Accepted Manuscript

Unified Subspace Learning for Incomplete and Unlabeled Multi-view Data

Qiyue Yin, Shu Wu, Liang Wang

PII: S0031-3203(17)30034-1
DOI: 10.1016/j.patcog.2017.01.035
Reference: PR 6055

To appear in: *Pattern Recognition*

Received date: 11 August 2016
Revised date: 15 January 2017
Accepted date: 19 January 2017

Please cite this article as: Qiyue Yin, Shu Wu, Liang Wang, Unified Subspace Learning for Incomplete and Unlabeled Multi-view Data, *Pattern Recognition* (2017), doi: 10.1016/j.patcog.2017.01.035

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Highlights

- Class indicator matrix is learned for incomplete and unlabeled multi-view data.
- Preserving the inter-view and intra-view data similarity can improve performance.
- Running time is in the same magnitudes with that of the mainstream methods.
- Obtain best results for incomplete multi-view clustering and cross-modal retrieval.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Unified Subspace Learning for Incomplete and Unlabeled Multi-view Data

Qiyue Yin, Shu Wu, Liang Wang*

Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{*qyyin, shu.wu, wangliang*}*@nlpr.ia.ac.cn*

Abstract

Multi-view data with each view corresponding to a type of feature set are common in real world. Usually, previous multi-view learning methods assume complete views. However, multi-view data are often incomplete, namely some samples have incomplete feature sets. Besides, most data are unlabeled due to a large cost of manual annotation, which makes learning of such data a challenging problem. In this paper, we propose a novel subspace learning framework for incomplete and unlabeled multi-view data. The model directly optimizes the class indicator matrix, which establishes a bridge for incomplete feature sets. Besides, feature selection is considered to deal with high dimensional and noisy features. Furthermore, the inter-view and intra-view data similarities are preserved to enhance the model. To these ends, an objective is developed along with an efficient optimization strategy. Finally, extensive experiments are conducted for multi-view clustering and cross-modal retrieval, achieving the state-of-the-art performance under various settings.

Keywords: Multi-view learning, Subspace learning, Incomplete and unlabeled data, Multi-view clustering, Cross-modal retrieval

*Corresponding author

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

1. Introduction

Various kinds of real-world data appear in multiple modalities or come from multiple channels. For example, a web page can be described by both images and texts, and an image can be encoded by different visual features such as SIFT and GIST. Such data are called multi-view data with each view representing a type of feature set and these views can be homogeneous descriptors or heterogeneous modalities. Usually, multiple views provide complementary information for the semantically same data, which motivates the multi-view learning to obtain better performance than using a single view [1]. Besides, Multi-view data describing the same content lead to the research of exploring consistent information between different views, which results in cross-modal matching tasks [2].

Recently, plenty of methods have been developed for multi-view data to explore complementarity and consistency characteristics. It should be noted that most methods focus on complete multi-view data, which means all data samples in the datasets have complete feature sets. However, in real applications, it is often the case that some views suffer from missing information leading to incomplete multi-view data. For example, given a two view dataset with visual and textual features, some samples have only either visual or textual feature with only part of them sharing both representations. Under such scenario, traditional multi-view learning methods usually face notable performance degeneration [3, 4]. Besides, real multi-view data are often unlabeled due to the expensive cost of manual annotation, which makes the learning of incomplete multi-view data a challenging problem.

Generally, to model incomplete and unlabeled multi-view data, we confront two basic challenges. The first one is how to handle incomplete multi-view data. Since some samples have incomplete feature representations, a naive strategy is to remove such examples and only use samples with complete feature sets. However, such methods are contradicting with some tasks such as clustering because we need to cluster all the data samples. More importantly, they cannot

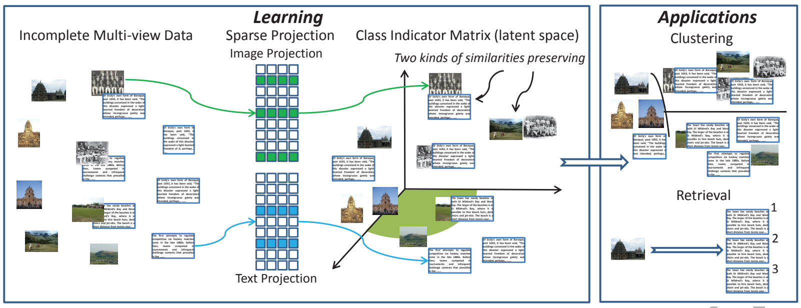

Figure 1: The overview of our model with two views, i.e., text and image. For the incomplete multi-view dataset, we use projection matrix to project the original features to the class indicator matrix, which explicitly captures the clustering structure and serves as the latent space. Besides, group sparsity is imposed on the projection matrices for feature selection. Furthermore, the inter-view and intra-view data similarities are preserved to enhance the model. Finally, our model can be applied for clustering and retrieval tasks.

make full use of the whole data to learn models. Another strategy is to fill missing information. For example, matrix completion based methods [5] utilize low rank structure of the matrix to fill missing entities. However, those methods usually cannot perform feature selection to deal with high dimensional and noisy features. Thus by filling missing information is not a satisfactory strategy. Overall, a suitable model should use samples with complete feature representations and meanwhile utilize examples with incomplete feature sets to enhance the learning process.

The second challenge is how to explore complementarity and consistency for unlabeled multi-view data. Usually, for multi-view data describing semantically same content, different views share common characteristics and have view-specific characteristics, which makes the modeling of those characteristics complex. Furthermore, given unlabeled data, we just have the corresponding relation between different views and this makes discover the structure of multi-view data harder. Most previous methods try to find a low dimensional subspace, where data samples under different views can be compared for exploring the above characteristics. For example, canonical correlation analysis (CCA) based approaches [6, 7] aim to find linear projections of different views with maximal mutual correlation, and multi-view non-negative matrix factorization based methods learn unified latent representations among multiple sources of

ACCEPTED MANUSCRIPT

information [8, 9]. However, those methods cannot thoroughly explore the data semantics in the learned subspace. To sum up, one good subspace should reflect such information and meanwhile make use of multiple views.

In this paper, we propose a novel subspace learning framework to alleviate the above problems, as shown in Figure 1. We directly optimize the class indicator matrix as a shared subspace through linear projection matrices, which has two advantages: 1) establishing a bridge for different views based on their optimized labels whether the multi-view data are complete or incomplete, and 2) the class indicator matrix in turn guides the subspace learning in a supervised manner to make the learning process more accurate. Since data are often with high dimensional and noisy features, the projection matrices are enforced to be sparse to select relevant features when learning the latent space. Furthermore, the inter-view and intra-view data similarities are preserved to enhance the subspace learning. To these ends, an objective is developed with an efficient optimization strategy and convergence analysis. The experimental results show that our method outperforms the state-of-the-art methods.

Our contributions can be summarized as follows. 1) We propose a novel subspace learning based incomplete and unlabeled multi-view learning method, which jointly considers feature selection and inter-view and intra-view similarity preserving to enhance the subspace learning. 2) We develop an iterative optimization algorithm to efficiently solve the proposed objective, and provide theoretical analysis to guarantee its convergence. 3) We validate our proposed method with extensive experiments under two settings in terms of two tasks, i.e., multi-view clustering and cross-modal retrieval, achieving better performance than the state-of-the-art methods.

The rest of the paper is organized as follows. In Section 2, we briefly review multi-view learning, especially multi-view clustering and cross-modal retrieval. Section 3 describes our model, along with its optimization and convergence analysis. In Section 4, we report experimental results on multi-view clustering and cross-modal retrieval. Finally, we draw the conclusion in Section 5.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

2. Related work

In this section, we briefly review general multi-view learning methods. Since we are focusing on two specific multi-view learning tasks, i.e., multi-view clustering and cross-modal retrieval, we also introduce recent progresses of them.

2.1. Multi-view learning

Multi-view learning deals with data represented by multiple distinct feature sets and aims at boosting learning performance or discovering correlation. It has a wide range of applications, such as dimensionality reduction, classification and clustering. Generally, existing multi-view learning algorithms can be categorized into three schemes [1]. Co-training [10] is one of the earliest framework, which alternately maximizes the agreement of two feature sets. Soon after, plenty of variants are developed, such as generalized expectation-maximization (EM) and methods fusing co-training and other algorithms [11]. Multiple kernel learning solves multi-view learning by regarding different kernels as different views and then combining those kernels through linear or non-linear strategies. Such framework is widely studied and readers can refer to [12] for more details. The last framework is subspace learning, which aims to find a low dimensional space to measure the consistency and complementarity among multi-view data. Typical examples such as Canonical Correlation Analysis (CCA) and its various extensions [13, 14, 15] have obtained promising results in various tasks. In this paper, a novel subspace learning framework is developed for learning incomplete and unlabeled multi-view data.

2.2. Multi-view clustering

Multi-view clustering, as one of basic tasks of multi-view learning, provides a natural way to cluster multi-view datasets [16, 17, 18]. Generally, the main challenge lies in the mining of the complementary information among multiple sources of information. Fortunately, a number of promising approaches have been proposed, which can be roughly classified into four categories [1]. Methods in the first category are subspace based ones [19, 8, 20, 21] and in the second

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

category are co-training based algorithms [22, 23], which are popular frameworks as mentioned in multi-view learning. The third category is called late fusion [24, 25], which combines the clustering results of different views by voting or other fusion strategies. The last category learns a unified similarity matrix among multi-view data [26, 27] based on subspace segmentation algorithms. Then the matrix serves as an affinity matrix for final clustering.

The existing multi-view clustering methods mainly focus on the data with complete views, i.e., every data example has complete feature sets. As for incomplete views, only a few works have been developed. Piyush et al. [28] and Shao et al. [29] proposed spectral-based multi-view clustering methods by filling kernel matrices of incomplete views through Laplacian regularization, which can only fit kernel-based multi-view clustering. Recently, Li et al. [30] and Shao et al. [31] proposed subspace learning based incomplete multi-view clustering method by using nonnegative matrix factorization (NMF). However, NMF cannot be utilized for data with negative feature values. Xu et al. [5] developed a matrix completion based incomplete multi-view learning method, but they cannot perform feature learning to deal with high dimensional and even noisy features. Hence, we propose a new subspace learning framework to consider all above factors.

2.3. Cross-modal retrieval

As a basic task of cross-modal matching, cross-modal retrieval plays an important role in many real applications [32]. Aiming to explore correlation between different modalities, different kinds of methods are developed. Probabilistic models are widely applied for specific cross-modal matching tasks, i.e., image annotation exploring relation between images and tags [33]. Metric learning approaches aim to learn a metric between different modalities. Usually, similar pairs and dissimilar pairs or ranking lists are considered for similarity calculation between different modalities [34, 35]. Recently, to speed up retrieval, binary representations of different modalities are learned. Those methods aiming to find a Hamming space usually sacrifice accuracy for speed with typical

ACCEPTED MANUSCRIPT

examples such as [36, 37].

