, \tag{19}$$

where $U \in \{0, 1\}^{n \times l}$ is the correspondence matrix. If the i-th column in F corresponds to the j-th column in Y, then $U_{ij} = 1$, and $U_{ik} = 0$, $k \neq j$. $\eta > 0$ is used to control the equality constraint, which should be large enough to ensure the equality constraint satisfied. Let the partial derivative

Algorithm 1: The algorithm to solve $S^{(v)}$.

Input: $H_m^{(v)}$, $P^{(v)}$, $W^{(v)}$, F, $Q^{(v)}$, α, λ, μ.

Initialization: $\forall j, B_j = R_j = 0$
while *not converged* **do**
  Update $S_i^{(v)}$ by (15);
  Update B_1, B_2, B_3, B_4 in Sec. 3.4;
  Update the Lagrange multipliers:
  $R_1 \leftarrow R_1 - (B_1 - H_m^{(v)} P^{(v)} S^{(v)})$;
  $R_2 \leftarrow R_2 - (B_2 - S^{(v)})$;
  $R_3 \leftarrow R_3 - (B_3 - W^{(v)} H_m^{(v)} P^{(v)} S^{(v)})$;
  $R_4 \leftarrow R_4 - (B_4 - S^{(v)})$;
end

Algorithm 2: The learning procedure of our approach.

Input: $\{X^{(v)}, P^{(v)}, Q^{(v)}\}_{v=1}^{V}$, m, λ, α, β_1, β_2.

Initialize $Z_i^{(v)}$, $H_m^{(v)}$ by pre-training.
while *not converged* **do**
  for $v = 1, ..., V$ **do**
    Update $\{Z_i^{(v)}\}_{i=1}^{m}$ in Sec. 3.2;
    Update $H_m^{(v)}$ by (10);
    Update $S^{(v)}$ by Algorithm 1;
    Update $W_p^{(v)}$ by solving (18);
  end
  Update W_s by solving (18);
  Update F by (20);
end

$\partial(J(F))/\partial(F) = 0$, the following update rule can be derived to solve F

$$F = (\sum_{v=1}^{V} W^{(v)} H_m^{(v)} P^{(v)} S^{(v)} Q^{(v)^T} + \eta Y U^T) (\sum_{v=1}^{V} Q^{(v)} Q^{(v)^T} + \eta U U^T)^{-1}. \quad (20)$$

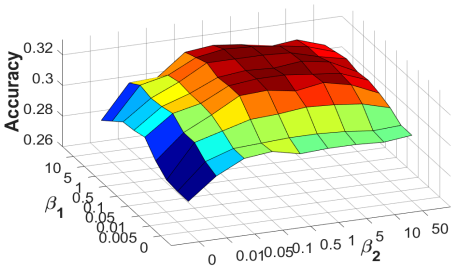
3.7 Complexity Analysis

Pre-training and fine-tuning are the two stages of our algorithm. The computational complexity for pre-training is of order $\mathcal{O}(ndk + nk^2 + kn^2)$, where n is the number of samples, d is the maximum dimension of multi-view data, and k is the maximum dimension of all the layers. All the variables are updated in the fine-tuning stage. The complexity for fine-tuning is of order $\mathcal{O}(ndk + nk^2 + kn^2 + n^3 + ck^2 + cn^2)$. The proposed algorithm is efficient and achieves comparative complexity with NMF and subspace clustering methods.

