Incomplete Multi-view Clustering via Graph Regularized Matrix Factorization

Jie Wen[1] ⋆, Zheng Zhang[1,2⋆], Yong Xu[1†], and Zuofeng Zhong[1]

[1] Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen, Shenzhen 518055, Guangdong, China
[2] The University of Queensland, Australia
wenjie@hrbeu.edu.cn; darrenzz219@gmail.com; yongxu@ymail.com; zfzhong2010@gmail.com

Abstract. Clustering with incomplete views is a challenge in multi-view clustering. In this paper, we provide a novel and simple method to address this issue. Specially, the proposed method simultaneously exploits the local information of each view and the complementary information among views to learn the common latent representation for all samples, which can greatly improve the compactness and discriminability of the obtained representation. Compared with the conventional graph embedding methods, the proposed method does not introduce any extra regularization term and corresponding penalty parameter to preserve the local structure of data, and thus does not increase the burden of extra parameter selection. By imposing the orthogonal constraint on the basis matrix of each view, the proposed method is able to handle the out-of-sample. Moreover, the proposed method can be viewed as a unified framework for multi-view learning since it can handle both incomplete and complete multi-view clustering and classification tasks. Extensive experiments conducted on several multi-view datasets prove that the proposed method can significantly improve the clustering performance.

Keywords: Multi-view clustering · incomplete view · common latent representation · out-of-sample

1 Introduction

Multi-view clustering has been achieved great development and has been successfully applied in many applications, such as image retrieval [9], webpage classification [1, 25], and speech recognition [12]. Recently, many methods have been proposed, such as multi-view k-means clustering [2], multi-view spectral clustering via bipartite graph [10], and co-regularized multi-view spectral clustering [8], etc. Compared with the single-view clustering, multi-view clustering can exploit the complementary information among multiple views, and thus has the potential to achieve a better performance [29].

⋆ ⋆ indicates equal contributions; † indicates the corresponding author.

For the conventional multi-view clustering, they commonly require that the available samples should have all of the views. However, it always happens that some views are missing for parts of samples in real world applications [18]. For example, the data obtained by the blood test and images scanned by the magnetic resonance can be regarded as two necessary views for diagnosing the disease. However, it is often the case that we only have the results of one view for some individuals since they would like to take only one of the two tests. In this case, the conventional methods fail. In this paper, we refer to the clustering task with incomplete views as incomplete multi-view clustering (IMC).

For IMC, a few methods have been proposed, which can be commonly categorized into two groups. The one group is based on completing the incomplete views. For example, Trivedi et al. proposed a kernel CCA based method, which tries to recover the kernel matrix of the incomplete view and then learns two projections for the two views, respectively [18]. However, it requires at least one complete view for reference. In other words, it is not applicable to the case that all views are incomplete. To address this issue, Gao et al. proposed a two-step approach, which first fills in the missing views with the corresponding average of all samples, and then learns the common representation for the two views based on the spectral graph theory [5]. The shortcoming of this approach is that it introduces some useless even noisy information to the data. For data with small incomplete percentages, this approach may be effective. However, for the data with large incomplete percentages, this approach is harmful to find the common representation since these useless information may dominate the representation learning [17]. The other group focuses on directly learning the common latent subspace or representation for all views, in which the most representative works are the partial multi-view clustering (PVC) [30], multi-incomplete-view clustering (MIC) [17], and incomplete multi-modality grouping (IMG) [28]. Based on the non-negative matrix factorization (NMF), PVC directly learns a common latent representation for two views by simply regularizing different views of the same sample to have the same representation [30]. MIC jointly learns the latent representation of each view and the consensus representation by utilizing the weighted NMF algorithm, in which the missing views are constrained with the small weight even 0 during learning [17]. IMG can be viewed as an extension of PVC, which further embeds an adaptively learned graph on the latent representation [28].

Although some methods have been proposed to address the IMC problem, several problems still exist which limit their performances. For example, these methods all ignore the geometric structure of data. This indicates that the intrinsic geometric structure of data may be destroyed in the representation space, which may lead to a bad performance. The second shortcoming of these methods, especially MIC and IMG, is that there are many penalty parameters (more than three) to be set. These tunable parameters directly influence the clustering performance and limit its real applications because it is still an open problem to adaptively select the optimal parameter for different datasets. The third shortcoming is that these methods all cannot handle the out-of-sample problem.