The most related kind of algorithms with ours are subspace based methods, such as Canonical Correlation Analysis (CCA) [7], Partial Least Squares (PLS) [38] and Bilinear Model (BLM) [39, 40]. Those methods are typical unsupervised algorithms with wide-spread applications. Besides, labels are considered to enhance the subspace learning [41]. For example, Lin and Tang [42] proposed to learn a latent subspace so as to maximize the difference between within scatter matrix and between scatter matrix. Sharma et al. [39] developed Generalized Multiview LDA and Generalized Multiview MFA, which are based on single view Linear Discriminant Analysis (LDA) and Marginal Fisher Analysis (MFA). Recently, deep learning methods are applied for cross-modal retrieval, which aim to learn features for multiple modalities and meanwhile to explore their correlation [2]. In [43], Kang et al. proposed a supervised subspace learning method under the incomplete scenario, but their approach cannot deal with unsupervised data. Generally, most above methods ignore the incomplete multi-view scenario, which, however, is our focus here.

This paper is built upon our preliminary conference version [44], and the main extensions are summarized as follows. 1) While the previous paper [44] mainly focuses on incomplete multi-view clustering, we now propose to model for incomplete and unlabeled multi-view data. Accordingly, the previous work is just a special case of this paper. 2) We extend previous objective for two views to a multi-view case, and more than two views experiments are conducted. 3) We conduct extensive experiments of a new task, i.e.,unsupervised cross-modal retrieval, which further validate the effectiveness of our model. Besides, more experiments, e.g., running time, are designed to improve incomplete multi-view clustering.

3. Model

3.1. Preliminaries

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

For the incomplete multi-view data, we focus on a scenario that features of all the views are available only for part of the dataset and the other samples only have partial views. Such a scenario is often appeared in webpage applications, where two/three views are frequently utilized. For example, in webpages clustering based on images, texts and hyperlinks, usually partial webpages have all the three feature sets, and the other webpages contain only one or two views.

Suppose we are given a l view dataset with n samples categorized into k clusters. We use $\mathbf{X}_C^{(g)}$ and $\hat{\mathbf{X}}^{(g)}$ to represent feature matrix of the g-th view for examples with complete views and the other samples in view g, respectively. Besides, their sizes are denoted as c and n_g satisfying $c < n$. We use d_g to indicate the feature dimensionality of the g-th view. Accordingly, feature matrix of examples under the g-th view are denoted as $\bar{\mathbf{X}}^{(g)} = [\mathbf{X}_C^{(g)}, \hat{\mathbf{X}}^{(g)}] \in R^{d_g \times (c+n_g)}$. Generally, multi-view data consisting of heterogeneous feature sets represent the same object, and therefore they share the same class labels. Similarly, $\mathbf{Y}^C$ and $\hat{\mathbf{Y}}^{(g)}$ are utilized to represent the class indicator of the g-th view. Finally we use $\mathbf{Y} \in \mathrm{R}^{n \times k}$ to denote the class indictor matrix of the n samples with the i-th row of $\mathbf{Y}$ satisfying $\mathbf{Y}(i,:) \in \{0,1\}^{1\times k}$ being the class indicator vector of the i-th sample. The notations are summarized in Table 1.

Table 1: Notations and Explanations.

notation	size	description
$\mathbf{X}_C^{(g)}$	$d_g \times c$	feature matrix of the g-th view for examples with complete views
$\hat{\mathbf{X}}^{(g)}$	$d_g \times n_g$	feature matrix of the g-th view for examples excluding $\mathbf{X}_C^{(g)}$
$\bar{\mathbf{X}}^{(g)}$	$d_g \times (c+n_g)$	feature matrix of the g-th view consisting of $\mathbf{X}_C^{(g)}$ and $\hat{\mathbf{X}}^{(g)}$
$\mathbf{Y}^C$	$c \times k$	class indicator matrix of samples with complete views
$\hat{\mathbf{Y}}^{(g)}$	$n_g \times k$	class indicator matrix of the g-th view for samples excluding $\mathbf{X}_C^{(g)}$
$\bar{\mathbf{Y}}^{(g)}$	$(c+n_g) \times k$	class indicator matrix of the g-th view for examples appearing in view g
$\mathbf{Y}$	$n \times k$	class indicator matrix of all the n samples
$\mathbf{U}_{(g)}$	$d_g \times k$	learned projection matrix for the g-th view

Since our incomplete multi-view data are unsupervised, we do not know exactly what the $\mathbf{Y}$ is, but we are aware of its structure and our task is to learn such $\mathbf{Y}$, which serves as a unified subspace for the incomplete and unlabeled multi-view data. Finally, based on the learned subspace, we can deal with various multi-view tasks, e.g., multi-view clustering and cross-modal retrieval.

3.2. Formulation

We aim to optimize the class indicator matrix $\mathbf{Y}$ for the incomplete and unlabeled multi-view data and the advantages are listed as follows. 1) $\mathbf{Y}$ reflects the class indicator of the multi-view data, which is a relatively higher level semantic representation of data. Even though data consist of multiple heterogeneous features, they potentially share the same semantic information. 2) By introducing the above semantic space, we construct a bridge for different heterogeneous feature sets even though some samples have incomplete views. 3) Given the optimized $\mathbf{Y}$, we can conduct multi-view learning in a supervised manner, which in turn enhances the learning process. For example, using such an indictor matrix, we can perform feature selection in a supervised manner.

To learn the class indictor matrix, we learn a projection matrix $\mathbf{U}_{(g)} \in R^{d_g \times k}$ for each view to project their original spaces to such a semantic space as always done in classification tasks. Then the objective can be formulated as:

$$\begin{array}{l} \min\limits_{\mathbf{U},\mathbf{Y}} \sum_{g=1}^{l} \ell\left((\bar{\mathbf{X}}^{(g)}, \mathbf{U}_{(g)}), \bar{\mathbf{Y}}^{(g)}\right) + \beta \sum_{g=1}^{l} \varphi\left(\mathbf{U}_{(g)}\right) + \gamma \Omega\left(\mathbf{U}_{(1)}, ..., \mathbf{U}_{(l)}\right) \\ s.t. \quad \mathbf{Y} \in \{0,1\}^{n \times k}; \quad \mathbf{Y}\mathbf{1}_k = \mathbf{1}_n \end{array} \tag{1}$$

In the above objective, there are four parts: feature projection for incomplete and unlabeled multi-view data, feature learning, data similarity preserving and constrains. As for the constraints, $\mathbf{1}_k$ and $\mathbf{1}_n$ are k and n dimensional column vectors with their values all being 1. Using the constraints, we force each data sample belong to only one class. Next, we elaborate different parts.

Feature projection: As stated in the introduction part, a good way to deal with incomplete dataset should make use of data examples whether they consist of complete feature sets or not. Thus, we project all samples under different views to the semantic space and establish the relation between views by enforcing the samples consisting of complete feature sets to share the same class indicator vectors. By doing so, we can learn the class indicator matrix for all data samples and learn projection matrices for a view based on all the examples in that view. Then the first part of Equation 1 is written as:

$$\ell\left((\bar{\mathbf{X}}^{(g)}, \mathbf{U}_{(g)}), \bar{\mathbf{Y}}^{(g)}\right) = \left\| [\mathbf{X}_C^{(g)}, \hat{\mathbf{X}}^{(g)}]^T \mathbf{U}_{(g)} - [\mathbf{Y}^C; \hat{\mathbf{Y}}^{(g)}] \right\|_F^2 \tag{2}$$

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

where $\mathbf{Y}^C \in R^{c\times k}$ and $\hat{\mathbf{Y}}^{(g)} \in R^{n_g\times k}$ are the learned class indicator matrices for data examples with complete views and only with the g-th view respectively.

Feature learning: In Equation 1, a commonly used regularizer for $\mathbf{U}_{(g)}$ is the F-norm to avoid over-fitting. However, we choose the l_{21}-norm here to perform feature selection like in supervised feature learning methods [45]. By doing so, we can well deal with high dimensional and noisy features of each view, and it is defined as:

$$\beta\varphi\left(\mathbf{U}_{(g)}\right) = \beta\left\|\mathbf{U}_{(g)}\right\|_{21} \tag{3}$$

where $||\mathbf{U}_{(g)}||_{21} = \sum_i ||\mathbf{U}_{(g)}(i,:)||_2$ and $\mathbf{U}_{(g)}(i,:)$ is the i-th row of $\mathbf{U}_{(g)}$. When β is big, only a small subset of features will be selected, otherwise a large subset will be chosen.

Similarity preserving: We hope to preserve the intra-view similarity and the inter-view similarity to further enhance the learning of projection matrices. More specifically, the neighborhood relationship between data samples under each view and the pairwise relationship for an example under different views should be preserved in the latent space. The data similarities are:

$$W_{ij}^{(g)} = \begin{cases} \exp(\frac{-z_{ij}^{(g)}}{2\sigma^2}), \bar{\mathbf{x}}_i^{(g)} \in N_m(\bar{\mathbf{x}}_j^{(g)}) \,\text{or}\, \bar{\mathbf{x}}_j^{(g)} \in N_m(\bar{\mathbf{x}}_i^{(g)}) \\ 0, \qquad \text{otherwise} \end{cases} \tag{4}$$

$$W_{ij}^{(pq)} = \begin{cases} 1, & \text{if } \bar{\mathbf{x}}_i^{(p)} \text{and } \bar{\mathbf{x}}_j^{(q)} \text{represent the same sample} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $\mathbf{W}^{(g)}, g = 1, ..., l$ is the similarity matrix of the g-th view calculated using the Gaussian kernel. $z_{ij}^{(g)}$ is the Euclidean distance between two data examples, σ is width parameter for the Gaussian kernel, and $N_m(\bar{\mathbf{x}}_i^{(g)})$ indicates the examples of m nearest neighbors of $\bar{\mathbf{x}}_i^{(g)}$. $\mathbf{W}^{(pq)}, p = 1, ..., l; q = 1, ..., l; p \neq q$ is the similarity matrix for view p to view q. When features $\bar{\mathbf{x}}_i^{(p)}$ and $\bar{\mathbf{x}}_j^{(q)}$ indicate the same example, 1 is given as the weight, otherwise 0. From the definition, we have $\mathbf{W}^{(pq)} = (\mathbf{W}^{(qp)})^T$.