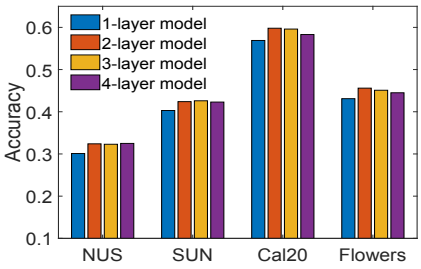
4 Experiments

4.1 Experimental Setup

Datasets. Four commonly used multi-view datasets are used to evaluate the proposed method. *NUS* [Chua *et al.*, 2009] is a web image dataset for object recognition. We adopt 31 categories and choose 100 images for each class. Five features are extracted to represent images. *SUN* [Xiao *et al.*, 2016] contains 899 categories and 130519 images. We randomly choose 30 classes, and each class has 100 images. Five types of visual features are adopted as different views. *Caltech* [Li *et al.*, 2007] is an object recognition data set containing 101 categories of images. We select the widely used 20 classes and get 1230 images. Five features are extracted from the images. *Flowers* [Nilsback and Zisserman, 2006] is composed of 17 flower categories, with 80 images for each category. Each image is described by three views using color, shape, and texture features. For each dataset, 70% data are randomly sampled for training and the remaining 30% data are used for testing. To create incomplete multi-view data scenarios, we randomly remove ε% samples from each view and ensures that each sample appears in at least one view.



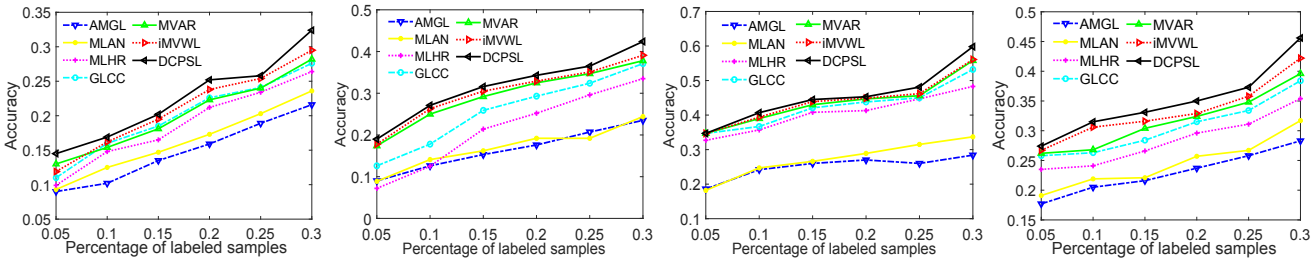
(a) Sensitivity analysis of β_1, β_2 on NUS.



(b) Sensitivity analysis of m on each dataset

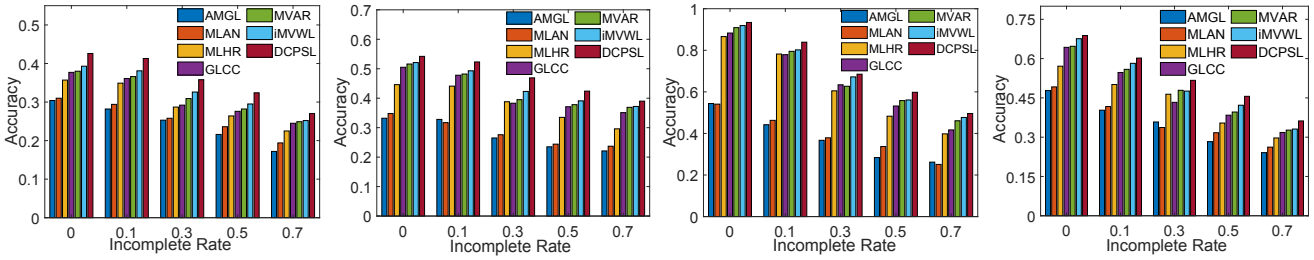
Figure 1: Parameter analysis w.r.t. β_1, β_2 and m.