In this paper, we propose a novel and simple IMC method, named incomplete multi-view clustering via graph regularized matrix factorization (IMC_GRMF), to solve the above problems and improve the performance. Similar to PVC, the matrix factorization technique is exploited to learn the common latent representation, in which the representation corresponding to those samples with all views are regularized to be consistent. In addition, a nearest neighbor graph is neatly imposed on the reconstruction errors of the matrix factorization to exploit the local geometric structure of data, which enables the method to learn a more compact and discriminative representation for clustering. Compared with the other methods, our approach does not introduce any extra regularization term and corresponding penalty parameter to preserve the locality structure of data. Extensive experimental results prove the effectiveness of the proposed method for incomplete multi-view clustering.

2 Notations and related work

2.1 Notations

Let $X^{(k)} = [X_c^{(k)T}; \bar{X}^{(k)T}]^T \in R^{(n_c+n_k)\times m_k}$ be the kth view of data, where each sample in the corresponding view is represented by a row vector with m_k features, n_c is the number of paired samples (*i.e.*, there are no missing views for these samples). $x_i^{(k)}$ denotes the features of the kth view of the ith sample. We refer to the kth view as $Vi\,(k)$. $\bar{X}^{(k)} \in R^{n_k\times m_k}$ represents that n_k samples only contain the features of $Vi\,(k)$ while the features of the other views are missing. The total samples of the data is $n = n_c + \sum_{k=1}^{v} n_k$. For a matrix $A \in R^{m\times n}$, its l_F norm and l_1 norm are defined as $\|A\|_F = \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{m} a_{i,j}^2}$ and $\|A\|_1 = \sum_{j=1}^{n}\sum_{i=1}^{m} |a_{i,j}|$, respectively, where $a_{i,j}$ denotes the ith row and jth column element of matrix A [23, 14]. $Tr\,(\cdot)$ is the trace operation. We use A^T to denote the transposition of matrix A [15]. I is the identity matrix. $A \geq 0$ means that all elements of matrix A are not less than zero.

2.2 Partial multi-view clustering (PVC)

For data with two incomplete views, PVC seeks to learn a common latent subspace for both two views, where different views of the same sample should have the same representation [14]. The learning model of PVC is formulated as follows:

$$\begin{aligned}\min_{P_c,\bar{P}^{(1)},\bar{P}^{(2)},U^{(1)},U^{(2)}} & \left\| \begin{bmatrix} X_c^{(1)} \\ \bar{X}^{(1)} \end{bmatrix} - \begin{bmatrix} P_c \\ \bar{P}^{(1)} \end{bmatrix} U^{(1)} \right\|_F^2 + \lambda \left\| \begin{bmatrix} P_c \\ \bar{P}^{(1)} \end{bmatrix} \right\|_1 \\ & + \left\| \begin{bmatrix} X_c^{(2)} \\ \bar{X}^{(2)} \end{bmatrix} - \begin{bmatrix} P_c \\ \bar{P}^{(2)} \end{bmatrix} U^{(2)} \right\|_F^2 + \lambda \left\| \begin{bmatrix} P_c \\ \bar{P}^{(2)} \end{bmatrix} \right\|_1 \\ & s.t.\, U^{(1)} \geq 0, U^{(2)} \geq 0, P_c \geq 0, \bar{P}^{(1)} \geq 0, \bar{P}^{(2)} \geq 0, \end{aligned} \tag{1}$$

where λ is the penalty parameter. $U^{(1)} \in R^{K\times m_1}$ and $U^{(2)} \in R^{K\times m_2}$ are the latent space basis matrices for the two views, $P_c \in R^{n_c\times K}$, $\bar{P}^{(1)} \in R^{n_1\times K}$, and $\bar{P}^{(2)} \in R^{n_2\times K}$ are the latent representations of the original data, K is the feature dimension in the latent space.

For PVC, the new representation corresponding to all samples can be expressed as $P = \begin{bmatrix} P_c \\ \bar{P}^{(1)} \\ \bar{P}^{(2)} \end{bmatrix} \in R^{n\times K}$. Then the conventional k-means can be performed on it to obtain the final clustering result.

3 The proposed method

For multi-view data, learning a common latent representation for all views is one of the most favorite approaches in the field of multi-view clustering. However, how to learn a compact and discriminative common representation for the incomplete multi-view data is a challenge task. In this section, a novel multi-view clustering framework shown in Fig.1 is provided to address this issue, in which the local information of each view and the complementary information across different views are jointly integrated.