Based on the above similarities, we define the regularization on the projection matrices as:

$$\begin{aligned}&\Omega(\mathbf{U}_{(1)},...,\mathbf{U}_{(l)}) = \\ &\sum_g \sum_{ij} W_{ij}^{(g)} \left\| \mathbf{U}_{(g)}^T \bar{\mathbf{x}}_i^{(g)} - \mathbf{U}_{(g)}^T \bar{\mathbf{x}}_j^{(g)} \right\|_F^2 + \sum_p \sum_{q \neq p} \sum_{ij} W_{ij}^{(pq)} \left\| \mathbf{U}_{(p)}^T \bar{\mathbf{x}}_i^{(p)} - \mathbf{U}_{(q)}^T \bar{\mathbf{x}}_j^{(q)} \right\|_F^2\end{aligned} \tag{6}$$

We define the overall similarity matrix $\mathbf{W}$ based on the above inter-view and intra-view similarities as:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}^{(1)} & \mathbf{W}^{(12)} & \dots & \mathbf{W}^{(1l)} \\ \mathbf{W}^{(21)} & \mathbf{W}^{(2)} & \dots & \mathbf{W}^{(2l)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}^{(l1)} & \mathbf{W}^{(l2)} & & \mathbf{W}^{(l)} \end{bmatrix} \tag{7}$$

Then Equation 6 can be rewritten as:

$$\Omega(\mathbf{U}_{(1)},...,\mathbf{U}_{(l)}) = \sum_{p=1}^{l} \sum_{q=1}^{l} Tr(\mathbf{U}_{(p)}^T \bar{\mathbf{X}}^{(p)} \mathbf{L}_{pq} (\bar{\mathbf{X}}^{(q)})^T \mathbf{U}_{(q)}) \tag{8}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix and $\mathbf{D}$ is a diagonal matrix with its i-th diagonal element defined as the sum of the i-th row in $\mathbf{W}$. Tr is the trace of a matrix.

Finally, our objective is rewritten as:

$$\begin{aligned}&\min_{\mathbf{U},\mathbf{Y}} \sum_{i=1}^{l} \left\| [\mathbf{X}_C^{(i)}, \hat{\mathbf{X}}^{(i)}]^T \mathbf{U}_{(i)} - [\mathbf{Y}^C; \hat{\mathbf{Y}}^{(i)}] \right\|_F^2 + \beta \sum_{i=1}^{l} \left\| \mathbf{U}_{(i)} \right\|_{21} \\ &\quad + \gamma \sum_{p=1}^{l} \sum_{q=1}^{l} Tr(\mathbf{U}_{(p)}^T \bar{\mathbf{X}}^{(p)} \mathbf{L}_{pq} (\bar{\mathbf{X}}^{(q)})^T \mathbf{U}_{(q)}) \\ &s.t. \quad \mathbf{Y} \in \{0,1\}^{n \times k}; \quad \mathbf{Y}\mathbf{1}_k = \mathbf{1}_n\end{aligned} \tag{9}$$

In our objective, we have four terms: using the projection matrix to project each incomplete view to the latent space defined by $\mathbf{Y}$; feature selection for each view using the $\ell 21$-norm based regularizer and the inter-view and intra-view similarity preserving term defined by the Laplacian matrix. Besides, the constraints imposed on $\mathbf{Y}$ guarantee that each example only belongs to one group.

3.3. Optimization

Since the variables in Equation 9, i.e., the projection matrix and the latent representation, are coupled together, it may be difficult to optimize them at the same time. Hence, we propose to alternatively optimize the variables to obtain a local solution.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

3.3.1. Optimize the class indicator matrix

Directly optimizing the $\mathbf{Y}$ is hard due to the discrete constraint, we follow previous methods to relax the constraint as [46]:

$$\mathbf{Y}^T\mathbf{Y} = \mathbf{I}_k; \quad \mathbf{Y} \geq \mathbf{0} \tag{10}$$

where $\mathbf{I}_k$ is an identity matrix. The constraints guarantee that there is only one positive value in each row of $\mathbf{Y}$, which is the ideal structure we need. However, different views only have part of all the latent representations, i.e., $[\mathbf{Y}^C; \hat{\mathbf{Y}}^{(g)}]$ for the g-th view is only part of $\mathbf{Y}$, which makes the optimization still not an easy problem. To handle this, we optimize $\mathbf{Y}^C$ and $\hat{\mathbf{Y}}^{(g)}$ separately and relax the constraints to the following form:

$$(\mathbf{Y}^C)^T\mathbf{Y}^C = \mathbf{I}_k, \mathbf{Y} \geq \mathbf{0} \tag{11}$$

Even though the orthogonal constraint on $\mathbf{Y}^C$ may not be rigorous when examples with complete feature sets do not have all kinds of class labels. We ignore this slight influence. In turn, it makes our optimization very compact. As for $\hat{\mathbf{Y}}^{(g)}$, since examples in the same view share the same projection matrix and the same data distribution, $\hat{\mathbf{Y}}^{(g)}$ will have similar characteristic with $\mathbf{Y}^C$. In summary, the relaxed constraints will have almost the same effect with that of the original ones and can make the optimization more succinct.

We denote the objective in Equation 9 as O and the part excluding the $\mathbf{Y}^C$ as $\hat{\mathbf{Y}}$ ($\mathbf{Y} = [\mathbf{Y}^C; \hat{\mathbf{Y}}]$). Then minimizing O over $\mathbf{Y}^C$ and $\hat{\mathbf{Y}}$ are simplified as:

$$\min_{\mathbf{Y}^C} \sum_{g=1}^{l} \left\| (\mathbf{X}_C^{(g)})^T \mathbf{U}_{(g)} - \mathbf{Y}^C \right\|_F^2 \; s.t. (\mathbf{Y}^C)^T\mathbf{Y}^C = \mathbf{I}_k, \mathbf{Y}^C \geq \mathbf{0} \tag{12}$$

$$\min_{\hat{\mathbf{Y}}_i, i=1,\dots,n-c} \sum\nolimits_{g=1}^{l} r_g \left\| (\mathbf{x}_i^{(g)})^T \mathbf{U}_{(g)} - \hat{\mathbf{Y}}_i \right\|^2 \; s.t. \; \hat{\mathbf{Y}}_i \geq 0 \tag{13}$$

where $\hat{\mathbf{Y}}_i$ is the i-th row of $\hat{\mathbf{Y}}$. r_g is an indicator, and its value is set to be 1 if example $\mathbf{x}_i$ has the g-th view, otherwise 0.

To optimize $\mathbf{Y}^C$, we bring in Lagrangian function as:

$$\begin{aligned} L(\mathbf{Y}^C, \Lambda, \Gamma) &= Tr(\Gamma((\mathbf{Y}^C)^T\mathbf{Y}^C - \mathbf{I}_k)) \\ &-Tr(\Lambda\mathbf{Y}^C) + \sum_g Tr(-2\mathbf{A}_g^T\mathbf{Y}^C + (\mathbf{Y}^C)^T\mathbf{Y}^C) \end{aligned} \tag{14}$$

where Γ and $\Lambda \geq 0$ are Lagrangian multipliers of the above function and $\mathbf{A}_g = (\mathbf{X}_C^{(g)})^T \mathbf{U}_{(g)}$. Applying the KKT condition, i.e., $\Lambda(s,t)\mathbf{Y}^C(s,t) = \mathbf{0}$, we obtain:

$$(\sum_g (-\mathbf{A}_g + \mathbf{Y}^C) + \mathbf{Y}^C\Gamma)\,(s,t)\,\mathbf{Y}^C\,(s,t) = \mathbf{0} \tag{15}$$

and we can obtain the following updating rule for $\mathbf{Y}^C$ [47, 48]:

$$\mathbf{Y}^C(s,t) = \mathbf{Y}^C(s,t)\sqrt{\frac{(\sum_g \mathbf{A}_g^+ + \mathbf{Y}^C\mathbf{\Gamma}^-)(s,t)}{(\sum_g (\mathbf{A}_g^- + \mathbf{Y}^C) + \mathbf{Y}^C\mathbf{\Gamma}^+)(s,t)}} \tag{16}$$

where for a matrix $\mathbf{C}$, $\mathbf{C}^+(s,t) = (|\mathbf{C}(s,t)| + \mathbf{C}(s,t))/2$, $\mathbf{C}^-(s,t) = (|\mathbf{C}(s,t)| - \mathbf{C}(s,t))/2$ and $\mathbf{C} = \mathbf{C}^+ - \mathbf{C}^-$. As for $\mathbf{\Gamma}$, its diagonal elements are obtained by summing s: $\mathbf{\Gamma}\,(s,s) = \sum_g ((\mathbf{Y}^C)^T\mathbf{A}_g - \mathbf{I}_k)(s,s)$. The off-diagonal elements of $\mathbf{\Gamma}$ are approximated by ignoring the non-negative values of $\mathbf{Y}^c$: $\mathbf{\Gamma}(s,t) = \sum_g ((\mathbf{Y}^C)^T\mathbf{A}_g - \mathbf{I}_k)(s,t)$. In summary, $\mathbf{\Gamma}$ is calculated by $\mathbf{\Gamma} = \sum_g ((\mathbf{Y}^C)^T\mathbf{A}_g - \mathbf{I}_k)$.

To optimize $\hat{\mathbf{Y}}$, we directly obtain its gradients and the updating rule is:

$$\hat{\mathbf{Y}}_i = \max\left((\sum\nolimits_g r_g (\mathbf{x}_i^{(g)})^T \mathbf{U}_{(g)})/(\sum\nolimits_g r_g), 0\right) \tag{17}$$

where max is an element-wise operator that returns the maximal value.

3.3.2. Optimize the projection matrix

Minimizing the objective O in Equation 9 with respect to $\mathbf{U}_{(g)}$ is rewritten as:

$$\min_{\mathbf{U}_{(g)}} \sum_{g=1}^{l} \left\|(\bar{\mathbf{X}}^{(g)})^T \mathbf{U}_{(g)} - \bar{\mathbf{Y}}^{(g)}\right\|_F^2 + \beta \sum_{g=1}^{l} \|\mathbf{U}_{(g)}\|_{21} + \gamma \sum_{p=1}^{l}\sum_{q=1}^{l} Tr(\mathbf{U}_{(p)}^T \bar{\mathbf{X}}^{(p)} \mathbf{L}_{pq} (\bar{\mathbf{X}}^{(q)})^T \mathbf{U}_{(q)}) \tag{18}$$

where $\bar{\mathbf{X}}^{(g)}, (g = 1, ..., l)$ and $\bar{\mathbf{Y}}^{(g)}, (g = 1, ..., l)$ are the feature matrix and the latent representation for the g-th view as described before. They consist of the data examples with all feature sets and only with the g-th view.