Methods and parameter setting. We compare six state-of-the-art multi-view learning methods to demonstrate the effectiveness of our method: AMGL [Nie *et al.*, 2016], M-LAN [Feiping Nie and Li, 2018], MLHR [Yang *et al.*, 2013], GLCC [Zhang and Zhang, 2016], MVAR [Tao *et al.*, 2017], iMVWL [Tan *et al.*, 2018]. The first five methods are conventional multi-view semi-supervised learning methods which are designed for complete view data. For a fair comparison, we adopt the matrix completion method [Lin *et al.*, 2010] by filling the missing information and then conduct classification for these methods. The parameters of all the compared methods are set as suggested in the corresponding papers. The parameters of our method are determined by five fold cross-validation. λ and α are selected from $\{10^{-3}, 10^{-2}, ..., 10^{2}\}$. β_1 and β_2 are selected from $\{0.005, 0.01, ..., 50\}$. All the experiments are repeated ten times and the averaged performance are reported. For the evaluation metric, we follow [Feiping Nie and Li, 2018] and use accuracy for performance evaluation, which calculates the proportion of the correctly classified samples.



(a) NUS (b) SUN (c) Cal20 (d) Flowers

Figure 2: Semi-supervised classification results comparison. The incomplete rate of multi-view data is $\varepsilon\% = 50\%$.



(a) NUS (b) SUN (c) Cal20 (d) Flowers

Figure 3: Semi-supervised classification results for different incomplete rates. The percentage of labeled samples is 0.3.

4.2 Experimental Results

Semi-supervised Classification. To illustrate the semi-supervised classification performance of DCPSL, we fix the incomplete rate of the multi-view data as $\varepsilon\% = 50\%$, and present the classification accuracy of all the methods with different percentages of labeled samples in Figure 2. In can be observed that DCPSL achieves better classification accuracy compared to all the other methods on each dataset. The largest performance improvements on the four datasets are: 2.9%, 3.3%, 3.7% and 3.4%, respectively. The classification results clearly verify that the learned deep correlated predictive subspace is helpful to improve the class label prediction performance for incomplete multi-view data. The conventional multi-view methods such as AMGL, MLAN, MLHR, GLCC, MVAR fail to achieve good performance due to they cannot well handle the incomplete multi-view data. Since DCPSL can jointly learn the proper subspace representation and discriminative label predictors for incomplete multi-view data, our method outperforms the other methods with different percentages of labeled samples.

Influence of Incomplete Rate. To evaluate the influence of incomplete rate $\varepsilon\%$ on classification, we conduct classification experiments by changing the incomplete rate $\varepsilon\%$ from $\{0, 10\%, 30\%, 50\%, 70\%\}$ while fixing the percentage of labeled samples to 0.3. The classification results are shown in Figure 3. It clearly shows that DCPSL performs better than the other methods on each dataset. Due to the influence of missing views, the performance of all the methods are declined with the increase of $\varepsilon\%$. In contrast, DCPSL achieves better performance than the others by effectively leveraging both data correlation and multi-view complementary information of the incomplete data.

Parameter Analysis. We conduct the sensitivity analysis with several critical parameters β_1, β_2, and m. $\varepsilon\%$ is set to 0.5, and the percentage of labeled samples is set to 0.3. The experimental results of β_1 and β_2 on NUS dataset are shown in Figure 1(a). DCPSL obtains competitive performance when $\beta_1 = \{0.05, ..., 5\}$ and $\beta_2 = \{0.1, ..., 10\}$. The sensitivity analysis of the number of layers m are shown in Figure 1(b). Better classification results can be obtained when $m = \{2, 3\}$ for most of the datasets. The shallow model ($m = 1$) fails to learn the discriminative subspace representation, so its classification performance is limited.

5 Conclusion

In this paper, a deep correlated predictive subspace learning method (DCPSL) is developed for incomplete multi-view semi-supervised classification. Our method is capable of jointly leveraging the data correlations and multi-view complementary information, which is achieved by integrating deep correlated predictive subspace learning and multi-view shared and private label prediction into a unified objective function. Compared with the state-of-the-art multi-view semi-supervised learning methods, DCPSL can better handle the incomplete multi-view data and achieves competitive classification results on various practical datasets.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China: 61802028, 61532006, 61772083 and 61877006, in part by the Fundamental Research Funds for the Central University (No. 2018RC44), in part by US NSF IIS-1816227.

References

[Cai *et al.*, 2010] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, page 48, 2009.