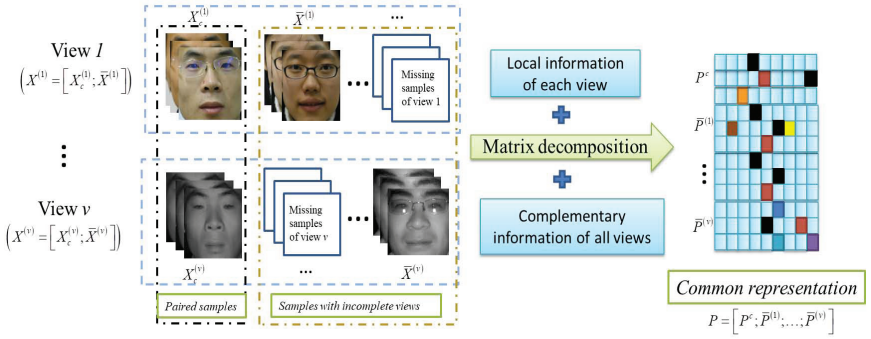


Fig. 1: The description of IMC_GRMF. In this work, we suppose that there are only n_c samples (paired samples) have features of all views.

3.1 Learning model of the proposed method

In past years, exploiting the locality geometric structure of data has been proved an effective approach for representation learning, which not only can improve the discriminability and compactness of the learned representation, but also avoids overfitting [26, 20, 3, 22, 27, 13, 16]. For example, in [13, 16], a nearest neighbor graph is introduced to constrain the new representation or basis for incomplete multi-view clustering. Although the purpose is realized, the complexity is also

increased because they commonly introduce at least one tunable penalty parameter to the model. Since some basic models already have two or more tuned parameters, introducing any extra tuned parameter to the model will greatly increase the burden in parameter selection. So the conventional graph embedding approaches are not a good choice to guide the representation learning. In this section, we propose a novel and simple approach to solve this challenge, in which the local information of each view are embedded into the learning model based on the following Lemma [21].

Lemma 1: For three samples $\{x_1, x_2, x_3\} \in R^m$, suppose x_1 and x_2 are the nearest neighbor to each other, x_3 is not the nearest neighbor to samples x_1 and x_2. If there is a complete dictionary $U \in R^{k\times m}$ that satisfies $x_i = p_i U$ $(i = \{1, 2, 3\})$, where $p_i \in R^k$ can be viewed as the reconstruction coefficient. Then we have the following conclusion: the reconstructed sample p_2U (p_1U) is also the nearest neighbor to the original sample x_1 (x_2) and is still not the nearest neighbor to sample x_3.

The proof to Lemma 1 is very simple and thus we omit it here. From Lemma 1, we know that the reconstruction operation does not destroy the local geometric structure of the original data. Inspired by this motivation, we design the following objective function to exploit the local information of data for common representation learning:

$$\min_{P^{(k)},U^{(k)}} \sum_{k=1}^{v} \sum_{j=1}^{n_c+n_k} \sum_{i=1}^{n_c+n_k} \left\| x_i^{(k)} - p_j^{(k)} U^{(k)} \right\|_2^2 w_{i,j}^{(k)} + \lambda_2 \sum_{k=1}^{v} \left\| P^{(k)} \right\|_1 \quad s.t. U^{(k)} U^{(k)T} = I, \tag{2}$$

where λ_2 is a penalty parameter. $p_j^{(k)}$ is the new representation of the jth sample in the kth view. $w_{i,j}^{(k)}$ is a binary value which is simply pre-defined as follows:

$$w_{i,j}^{(k)} = \begin{cases} 1, if\ x_i^{(k)} \in \Phi\left(x_j^{(k)}\right)\ or\ x_j^{(k)} \in \Phi\left(x_i^{(k)}\right) \\ 0, \qquad\qquad otherwise, \end{cases} \tag{3}$$

where $\Phi\left(x_j^{(k)}\right)$ denotes the sample set of nearest neighbors to sample $x_j^{(k)}$.