Differentiating the objective function in Equation 18 with respect to $\mathbf{U}_{(g)}$ and setting it to zero, we have the following equation:

$$\begin{aligned} &\bar{\mathbf{X}}^{(g)}((\bar{\mathbf{X}}^{(g)})^T\mathbf{U}_{(g)} - \bar{\mathbf{Y}}^{(g)}) + \beta\mathbf{D}_{(g)}\mathbf{U}_{(g)} \\ &+\gamma\bar{\mathbf{X}}^{(g)}\mathbf{L}_{gg}(\bar{\mathbf{X}}^{(g)})^T\mathbf{U}_{(g)} + \gamma\sum_{t\neq g}\bar{\mathbf{X}}^{(g)}\mathbf{L}_{gt}(\bar{\mathbf{X}}^{(t)})^T\mathbf{U}_{(t)} = 0 \end{aligned} \tag{19}$$

ACCEPTED MANUSCRIPT

where $\mathbf{D}_{(g)}$ is a diagonal matrix with its i-th diagonal element calculated as $\mathbf{D}_{(g)}(i,i) = 1/(2||\mathbf{U}_{(g)}(i,:)||)$, and $\mathbf{U}_{(g)}(i,:)$ is the i-th row of $\mathbf{U}_{(g)}$. Practically, $\mathbf{D}_{(g)}(i,i)$ is calculated by[1]:

$$\mathbf{D}_{(g)}(i,i) = \frac{1}{2\sqrt{||\mathbf{U}_{(g)}(i,:)||^2 + \varepsilon}} \tag{20}$$

where ε is a smoothing term, which is usually set to be a small positive value.

Then Equation 19 is optimized as:

$$\begin{aligned} \mathbf{U}_{(g)} = &(\bar{\mathbf{X}}^{(g)}(\bar{\mathbf{X}}^{(g)})^T + \beta\mathbf{D}_{(g)} + \gamma\bar{\mathbf{X}}^{(g)}\mathbf{L}_{gg}(\bar{\mathbf{X}}^{(g)})^T)^{-1} \\ &(\bar{\mathbf{X}}^{(g)}\bar{\mathbf{F}}^{(g)} - \gamma\sum_{g\neq s}\bar{\mathbf{X}}^{(g)}\mathbf{L}_{gt}(\bar{\mathbf{X}}^{(t)})^T\mathbf{U}_{(t)}) \end{aligned} \tag{21}$$

Finally, Algorithm 1 gives the overall optimization for equation 9. In Step 3, we calculate the latent representation for the incomplete multi-view dataset. In Steps 4 and 5, we optimize the projection matrices $\mathbf{U}_{(g)}, (g = 1, ..., l)$. Finally Steps 3, 4 and 5 are repeated until convergence. Based on the latent representation, we can obtain the final clustering results directly based on the max value of each row or use regular clustering algorithms, e.g., k-means imposed on the latent representation. Besides, based on the learned projection matrix, we can project new multi-modal data to a common space to perform cross-modal retrieval.

Algorithm 1 Optimization for Equation 9

Input:
Incomplete dataset $\mathbf{X}$, parameters β and γ, and the number of classes.
1: Initialize $\mathbf{U}_{(g)}, (g = 1, ..., l)$ and $\mathbf{Y}$ randomly from $[0, 1]$;
2: **while** not converge **do**
3: Calculate $\mathbf{Y}^C$, $\hat{\mathbf{Y}}$ using Equation 16 and 17 respectively;
4: Solve $\mathbf{D}_{(g)}, (s = 1, ..., l)$ using Equation 20;
5: Calculate $\mathbf{U}_{(g)}, (g = 1, ..., l)$ using Equation 21 respectively;
6: **end while**
Output:
The latent representation and projection matrices for the incomplete multi-view dataset: $\mathbf{Y}$ and $\mathbf{U}_{(g)}, (g = 1, ..., l)$.

[1] $||\mathbf{U}_{(g)}(i,:)||$ can be zero, which cannot guarantee the convergence of the algorithm. Similar to [49], we add a smoothing term as in Equation 20.

ACCEPTED MANUSCRIPT

3.4. Convergence and complexity analysis

We prove the proposed iterative optimization strategy in Algorithm 1 will monotonically decrease the objective function in Equation 9 in each iteration until convergence.

3.4.1. Convergence for the indicator matrix

In Step 3 of Algorithm 1, we will resort to auxiliary function approach [47] to validate that the updating rule for $\mathbf{Y}^C$ will monotonically decrease the objective value.

Let

$$H(\mathbf{Y}^C) = Tr(\sum_g (-2\mathbf{A}_g^T \mathbf{Y}^C + (\mathbf{Y}^C)^T \mathbf{Y}^C) + \mathbf{\Gamma}((\mathbf{Y}^C)^T \mathbf{Y}^C - \mathbf{I}_k)) \tag{22}$$

and it is further rewritten as:

$$\begin{aligned} H(\mathbf{Y}^C) = Tr(\sum_g (2(\mathbf{A}_g^-)^T \mathbf{Y}^C + (\mathbf{Y}^C)^T \mathbf{Y}^C) + \mathbf{\Gamma}^+ (\mathbf{Y}^C)^T \mathbf{Y}^C \\ -Tr(\sum_g (2(\mathbf{A}_g^+)^T \mathbf{Y}^C + \mathbf{\Gamma}^- (\mathbf{Y}^C)^T \mathbf{Y}^C) \end{aligned} \tag{23}$$

Then the following function

$$\begin{aligned} & h(\mathbf{Y}^C, \tilde{\mathbf{Y}}^C) = \\ & \sum_{g,s,t} (\mathbf{A}_g^-(s,t) \frac{\mathbf{Y}^C(s,t)^2 + \tilde{\mathbf{Y}}^C(s,t)^2}{\tilde{\mathbf{Y}}^C(s,t)} + \frac{\tilde{\mathbf{Y}}^C(s,t)\mathbf{Y}^C(s,t)^2}{\tilde{\mathbf{Y}}^C(s,t)}) \\ & - \sum_{st} (\sum_g 2\mathbf{A}_g(s,t)) \tilde{\mathbf{Y}}^C(s,t)(1 + \log \frac{\mathbf{Y}^C(s,t)}{\tilde{\mathbf{Y}}^C(s,t)}) + \sum_{st} \frac{(\tilde{\mathbf{Y}}^C \mathbf{\Gamma}^+)(s,t) \mathbf{Y}^C(s,t)^2}{\tilde{\mathbf{Y}}^C(s,t)} \\ & - \sum_{gst} \mathbf{\Gamma}^-(s,t) \tilde{\mathbf{Y}}^C(g,s) \tilde{\mathbf{Y}}^C(g,t)(1 + \log \frac{\mathbf{Y}^C(g,s)\mathbf{Y}^C(g,t)}{\tilde{\mathbf{Y}}^C(g,s)\tilde{\mathbf{Y}}^C(g,t)}) \end{aligned} \tag{24}$$

is an auxiliary function of $H(\mathbf{Y}^C)$ (see in the appendix). Besides, it is easy to verify that the Hessian matrix of $h(\mathbf{Y}^C, \tilde{\mathbf{Y}}^C)$ is a positive definite matrix, thus, $h(\mathbf{Y}^C, \tilde{\mathbf{Y}}^C)$ is convex and its global minimum is obtained as in Equation 16.

Through the definition of the auxiliary function and the above derivation, we can obtain the following inequality:

$$H(\mathbf{Y}_0^C) = h(\mathbf{Y}_0^C, \mathbf{Y}_0^C) \geq h(\mathbf{Y}_0^C, \mathbf{Y}_1^C) \geq H(\mathbf{Y}_1^C) ... \tag{25}$$

Thus, the updating rule for $\mathbf{Y}^C$ will monotonically decrease the objective value.

3.4.2. Convergence for the projection matrix

In Step 5 of Algorithm 1, we will prove that the updating rule in Equation 21 for $\mathbf{U}_{(g)}, (g = 1, ..., l)$ decreases the objective monotonically.

Taking $\mathbf{U}_{(1)}$ as an example, we can derive that:

$$\begin{aligned}\mathbf{U}_{(1)}^{t+1} = \min_{\mathbf{U}_{(1)}} ||(\bar{\mathbf{X}}^{(1)})^T\mathbf{U}_{(1)} - \bar{\mathbf{Y}}^{(1)}||^2 + \beta tr(\mathbf{U}_{(1)}^T \mathbf{D}_{(1)}^{t+1}\mathbf{U}_{(1)}) \\ +\gamma \sum_{s=1}^{l} Tr(\mathbf{U}_{(1)}^T \bar{\mathbf{X}}^{(1)} \mathbf{L}_{1s} (\bar{\mathbf{X}}^{(s)})^T \mathbf{U}_{(s)})\end{aligned} \tag{26}$$

and Equation 21 is the analytic solution of the above function. Then we have:

$$z_{t+1} + \beta tr((\mathbf{U}_{(1)}^T)^{t+1}\mathbf{D}_{(1)}^{t+1}\mathbf{U}_{(1)}^{t+1}) \leq z_t + \beta tr((\mathbf{U}_{(1)}^T)^{t}\mathbf{D}_{(1)}^{t+1}\mathbf{U}_{(1)}^{t}) \tag{27}$$

where

$$z_{t+1} = ||(\bar{\mathbf{X}}^{(1)})^T\mathbf{U}_{(1)}^{t+1} - \bar{\mathbf{Y}}^{(1)}||^2 + \gamma \sum_{s=1}^{l} Tr(\mathbf{U}_{(1)}^T \bar{\mathbf{X}}^{(1)} \mathbf{L}_{1s} (\bar{\mathbf{X}}^{(s)})^T \mathbf{U}_{(s)}) \tag{28}$$

Substituting $\mathbf{D}_{(1)}^{t+1}$ into the above inequality, we have:

$$z_{t+1} + \sum_i \sum_j \frac{\mathbf{U}_{(1)}^{t+1}(i,j)\mathbf{U}_{(1)}^{t+1}(i,j)}{2||\mathbf{U}_{(1)}^{t}(i,:)||} \leq z_t + \sum_i \sum_j \frac{\mathbf{U}_{(1)}^{t}(i,j)\mathbf{U}_{(1)}^{t}(i,j)}{2||\mathbf{U}_{(1)}^{t}(i,:)||} \tag{29}$$

Here we introduce a function $f(x) = x - x^2/(2a)$, which satisfies $\{\forall x \in R, f(x) \leq f(a) | a > 0\}$. Then we make x and a be $||\mathbf{U}_{(1)}^{t+1}(i,:)||$ and $||\mathbf{U}_{(1)}^{t}(i,:)||$ respectively, we have the following inequality:

$$||\mathbf{U}_{(1)}^{t+1}(i,:)|| - \sum_j \frac{\mathbf{U}_{(1)}^{t+1}(i,j)\mathbf{U}_{(1)}^{t+1}(i,j)}{2||\mathbf{U}_{(1)}^{t}(i,:)||} \leq \sum_j ||\mathbf{U}_{(1)}^{t}(i,:)|| - \frac{\mathbf{U}_{(1)}^{t}(i,j)\mathbf{U}_{(1)}^{t}(i,j)}{2||\mathbf{U}_{(1)}^{t}(i,:)||} \tag{30}$$

Add both sides of the above inequality to Equation 29, we obtain the following inequality:

$$z_{t+1} + \beta ||\mathbf{U}_{(1)}^{t+1}||_{21} \leq z_t + \beta ||\mathbf{U}_{(1)}^{t}||_{21} \tag{31}$$

Thus the updating rule for $\mathbf{U}$ will decrease the objective function monotonically.