[Elhamifar and Vidal, 2013] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *TPAMI*, 35(11):2765–2781, 2013.

[Feiping Nie and Li, 2018] Jing Li Feiping Nie, Gguohao Cai and Xuelong Li. Auto-weighted multi-view learning for image clustering and semi-supervised classification. *TIP*, 27(3):1501–1511, 2018.

[Hu and Chen, 2018] Menglei Hu and Songcan Chen. Doubly aligned incomplete multi-view clustering. In *IJCAI*, pages 2262–2268, 2018.

[Lange, 2008] Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, 2(1):224–244, 2008.

[Li *et al.*, 2007] Fei-Fei Li, Robert Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 106(1):59–70, 2007.

[Li *et al.*, 2014] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial multi-view clustering. In *AAAI*, pages 1968–1974, 2014.

[Lin *et al.*, 2010] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *CoRR*, abs/1009.5055, 2010.

[Liu *et al.*, 2012] Haifeng Liu, Zhaohui Wu, Xuelong Li, Deng Cai, and Huang Thomas S. Constrained nonnegative matrix factorization for image representation. *TPAMI*, 34(7):1299–1311, 2012.

[Liu *et al.*, 2013] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *TPAMI*, 35(1):171–184, 2013.

[Nie *et al.*, 2016] Feiping Nie, Jing Li, and Xuelong Li. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *IJCAI*, pages 1881–1887, 2016.

[Nilsback and Zisserman, 2006] Maria-Elena Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *CVPR*, pages 1447–1454, 2006.

[Shao *et al.*, 2015] Weixiang Shao, Lifang He, and Philip Yu. Multiple incomplete views clustering via weighted nonnegative matrix factorization with $l_{2,1}$ regularization. In *ECML PKDD*, pages 318–334, 2015.

[Tan *et al.*, 2018] Qiaoyu Tan, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang. Incomplete multi-view weak-label learning. In *IJCAI*, pages 2703–2709, 2018.

[Tao *et al.*, 2017] Hong Tao, Chenping Hou, Feiping Nie, Jubo Zhu, and Dongyun Yi. Scalable multi-view semi-supervised classification via adaptive regression. *TIP*, 26(9):4283–4296, 2017.

[Trigeorgis *et al.*, 2014] George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Björn W Schuller. A deep semi-nmf model for learning hidden representations. In *ICML*, pages 1692–1700, 2014.

[Vidal, 2011] René Vidal. Subspace clustering. *SPM*, 28(2):52–68, 2011.

[Wen *et al.*, 2019] Jie Wen, Yong Xu, and Hong Liu. Incomplete multiview spectral clustering with adaptive graph learning. *TCYB*, pages 1–12, 2019.

[Xiao *et al.*, 2016] Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. SUN database: Exploring a large collection of scene categories. *IJCV*, 119(1):3–22, 2016.

[Xu *et al.*, 2018] Nan Xu, Yanqing Guo, Xin Zheng, Qianyu Wang, and Xiangyang Luo. Partial multi-view subspace clustering. In *ACM MM*, pages 1794–1801, 2018.

[Yang *et al.*, 2013] Yi Yang, Jingkuan Song, Zi Huang, Zhigang Ma, Nicu Sebe, and Alexander G Hauptmann. Multi-feature fusion via hierarchical regression for multimedia analysis. *TMM*, 15(3):572–581, 2013.

[Yang *et al.*, 2018] Yang Yang, De-Chuan Zhan, Xiang-Rong Sheng, and Yuan Jiang. Semi-supervised multi-modal learning with incomplete modalities. In *IJCAI*, pages 2998–3004, 2018.

[Zhang and Zhang, 2016] Lei Zhang and David Zhang. Visual understanding via multi-feature shared learning with global consistency. *TMM*, 18(2):247–259, 2016.

[Zhao *et al.*, 2016] Handong Zhao, Hongfu Liu, and Yun Fu. Incomplete multi-modal visual data grouping. In *IJCAI*, pages 2392–2398, 2016.