By introducing the binary weight to regularize the data reconstruction, the locality structure of the original data in each view can be well preserved. Meanwhile, from (2), we can find that the proposed method does not introduce any extra regularization term and corresponding tuned parameter to preserve such locality property, which greatly reduces the complexity of penalty parameter selection in comparison with the other graph regularized IMC methods, such as DCNMF [13] and GPMVC [16] which all commonly introduce at least an extra tuned penalty parameter to preserve such locality property. For the paired samples across different views, their new representation should be consensus. To this end, we further add a regularization term based on the paired information

of different views as follows:

$$\begin{aligned}\min_{P^{(k)},P^c,U^{(k)}} \sum_{k=1}^{v}\sum_{j=1}^{n_c+n_k}\sum_{i=1}^{n_c+n_k}\left\|x_i^{(k)}-p_j^{(k)}U^{(k)}\right\|_2^2 w_{i,j}^{(k)} \\ +\lambda_1\sum_{k=1}^{v}\left\|G^{(k)}P^{(k)}-P^c\right\|_F^2+\lambda_2\sum_{k=1}^{v}\left\|P^{(k)}\right\|_1 \\ s.t. U^{(k)}U^{(k)T}=I,\end{aligned} \tag{4}$$

where λ_1 is a penalty parameter. $P^c \in R^{c\times K}$ is the common latent representation for the paired samples of different views. $G^{(k)} \in R^{n_c\times(n_c+n_k)}$ can be viewed as an index matrix used to remove the unpaired representation $\bar{P}^{(k)}$ from $P^{(k)}=\begin{bmatrix}P_c^{(k)}\\ \bar{P}^{(k)}\end{bmatrix}$. Since the first n_c samples of each view are regarded as the paired samples, matrix $G^{(k)}$ can be simply defined as follows:

$$G_{i,j}^{(k)}=\begin{cases}1, & if\ i=j\\ 0, & otherwise.\end{cases} \tag{5}$$

For model (4), $P=[P^{cT},\bar{P}^{(1)T},\ldots,\bar{P}^{(v)T}]^T$ can be viewed as the new representations for all samples. After obtaining the new representations, we use k-means algorithm to partition those samples into their respective groups. Several good properties of the proposed model (4) are summarized as follows.

Remark 1: The proposed method is not only a clustering algorithm, but also an unsupervised classification algorithm because it can handle the out-of-sample. In essence, for any sample $x_i^{(k)}$ in the kth view, its new representation is obtained by the matrix factorization $x_i^{(k)}=p_i^{(k)}U^{(k)}$, which is equivalent to $x_i^{(k)}U^{(k)T}=p_i^{(k)}$ since $U^{(k)}U^{(k)T}=I$. Therefore, when the basis matrix $U^{(k)}$ is obtained, we can first achieve the discriminative representation for any new coming sample $y^{(k)}$ by projecting it onto the basis matrix as $p_y^{(k)}=y^{(k)}U^{(k)T}$, and then use the conventional unsupervised classification methods like k-nearest neighbor classify to predict its label.

Remark 2: The proposed model (4) is a unified multi-view learning framework, which can be applied to the incomplete and complete cases by defining different index matrixes $G^{(k)}$.

Remark 3: The proposed method simultaneously exploits the local information of each view and the complementary information across different views, which is beneficial to learn a more compact and discriminative representation for clustering, and thus has the potential to perform better. Moreover, embedding the local information into the model can avoid the overfitting in handing the new sample.

Remark 4: Most importantly, we do not introduce any extra regularization term to preserve the local geometric structure of data. In other words, compared with the conventional graph embedding methods, the proposed method does not increase the burden of parameter tuning.

Remark 5: The proposed method has the potential to recover the missing views. Specifically, for a sample with only the kth view $x^{(k)}$, when its new rep-

resentation $p_{x^{(k)}}$ is obtained via the proposed method, we can recover its fth missing view via $x^{(f)} = p_{x^{(k)}} U^{(f)}$.

3.2 Solution to IMC_GRMF

For the first term of (4), we can rewrite it into the following equivalent formula

$$\begin{aligned} &\sum_{k=1}^{v} \sum_{j=1}^{n_c+n_k} \sum_{i=1}^{n_c+n_k} \left\| x_i^{(k)} - p_j^{(k)} U^{(k)} \right\|_2^2 w_{i,j}^{(k)} \\ =&\sum_{k=1}^{v} \begin{pmatrix} Tr\left(X^{(k)T} D^{(k)} X^{(k)}\right) + Tr\left(U^{(k)T} P^{(k)T} D^{(k)} P^{(k)} U^{(k)}\right) \\ -2Tr\left(X^{(k)T} W^{(k)} P^{(k)} U^{(k)}\right) \end{pmatrix}, \end{aligned} \tag{6}$$

where $D^{(k)}$ is a diagonal matrix with each diagonal element $D_{i,i}^{(k)} = \sum_{j=1}^{n_c+n_k} W_{i,j}^{(k)}$. Considering that the first term of (6) is constant and condition $U^{(k)} U^{(k)T} = I$, we can simplify (4) as follows according to (6):

$$\begin{aligned} L\left(P^{(k)}, P^c, U^{(k)}\right) = \lambda_1 \sum_{k=1}^{v} \left\| G^{(k)} P^{(k)} - P^c \right\|_F^2 + \lambda_2 \sum_{k=1}^{v} \left\| P^{(k)} \right\|_1 \\ + \sum_{k=1}^{v} \left(Tr\left(P^{(k)T} D^{(k)} P^{(k)}\right) - 2Tr\left(X^{(k)T} W^{(k)} P^{(k)} U^{(k)}\right)\right). \end{aligned} \tag{7}$$

Then all variables can be calculated alternatively as follows.