Combining the above derivations, we prove that Algorithm 1 converges to a local minimum.

3.4.3. Complexity analysis

We briefly discuss the computational complexity of our algorithm. As for the optimization of $\mathbf{Y}$, the main computation lies in the updating for $\mathbf{Y}^C$ as in Equation 16, which mainly consists of some matrix multiplication operations. When optimizing $\mathbf{U}$, we need to compute the overall multi-view similarity matrix, whose complexity is about $O(d_g N_g^2)$, where $d_g N_g^2$ being the product of the dimensionality and the square of the number of examples for the g-th view is the largest one among all views. However, it is a constant matrix and can be computed before the optimization of the variables. Besides, we need to use Equation

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

21 to calculate $\mathbf{U}$, which solves an inverse problem. Instead, we can update the projection matrices by solving a linear system for $O(\hat{d}^2)(\hat{d} = \max(d_1, ..., d_l))$ complexity.

4. Experiments

4.1. Datasets

Seven public datasets are utilized, and their statistics are listed in Table 2.

Table 2: Information of the multi-view datasets.

Dataset	USPS	Cora	BBC	3Source	VOC	Wiki	NUS
# size	2,000	2,708	2,012	169	9,963	2,866	60,960
# cluster	10	7	5	6	20	10	10
# view	2	2	2	3	2	2	2
# feature size	76+216	2708+1433	6838+6790	3560+3631+3068	512+399	128+10	500+1000

USPS Dataset[2] It consists of feature sets of handwritten numerals (0-9) extracted from Dutch utility maps. The database has 2,000 examples even-distributed in ten categories and is represented in terms of six visual features. Being same in [50], we use the 76 Fourier coefficients and the 216 profile correlations as two views.

Cora Dataset[3] It contains 2,708 scientific publications divided into 7 classes. Two heterogeneous feature sets, i.e., citations and content are utilized here for experiments, where the content feature is represented by 0/1-valued word vector indicating the absence/presence of the corresponding word from a constructed dictionary.

BBC Dataset[4] It is a synthetic multi-view text database, which is constructed using single view BBC and BBCSport corpora. In total, it consists of 2,012 data examples categorized into 5 classes. The two views used here are the segments representations of the same document with the dimensionalities being 6,838 and 6,790 respectively.

[2]http://archive.ics.uci.edu/ml/datasets/Multiple+Features.
[3]http://lig-membres.imag.fr/grimal/data.html.
[4]http://mlg.ucd.ie/datasets/segment.html.

3Source dataset[5]It is constructed using three well-known online news sources, i.e., BBC, Reuters and the Guardian. In total, there are 416 distinct news divided into six categories. Among them, 169 news are reported by all the three sources and are used as in [8] with each source serving as one view.

VOC Dataset[6] It consists of 5,011 training and 4,952 testing image-tag pairs categorized into 20 classes. We use the 512-dimensional Gist features and 399-dimensional word frequency features here. Some of the pairs are multi-labeled, so we select those with only one label. Besides, those tag features with only zeros are deleted. Finally, we have 2,799 training and 2,820 testing pairs.

Wiki Dataset[7] It is a widely used dataset for cross-modal retrieval, which consists of 2,173/693 training/testing image-text pairs divided into 10 categories. In each pair, the image is encoded by the 128 dimensional SIFT descriptors and the text is 10 dimensional topics derived from a Latent Dirichlet Allocation model.

NUS-WIDE Dataset[8] It is collected from Flickr and consists of 270k images in 81 categories. The images are represented by 500 dimensional SIFT descriptors together with an 1,000 dimensional textual feature constructed by the tags annotated to the images. Similar to [51], we select the pairs that belong to one of the 10 largest classes as a subset for evaluation, which results in 60k image-text pairs.

4.2. Settings

Similar to [30], two different settings are considered and listed as follows. **The first setting**: features from all the views are incomplete. **The second setting**: at least one view is complete. For the above two settings, we randomly select 10% to 90% of the total examples, with 20% as interval, to have incomplete feature set. And this process is repeated 10 times with the average to be

[5]http://mlg.ucd.ie/datasets/3sources.html.
[6]http://www.pascalnetwork.org/challenges/voc/voc2007/workshop/index.html.
[7]http://www.svcl.ucsd.edu/projects/crossmodal/.
[8]http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

reported. In the first setting, we evenly distribute the number of examples with incomplete views for simplicity [30].

4.3. Multi-view clustering

4.3.1. Compared methods and settings

SingleV1, SingleV2: We run spectral clustering [52] on the two views under the condition that all views have complete data examples. **CCA:** We use the canonical correlation analysis to obtain the latent representation of multi-view data and then apply k-means on the obtained representation. **PairwiseSC, CentroidSC:** Two regularization frameworks developed by Kumar et al. [50] for multi-view spectral clustering. **MultiCF:** Wang et al. [20] proposed a structure sparsity based multi-view clustering method. **RMSC:** Xia et al. [26] developed a multi-view spectral clustering method, which is based on low rank and sparse decomposition of the transition matrix. **PVC:** Li et al. [30] proposed a non-negative matrix factorization based incomplete multi-view clustering method. **PairwiseSC++, CentroidSC++, RMSC++:** We denote the PairwiseSC, CentroidSC and RMSC methods with the preprocessing of the kernel matrix under the two settings using [28, 29] as PairwiseSC++, CentroidSC++, RMSC++ respectively.

For the compared methods without preprocessing of kernel matrices, we use zeros to replace incomplete feature sets. This may be a little arbitrary, but we find possibly no methods can well fill various types of features at the same time, e.g., visual features and textual features. Besides, it may be fair enough since our method does not preprocess the data at all. For our method, we use Gaussian kernel to construct the intra-view similarity matrix, where the neighbors number (m) and the width parameter (σ) in Equation 4 are empirically selected as ten percent of the dataset size and 1 respectively in all the experiments. Since k-means is used in all the experiments, it is run 20 times with random initialization and the mean value is reported.

Following [30], the normalized mutual information (NMI), one of the most famous clustering evaluation measures, is utilized [53]. Usually, the larger the

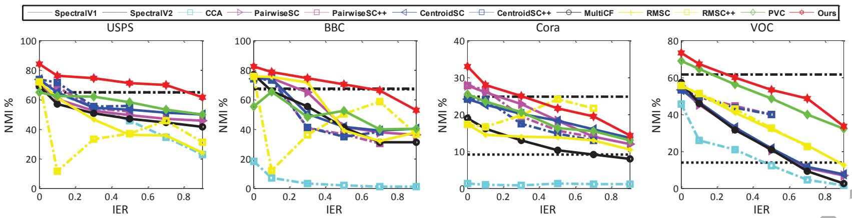

Figure 2: The NMI results on the four databases when both views suffer from the loss of examples. IER (incomplete example ratio) is the ratio of examples with only one feature set.

NMI, the better the clustering performance.

4.3.2. Experimental results

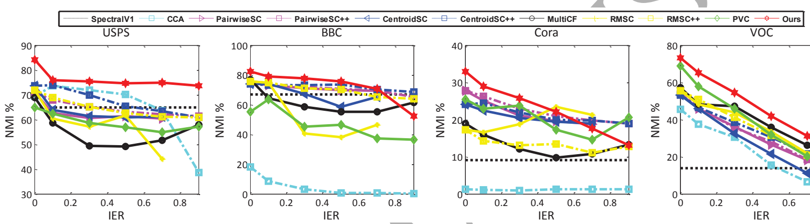

Figure 3: The NMI results on the four databases when the first view suffer from the loss of examples.

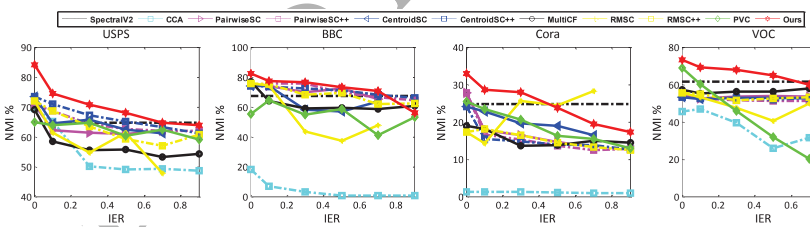

Figure 4: The NMI results on the four databases when the second view suffer from the loss of examples.

Figure 2 display the clustering performance on the two-view datasets USPS, BBC, Cora and VOC under the first setting, and Figures 3 and 4 show the results under the second setting with the first and second view suffering from incomplete examples respectively. *IER* (incomplete example ratio) indicates the percentage of examples having only one feature set. Besides, the results of all methods with *IER* being zero are also reported as the upper bound of each method. Comparing the three figures, It can be seen that similar results are

ACCEPTED MANUSCRIPT

obtained for the two settings. Overall, our method performs better than all the competing methods under different settings on the four databases.

As for PVC, it uses non-negative matrix factorization to find a unified low dimensional space. Compared with it, we also apply feature selection to select relevant features when learning the low dimensional subspace, which works confronting the high dimensional and noisy features. Besides, the multi-view data similarities are also explored in the proposed method. Thus our method performs better than PVC. One of the major differences between our method and the MultiCF method under complete views is the constraint imposed on the learned latent representation. We add the non-negative constraint, which is more reasonable to approach the class indicator matrix. It may be the reason that our method performs better when the incomplete example ratio is zero. Since MultiCF is not designed for incomplete multi-view data, our method also outperforms it when *IER* is greater than zero.

As for PairwiseSC, CentroidSC and RMSC, we utilize the method proposed in [29] to fill the kernel matrices of the incomplete views and accordingly PairwiseSC++, CentroidSC++, RMSC++ are developed in the first setting. From Figure 2, they perform better than their original ones in some databases and the performance gain seems not very significant especially when *IER* being large. Furthermore, we apply the method in [28] to fill the kernel matrix of the incomplete view using that of the complete views for PairwiseSC, CentroidSC and RMSC to obtain the PairwiseSC++, CentroidSC++, RMSC++ methods under the second setting. It can be seen the modified methods obtain relatively better performance compared with the original ones due to the use of at least one complete view. In summary, our method performs better although these kernel based multi-view clustering methods are preprocessed.

Finally, we conduct experiments on a more than two-view dataset, i.e., 3Source dataset. In the first setting, examples with incomplete views are enforced to have only one feature set for simplicity. In the second setting, examples with incomplete views are evenly distributed. The results are displayed in Figure 7. It can be seen that similar results are obtained as in other datasets, which

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

further validates our method. It should be noted that the method proposed in [29] cannot deal with more than two-view data, so there are no results of PairwiseSC++, CentroidSC++ and RMSC++ under the first setting. As for the second setting, the method developed in [28] can handle three or more views, so the results of modified versions of PairwiseSC, CentroidSC and RMSC, i.e., PairwiseSC++, CentroidSC++ and RMSC++, are displayed.

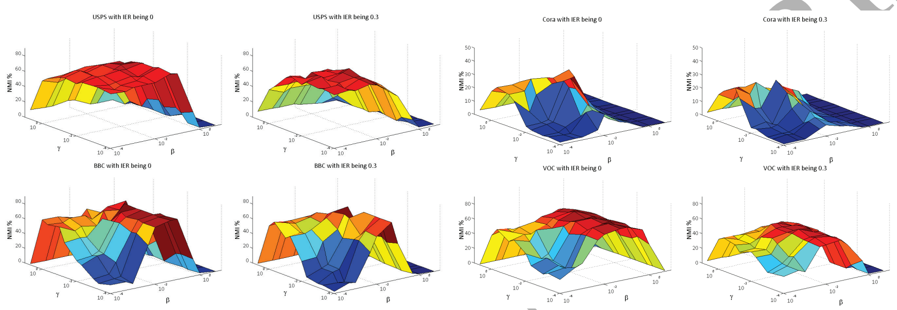

Figure 5: The NMI results on the four databases under the first setting with the IER being 0 and 0.3 respectively.

4.3.3. Parameter selection

In our model, β and γ balance the effect of feature projection term, $\ell 21$-norm based feature selection term and graph regularization based similarity preserving term. In this section, we investigate how the performance varies with the changes of the above two parameters. The results are shown in Figure 5. When β is small, the regularizer will lose the effect of feature selection. In the case when β is too big, the sparse characteristic will lead to the loss of useful features and harm the learned latent representations. As for γ, when it is too big, it may rely on too much of the neighborhood relationship obtained using the similarity metric and this may harm the intrinsical data structure because of the possible inaccuracy of the calculated similarity matrix. In summary, β and γ should be carefully selected and [0.001,0.01] is an optimal interval when the multi-view data are normalized.

ACCEPTED MANUSCRIPT

4.3.4. Convergence study

As discussed in previous section, the optimization strategy converges to a local minima. In this section, we give the convergence and the corresponding NMI curves with the varying updating iterations. Due to space limitation, we only give the results under the first setting with incomplete example ratio being 30% and similar results can be achieved under the second setting. From Figure 6, it can be seen that the objective function converges fast, and the clustering performance needs about 100 iterations to reach the best results. This may because the initial values of the variables in Algorithm 1 are randomly set. In the future, we may consider a better initialization method to reduce the number of iterations.

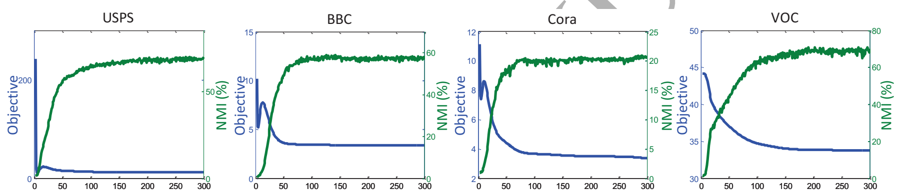

Figure 6: Convergence and the corresponding NMI curves for the four databases under the first setting with IER being 0.3.

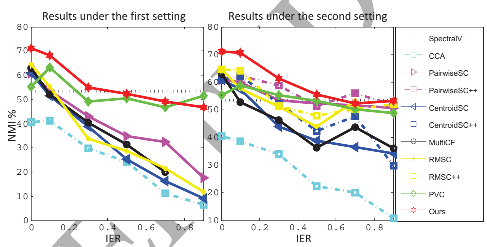

Figure 7: Results on the 3Source dataset.

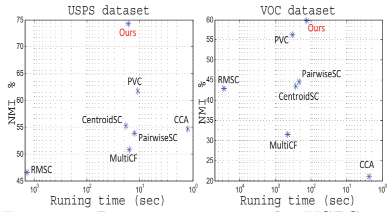

Figure 8: Running time on the USPS and VOC datasets under the first setting with IER being 0.3.

4.3.5. Running time

We show the running time for obtaining the subspaces of all the methods on the USPS and VOC datasets, where all the methods are run on the same machine (Intel CPU 3.1GHz and 12 GB memory). All the methods are implemented using MATLAB except that the main parts of PVC is C++ (provided

ACCEPTED MANUSCRIPT

by their authors). The experimental results are shown in Figure 8. It can be seen that our method obtains the best results, and the time used is in the same magnitudes with the mainstream methods. For methods PairwiseSC and CentroidSC, eigenvalue decomposition needs to be performed in every iteration. For RMSC, the Augmented Lagrangian Multiplier is utilized for optimization, which brings more auxiliary variables and is thus time consuming. For our method, not very large number of iterations is needed for acceptable results, which is shown in Figure 6.

4.4. Cross-modal retrieval

We conduct experiments on the VOC, Wiki and NUS WIDE datasets. For the VOC dataset, we follow the natural training and testing split criterion. For the Wiki database, similar to [51], we split it into a training set of 1,300 pairs and a testing set of 1,566 pairs. For the NUS WIDE database, we take 50% of total points as the training set and the remaining as the testing set.

To evaluate the performance of our method, we conduct two cross-modal retrieval tasks, i.e., Image query vs. Text database and Text query vs. Image database. More specifically, we map the testing multi-modal data into the common space, and then take one modality of the testing data as the query set to retrieve another modality. Finally, the cosine distance is utilized to measure the similarity between different modalities.

4.4.1. Compared methods and settings

We compare our method with the following representative cross-modal retrieval methods, i.e., **PLS** [38], **BLM** [40], **CCA** [7], **CDFE** [42], **GMLDA** [39], **GMMFA** [39], **CorrAE** [2] and **DCCAE** [54]. Among them, PLS, BLM and CCA are classical unsupervised methods that use pairwise information for the common latent space learning. CorrAE and DCCAE are typical deep learning methods that jointly learn high level features and cross-modal matching between different modalities. CDFE, GMLDA and GMMFA are three typical supervised methods that utilize label information. Different from unsupervised

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

methods, those methods can obtain relatively discriminative subspaces due to the guidance of labels.

For CDFE, GMLDA, GMMFA and DCCAE, we use the codes the authors released, and for PLS, BLM, CCA and CorrAE, we obtain their results based on suggestions described by the papers. As for our method, similar parameter settings are designed as in multi-view clustering, namely, we use KNN based Gaussian kernel to construct the intra-view similarity matrix and the number of the KNN neighbors and the width parameter for the Gaussian kernel are empirically selected as ten percent of the total examples of the database and one respectively in all the experiments. As for the trade off parameters β and γ, they are empirically selected to achieve the best results.

We use mean average precision (MAP) to evaluate the overall performance, which is one of the most popular metrics for retrieval tasks. Usually, the larger the MAP, the better the retrieval performance. Besides the MAP, we use precision-recall curve to further evaluate the effectiveness of different methods. For their detailed definition, readers can refer to [55].

Table 3: MAP under different incomplete example ratios on the VOC datasets. I, T and M represent Image query, Text query and Mean result, respectively.

Methods	0% IER			30% IER of I+T			30% IER of I			30% IER of T		
	I	T	M	I	T	M	I	T	M	I	T	M
PLS	27.6	20.0	23.8	27.4	19.9	23.7	27.6	19.7	23.7	27.0	19.8	23.4
BLM	30.6	23.1	26.9	30.1	22.5	26.3	30.3	22.5	26.4	29.8	22.4	26.1
CCA	26.7	22.2	24.5	25.3	21.6	23.5	25.1	21.2	23.2	25.0	20.6	22.8
CorrAE	26.4	23.8	25.1	27.6	21.3	24.5	27.6	21.5	24.6	26.9	20.2	23.6
DCCAE	24.2	20.1	22.2	22.3	19.1	20.7	20.2	19.5	19.9	23.6	18.5	21.1
CDFE	30.0	22.5	26.3	28.1	20.6	24.4	27.7	20.4	24.1	27.9	20.6	24.3
GMLDA	31.1	24.6	27.9	28.6	22.6	25.6	28.5	22.6	25.6	28.7	23.2	26.0
GMMFA	30.6	24.3	27.5	28.1	22.1	25.1	27.9	21.9	24.9	27.6	21.5	24.6
Ours	37.5	29.7	33.6	35.9	27.5	31.7	36.5	28.1	32.3	35.0	26.0	30.5

4.4.2. Results on the VOC dataset

Since methods PLS, BLM, CCA, CDFE, GMLDA and GMMFA mainly focus on learning the latent subspaces and perform no feature selection, we utilize Principal Component Analysis (PCA) to remove the redundancy in the original features as did in [51], which shows better results than the one without conducting PCA. CorrAE, DCCAE and our method performs feature learning and subspace learning simultaneously, so we do not use PCA as preprocessing.

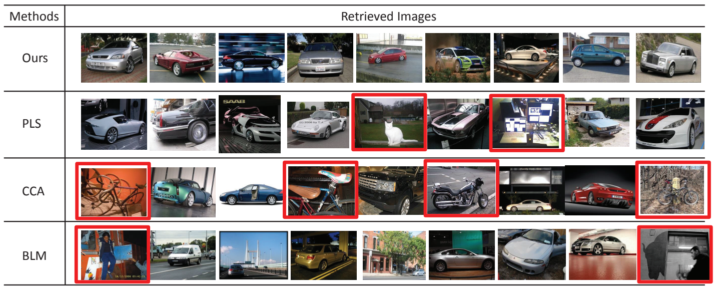

Figure 9: Cross-modal retrieval using text query (car+window+tire+rims) on the Pascal VOC dataset. Red rectangles indicate incorrect retrieval results.