Step 1: Calculate $U^{(k)}$. The basis matrix $U^{(k)}$ for each view can be calculated by optimizing the following problem:

$$\min_{U^{(k)} U^{(k)T} = I} -2Tr\left(X^{(k)T} W^{(k)} P^{(k)} U^{(k)}\right). \tag{8}$$

Then we can obtain the optimum value of $U^{(k)}$ as [31, 19]:

$$U^{(k)} = J^{(k)} B^{(k)T}, \tag{9}$$

where $J^{(k)}$ and $B^{(k)}$ are the right and left singular matrices of $(X^{(k)T} W^{(k)} P^{(k)})$, *i.e.*, $X^{(k)T} W^{(k)} P^{(k)} = B^{(k)} \Sigma^{(k)} J^{(k)T}$.

Step 2: Calculate $P^{(k)}$. Fixing the other variables, variable $P^{(k)}$ can be calculated by minimizing the following problem:

$$\begin{aligned} \min_{P^{(k)}} \lambda_1 \left\| G^{(k)} P^{(k)} - P^c \right\|_F^2 + \lambda_2 \left\| P^{(k)} \right\|_1 \\ + Tr\left(P^{(k)T} D^{(k)} P^{(k)}\right) - 2Tr\left(X^{(k)T} W^{(k)} P^{(k)} U^{(k)}\right). \end{aligned} \tag{10}$$

Define $A^{(k)} = U^{(k)} X^{(k)T} W^{(k)} + \lambda_1 P^{cT} G^{(k)}$, $M^{(k)} = D^{(k)} + \lambda_1 G^{(k)T} G^{(k)}$. Obviously, $M^{(k)}$ is still a diagonal matrix with all diagonal elements $M_{i,i}^{(k)} > 0$.

Thus, (10) can be rewritten into the following equivalent problem:

$$\min_{P^{(k)}} \left\| \left(M^{(k)}\right)^{\frac{1}{2}} P^{(k)} - \left(A^{(k)}\left(M^{(k)}\right)^{-\frac{1}{2}}\right)^T \right\|_F^2 + \lambda_2 \left\| P^{(k)} \right\|_1. \tag{11}$$

Define $C^{(k)} = (A^{(k)}\left(M^{(k)}\right)^{-\frac{1}{2}})^T$, problem (11) can be rewritten as follows

$$\min_{P^{(k)}} \sum_{i=1}^{n_c+n_k} M_{i,i}^{(k)} \left\| P_{i,:}^{(k)} - C_{i,:}^{(k)} \Big/ \sqrt{M_{i,i}^{(k)}} \right\|_2^2 + \lambda_2 \left\| P_{i,:}^{(k)} \right\|_1. \tag{12}$$

where $P_{i,:}^{(k)}$ and $C_{i,:}^{(k)}$ denote the ith row vector of matrices $P^{(k)}$ and $C^{(k)}$, respectively. For problem (12), its solution can be computed independently to each row by the conventional shrinkage operation as follows [19]:

$$P_{i,:}^{(k)} = \Theta_{\lambda_2 \Big/ 2M_{i,i}^{(k)}} \left(C_{i,:}^{(k)} \Big/ \sqrt{M_{i,i}^{(k)}} \right), \tag{13}$$

where Θ denotes the shrinkage operator.

Step 3: Calculate P^c. Fixing the other variables, the common latent representation P^c can be calculated by solving the following minimization problem:

$$\min_{P^c} \sum_{k=1}^{v} \left\| G^{(k)} P^{(k)} - P^c \right\|_F^2. \tag{14}$$

Problem (14) has the following closed form solution:

$$P^c = \sum_{k=1}^{v} G^{(k)} P^{(k)} / v. \tag{15}$$

Algorithm 1 summarizes the computing procedures of IMC_GRMF.