Table 3 gives the results of MAP under the two settings, where the incomplete example ratios are 0% and 30%. Overall, our algorithm outperforms all the compared methods under all the settings. BLM, CCA and PLS are unsupervised methods, compared with them, we conduct feature selection and consider similarity preserving. Compared with CorrAE and DCCAE, our method learns the class information and preserves the inter-view and intra-view data structure, thus our method performs better. Though CDFE, GMLDA and GMMFA are supervised cross-modal retrieval methods, our model outperforms them. This may be because our algorithm can learn the class indicator matrix, which in turn guides the learning of the subspace.

We also give the precision-recall curves for image query and text query under the two settings with the incomplete example ratio being 0.3, which are shown in Figure 10. Overall, it can be seen that our method performs better than all the compared methods. Figure 9 shows an example of the top nine retrieved images by three unsupervised methods, i.e., CCA, PLS, BLM and our method using the tags "car+window+tire+rims".

4.4.3. Results on the Wiki dataset

Since the dimensionalities of images and texts on the Wiki dataset are low, PCA is not utilized for the compared methods as did in [51]. Table 4 gives the MAP scores with incomplete example ratios being 0 and 0.3 under the two settings. Overall, our method outperforms all the compared methods as did

ACCEPTED MANUSCRIPT

Table 4: MAP under different incomplete example ratios on the Wiki datasets. I, T and M represent Image query, Text query and Mean result, respectively.

Methods	0% IER			30% IER of I+T			30% IER of I			30% IER of T		
	I	T	M	I	T	M	I	T	M	I	T	M
PLS	24.0	16.3	20.2	22.4	16.3	19.4	23.4	16.2	19.8	23.6	16.3	20.0
BLM	25.7	20.4	23.1	25.3	19.8	22.6	25.6	20.0	22.8	25.7	20.6	23.2
CCA	26.3	20.7	23.5	23.5	18.8	21.2	23.5	18.7	21.1	24.1	19.00	21.6
CorrAE	25.4	20.4	22.9	25.3	20.5	22.9	25.1	19.8	22.5	25.3	19.9	22.6
DCCAE	24.2	20.2	22.2	23.7	19.9	21.8	21.3	18.7	20.0	24.2	19.6	21.9
CDFE	26.9	20.6	23.8	25.6	19.3	22.5	25.00	18.2	21.6	26.4	19.4	22.9
GMLDA	27.4	21.2	24.3	25.9	20.1	23.0	26.0	20.4	23.2	26.8	20.3	23.6
GMMFA	27.4	21.7	24.6	25.8	20.0	22.9	25.9	20.4	23.2	26.8	20.7	23.8
Ours	28.2	22.3	25.3	27.7	21.6	24.7	27.9	22.4	25.2	27.6	22.0	24.8

on the VOC database. Similarly, Figure 11 shows the precision-recall curves of 30% incomplete example ratio under the two settings, which further validates the advantages of our method. It should be noted that similar results can be obtained under other IERs, but we omit them due to space limitation.

4.4.4. *Results on the NUS WIDE dataset*

Similar to the pre-processing of VOC dataset, Principal Component Analysis (PCA) is conducted on the original features. Table 5 shows the results of all methods under the two settings. It can be seen that our method performs better than all the unsupervised algorithms, i.e., CCA, BLM, PLS, CorrAE and DCCAE. As for the two popular supervised methods, i.e., GMMFA and GMLDA, our algorithm obtains similar results. Compared with them, we use no labels, which shows the advantages than the supervised methods. Finally, the precision-recall curves in Figure 12 further validate the above results.

Table 5: MAP under different incomplete example ratios on the NUS 60k datasets. I, T and M represent Image query, Text query and Mean result, respectively.

Methods	0% IER			30% IER of I+T			30% IER of I			30% IER of T		
	I	T	M	I	T	M	I	T	M	I	T	M
PLS	46.9	45.5	46.2	47.1	46.6	46.9	46.6	45.0	45.8	46.2	44.9	45.6
BLM	50.3	49.4	49.9	50.3	49.3	49.8	49.5	48.5	49.0	49.5	48.4	49.0
CCA	47.8	47.0	47.4	47.4	46.6	47.0	46.9	46.1	46.5	46.7	45.9	46.3
CorrAE	49.4	48.5	49.0	48.4	48.0	48.2	47.2	47.7	47.5	46.9	48.7	47.8
DCCAE	51.2	48.7	50.0	50.3	47.9	49.1	48.8	47.2	48.0	49.5	47.1	48.3
CDFE	44.9	46.4	45.7	44.9	45.4	45.2	44.1	44.1	44.1	43.4	43.6	43.5
GMLDA	52.5	50.5	51.5	52.2	50.2	51.2	51.7	50.0	50.9	51.8	49.8	50.8
GMMFA	49.8	49.2	49.5	50.2	49.4	49.8	50.9	49.4	50.2	51.0	49.3	50.2
Ours	51.2	53.0	52.1	50.6	52.8	51.7	50.9	51.2	51.1	50.9	51.0	51.0

ACCEPTED MANUSCRIPT

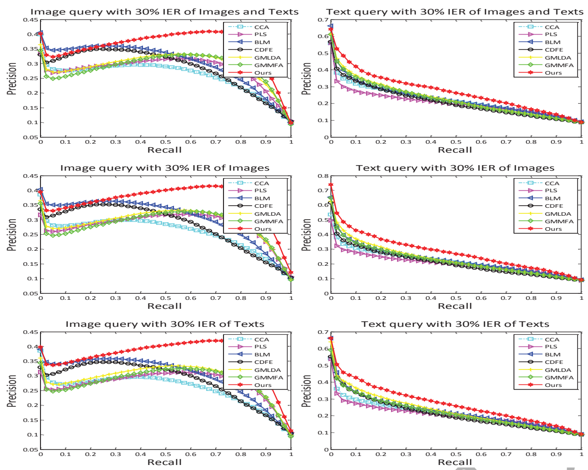

Figure 10: Precision-recall curves on the VOC datasets.

5. Conclusion and future work

In this paper, we have proposed a novel subspace learning framework for incomplete and unlabeled multi-view data. In our modal, we directly learn the class indicator matrix, which serves as a latent space for bridging heterogeneous feature sets. By utilizing all data samples in a view to learn the projection matrix and making data examples consisting of complete feature sets to learn the shared class indicator matrix, the proposed model can well use the incomplete data. Furthermore, feature selection and inter-view and intra-view data similarities are considered to enhance our framework. To these ends, an objective is developed with an efficient optimization strategy and convergence analysis. Extensive experiments including multi-view clustering and cross-modal retrieval have validated our method compared with the state-of-the-art methods.

In real applications, it may be easy to obtain some supervised or weak supervised information, such as partial labels and the pairwise relationship (must-link and cannot-link) between two data samples. This knowledge, serving as the true semantic information, can guide the learning of unsupervised multi-view data. In the future, we may consider adding such information to promote the learning of incomplete and unlabeled multi-view data.

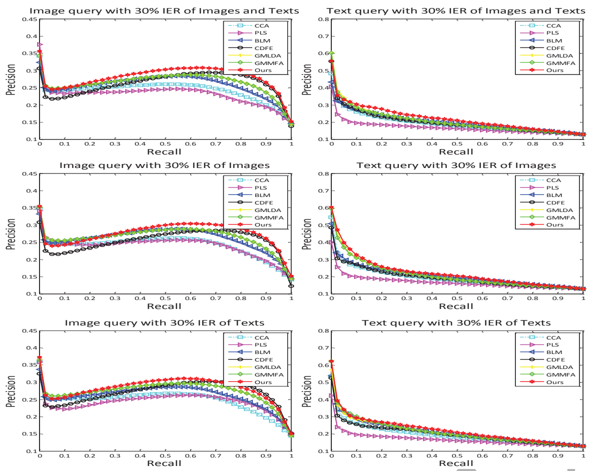

Figure 11: Precision-recall curves on the Wiki datasets.

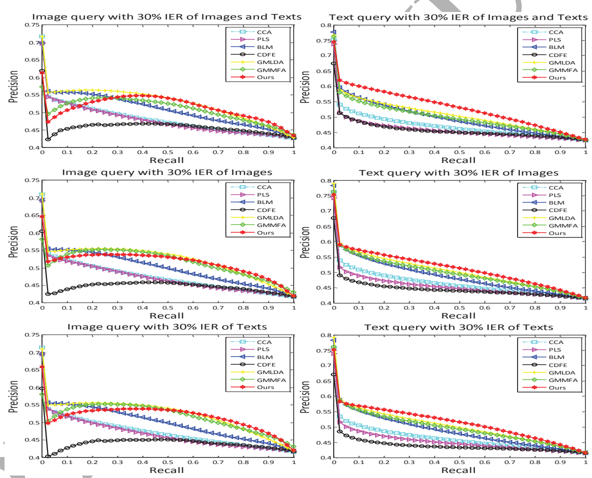

Figure 12: Precision-recall curves on the NUS datasets.

6. Acknowledgements

This work was supported by the state key development program [2016YFB1001000]; National Natural Science Foundation of China [61403390, U1435221].

References

[1] S. Sun, A survey of multi-view machine learning, Neural Computing and Applications 23 (7-8) (2013) 2031–2038.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

[2] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder, ACM International Conference on Multimedia (2014) 7–16.

[3] C. H. Lampert, O. Krmer, Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning, European Conference on Computer Vision (2010) 566–579.

[4] X. Chen, S. Chen, H. Xue, X. Zhou, A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data, Pattern Recognition 45 (5) (2012) 2005–2018.

[5] C. Xu, D. Tao, C. Xu, Multi-view learning with incomplete views, IEEE Transactions on Image Processing 24 (12) (2015) 5812–5825.

[6] K. Chaudhuri, S. M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, International Conference on Machine Learning (2009) 129–136.

[7] T.-K. Kim, J. Kittler, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (6) (2007) 1005–1018.

[8] J. Liu, C. Wang, J. Gao, J. Han, Multi-view clustering via joint nonnegative matrix factorization, SIAM International Conference on Data Mining (2013) 252–260.

[9] X. He, M.-Y. Kan, P. Xie, X. Chen, Comment-based multi-view clustering of web 2.0 items, International Conference on World Wide Web (2014) 771–782.

[10] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, Annual Conference on Computational Learning Theory (1998) 92–100.

[11] S. Yu, B. Krishnapuram, R. Rosales, R. B. Rao, Bayesian co-training, Journal of Machine Learning Research 12 (2011) 2649–2680.

[12] M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, Journal of Machine Learning Research 12 (2011) 2211–2268.

[13] F. R. Bach, M. I. Jordan, A probabilistic interpretation of canonical correlation analysis, University of California, Berkeley, Tech. Rep.

[14] O. Arandjelovi, Discriminative extended canonical correlation analysis for pattern set matching, Machine Learning 94 (3) (2014) 353–370.

[15] O. Arandjelovic, R. Cipolla, Face set classification using maximally probable mutual modes, International Conference on Pattern Recognition (2006) 511–514.

[16] C.-D. Wang, J.-H. La, P. S. Yu, Multi-view clustering based on belief propagation, IEEE Transactions on Knowledge and Data Engineering 28 (4) (2016) 1007–1021.

[17] J. Xu, J. Han, F. Nie, Discriminatively embedded k-means for multi-view clustering, IEEE Conference on Computer Vision and Pattern Recognition (2016) 5356–5364.

[18] Y.-M. Xu, C.-D. Wang, J.-H. Lai, Weighted multi-view clustering with feature selection, Pattern Recognition 53 (2016) 25–35.

[19] Z. Zhang, Z. Zhai, L. Li, Uniform projection for multi-view learning, IEEE Transactions on Pattern Analysis and Machine Intelligence (2016).

[20] H. Wang, F. Nie, H. Huang, Multi-view clustering and feature learning via structured sparsity, International Conference on Machine Learning (2013) 352–360.

[21] F. Nie, W. Zhu, X. Li, Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification, International Joint Conference on Artificial Intelligence (2016) 1881–1308.

ACCEPTED MANUSCRIPT

[22] S. Bickel, T. Scheffer, Multi-view clustering, International Conference on Data Mining (2004) 19–26.

[23] A. Kumar, H. Daumé, A co-training approach for multi-view spectral clustering, International Conference on Machine Learning (2011) 393–400.

[24] E. Bruno, S. Marchand-Maillet, Multiview clustering: a late fusion approach using latent models, ACM SIGIR Conference on Research and Development in Information Retrieval (2009) 736–737.

[25] S. F. Hussain, M. Mushtaq, Z. Halim, Multi-view document clustering via ensemble method, Journal of Intelligent Information Systems 43 (1) (2014) 81–99.

[26] R. Xia, Y. Pan, L. Du, J. Yin, Robust multi-view spectral clustering via low-rank and sparse decomposition, AAAI Conference on Artificial Intelligence (2014) 2149–2155.

[27] X. Cao, C. Zhang, H. Fu, S. Liu, H. Zhang, Diversity-induced multi-view subspace clustering, IEEE Conference on Computer Vision and Pattern Recognition (2015) 586–594.

[28] P. Rai, A. Trivedi, H. Daumé III, S. L. DuVall, Multiview clustering with incomplete views, NIPS Workshop on Machine Learning for Social Computing (2010).

[29] W. Shao, X. Shi, P. S. Yu, Clustering on multiple incomplete datasets via collective kernel learning, International Conference on Data Mining (2013) 1181–1186.

[30] S. Li, Y. Jiang, Z. Zhou, Partial multi-view clustering, AAAI Conference on Artificial Intelligence (2014) 1968–1974.

[31] W. Shao, L. He, P. S. Yu, Multiple incomplete views clustering via weighted nonnegative matrix factorization with l21 regularization, European Conference on Principles of Data Mining and Knowledge Discovery (2015) 318–334.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

[32] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Pl-ranking: A novel ranking method for cross-modal retrieval, ACM on Multimedia (2016) 1355–1364.

[33] D. M. Blei, M. I. Jordan, Modeling annotated data, ACM SIGIR conference on Research and development in informaion retrieval (2003) 127–134.

[34] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, Y. Zhuang, A low rank structural large margin method for cross-modal ranking, ACM SIGIR Conference on Research and Development in Information Retrieval (2013) 433–442.

[35] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, Y. Zhuang, Cross-media semantic representation via bi-directional learning to rank, ACM International Conference on Multimedia (2013) 877–886.

[36] K. Li, G. Qi, J. Ye, K. Hua, Linear subspace ranking hashing for cross-modal retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence (2016).

[37] D. Wang, X. Gao, X. Wang, L. He, B. Yuan, Multimodal discriminative binary embedding for large-scale cross-modal retrieval, IEEE Transactions on Image Processing 25 (10) (2016) 4540–4554.

[38] R. Rosipal, N. Krämer, Overview and recent advances in partial least squares, Subspace, Latent Structure and Feature Selection (2006) 34–51.

[39] A. Sharma, A. Kumar, H. Daume III, D. W. Jacobs, Generalized multiview analysis: A discriminative latent space, IEEE Conference on Computer Vision and Pattern Recognition (2012) 2160–2167.

[40] J. B. Tenenbaum, W. T. Freeman, Separating style and content with bilinear models, Neural Computation 12 (6) (2000) 1247–1283.

[41] C. Deng, X. Tang, J. Yan, W. Liu, G. Xinbo, Discriminative dictionary learning with common label alignment for cross-modal retrieval, IEEE Transactions on Multimedia 18 (2) (2016) 208–218.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

[42] D. Lin, X. Tang, Inter-modality face recognition, European Conference on Computer Vision (2006) 13–26.

[43] C. Kang, S. Xiang, S. Liao, C. Xu, C. Pan, Learning consistent feature representation for cross-modal multimedia retrieval, IEEE Transactions on Multimedia 45 (5) (2012) 2005–2018.

[44] Q. Yin, S. Wu, L. Wang, Incomplete multi-view clustering via subspace learning, ACM International on Conference on Information and Knowledge Management (2015) 383–392.

[45] F. Nie, H. Huang, X. Cai, C. H. Ding, Efficient and robust feature selection via joint l21-norms minimization, Advances in Neural Information Processing Systems (2010) 1813–1821.

[46] Z. Li, J. Liu, Y. Yang, X. Zhou, H. Lu, Clustering-guided sparse structural learning for unsupervised feature selection, IEEE Transactions on Knowledge and Data Engineering 26 (9) (2014) 2138–2150.

[47] C. Ding, T. Li, M. I. Jordan, Convex and semi-nonnegative matrix factorizations, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (1) (2010) 45–55.

[48] J. Tang, X. Hu, H. Gao, H. Liu, Unsupervised feature selection for multi-view data in social media, SIAM International Conference on Data Mining (2013) 270–278.

[49] I. F. Gorodnitsky, B. D. Rao, Sparse signal reconstruction from limited data using focuss: a re-weighted minimum norm algorithm, IEEE Transactions on Signal Processing 45 (3) (1997) 600–616.

[50] A. Kumar, P. Rai, H. Daume, Co-regularized multi-view spectral clustering, Advances in Neural Information Processing Systems (2011) 1413–1421.

[51] K. Wang, R. He, W. Wang, L. Wang, T. Tan, Learning coupled feature spaces for cross-modal matching, IEEE International Conference on Computer Vision (2013) 2088–2095.

ACCEPTED MANUSCRIPT

[52] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 888–905.

[53] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, E. Y. Chang, Parallel spectral clustering in distributed systems, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (3) (2011) 568–586.

[54] W. Wang, R. Arora, K. Livescu, J. Bilmes, On deep multi-view representation learning: Objectives and optimization, arXiv:1602.01024v1.

[55] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, International Conference on Multimedia (2010) 251–260.

Appendix

We prove that Equation 24 is an auxiliary function of $H(\mathbf{Y}^C)$. By the following inequality,

$$2a \leq \frac{a^2+b^2}{b}, \quad \forall a \geq 0, b \geq 0 \tag{32}$$

then,

$$Tr(\sum_g 2\mathbf{A}_g^-(\mathbf{Y}^C)^T) = \sum_{gst} 2\mathbf{A}_g^-(s,t)\mathbf{Y}^C(s,t) \leq \sum_{gst} (\mathbf{A}_g^-(s,t)\frac{\mathbf{Y}^C(s,t)^2 + \tilde{\mathbf{Y}}^C(s,t)^2}{\tilde{\mathbf{Y}}^C(s,t)}) \tag{33}$$

It is easy to obtain the following inequality,

$$Tr(\sum_g (\mathbf{Y}^C)^T\mathbf{Y}^C + \mathbf{\Gamma}^+(\mathbf{Y}^C)^T\mathbf{Y}^C) \leq \sum_{gst} \frac{\tilde{\mathbf{Y}}^C(s,t)\mathbf{Y}^C(s,t)^2}{\tilde{\mathbf{Y}}^C(s,t)} + \sum_{st} \frac{(\tilde{\mathbf{Y}}^C\mathbf{\Gamma}^+)(s,t)\mathbf{Y}^C(s,t)^2}{\tilde{\mathbf{Y}}^C(s,t)} \tag{34}$$

Due to $z \geq 1 + \log z, \forall z \geq 0$, we have:

$$\begin{aligned} -Tr(\sum_g 2\mathbf{A}_g^+(\mathbf{Y}^C)^T + \mathbf{\Gamma}^-(\mathbf{Y}^C)^T\mathbf{Y}^C) \leq & -\sum_{st}(\sum_g 2\mathbf{A}_g^+(s,t))\tilde{\mathbf{Y}}^C(s,t)(1+\log\frac{\mathbf{Y}^C(s,t)}{\tilde{\mathbf{Y}}^C(s,t)}) \\ & -\sum_{gst}\mathbf{\Gamma}^-(s,t)\tilde{\mathbf{Y}}^C(g,s)\tilde{\mathbf{Y}}^C(g,t)(1+\log\frac{\mathbf{Y}^C(g,s)\mathbf{Y}^C(g,t)}{\tilde{\mathbf{Y}}^C(g,s)\tilde{\mathbf{Y}}^C(g,t)}) \end{aligned} \tag{35}$$

By summing the above equations, we have $h(\mathbf{Y}^C, \tilde{\mathbf{Y}}^C) \geq H(\mathbf{Y}^C)$ and $h(\mathbf{Y}^C, \mathbf{Y}^C) = H(\mathbf{Y}^C)$. Thus, Equation 24 is an auxiliary function of $H(\mathbf{Y}^C)$.

Qiyue Yin received the B.S. degree in automation control from Harbin Engineering University, Harbin, China in 2012. He is currently pursuing the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition and computer vision.

Shu Wu received the Ph.D. degree in computer science from University of Sherbrooke, Canada, in 2012. He is an assistant professor in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include data mining, recommendation systems, and pervasive computing.

Liang Wang received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2004. Currently, he is a full Professor of Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition and computer vision.

ACCEPTED MANUSCRIPT