3.3 Computational complexity and convergence property

For Algorithm 1, it is obvious that the biggest computational cost is the singular value decomposition (SVD) in Step 1. Note that the computational complexities of matrix multiplication and addition are ignored since their computational costs are far less than SVD. Thus, we only take into account the computational complexity of Step 1. Generally, the computational complexity of SVD is $O(mn^2)$ for a $m \times n$ matrix [11]. Therefore, the computational complexity of Step 1 is about $O\left(vmK^2\right)$. v is the number of views, K is the reduced dimension or the number of clusters. Therefore, the computation complexity of the proposed method listed in Algorithm 1 is about $O\left(\tau vmK^2\right)$, where τ is the iteration number.

From the above presentations, it is obvious to see that the proposed optimization problem (7) is convex with respect to variables $P^{(k)}, P^c, U^{(k)}$, respectively. Then we have the following **Theorem 1**.

Algorithm 1 : IMC_GRMF (solving problem (4))

Input: Multi-view $X^{(k)}$, index matrix $G^{(k)}$, $k \in [1, v]$, parameters λ_1, λ_2.
Initialization: Initialize $P^{(k)}$ and $U^{(k)}$ with random values, construct the nearest neighbor graph $W^{(k)}$, $P^c = \sum_{k=1}^{v} G^{(k)} P^{(k)} \Big/ v$.
while not converged **do**
  for k from 1 to v
    Update $U^{(k)}$ using (9).
    Update $P^{(k)}$ using (13).
  end
  Update P^c using (15).
end while
Output: $P^c, P^{(k)}, U^{(k)}$

Theorem 1: The objective function value of problem (4) is monotonically decreasing during the iteration.

Proof. Suppose $\Upsilon\left(P_t^{(k)}, P_t^c, U_t^{(k)}\right)$ denotes the objective function value at the tth iteration. Since all sub-problems with respect to variables $P^{(k)}, P^c, U^{(k)}$ are convex and have the closed form solution, the following inequations are satisfied:

$$\begin{aligned} &\Upsilon\left(P_t^{(k)}, P_t^c, U_t^{(k)}\right) \geq \Upsilon\left(P_t^{(k)}, P_t^c, U_{t+1}^{(k)}\right) \\ &\geq \Upsilon\left(P_{t+1}^{(k)}, P_t^c, U_{t+1}^{(k)}\right) \geq \Upsilon\left(P_{t+1}^{(k)}, P_{t+1}^c, U_{t+1}^{(k)}\right). \end{aligned} \tag{16}$$

This inequation illustrates that the objective function value of problem (4) is monotonically decreasing during the iteration. Thus we complete the proof.

Meanwhile, we can find that problem (4) is lower bounded because it at least satisfies the condition $\Upsilon\left(P_t^{(k)}, P_t^c, U_t^{(k)}\right) \geq 0$, thus **Theorem 1** guarantees that the proposed method will finally converge to the local optimal solution after a few iterations.

4 Experiments and analysis

4.1 Experimental settings

Dataset: (1) ***Handwritten digit*** dataset [2]: The used handwritten digit is composed of 2000 samples from 10 digits, *i.e.*, 0-9. Each sample is represented by two views, in which the one is represented by a feature vector with 240 features obtained by the average of pixels in 2×3 windows, and the other one is represented by the Fourier coefficient vector with 76 features. (2) ***BUAA-visnir face dataset (BUAA)*** [7]: Following the experimental settings in [28], we evaluate different methods on the first 10 persons with 90 visual images and 90 near infrared images. Each image was pre-resized to a 10×10 matrix and then transformed into the vector. (3) ***Cornell*** dataset [1, 6]: This dataset contains 195 webpages collected from the Cornell University. Webpages in the dataset are

Table 1: Description of the used benchmark datasets.

Database	Class No.	No. of view	No. of samples	Feature No. of *Vi(1)*/ *Vi(2)*
handwritten digit	10	2	2000	240/76
BUAA	10	2	90	100/100
Cornell	5	2	195	195/1703
Caltech7	7	2	1474	512/928

partitioned into five classes and each webpage is represented by two views, *i.e.*, the content view and citation view. (4) ***Caltech101*** dataset [4]: The original Caltech101 dataset contains 8677 images from 101 objects. In the experiments, a subset named **Caltech7** [10], which is composed of 1474 images from 7 classes, is used to compare different methods. The popular two types of features, *i.e.*,GIST and LBP, are extracted from each image as the two views. The above used datasets are briefly summarized in Table 1.

Evaluation: Three well-known matrices, *i.e.*, clustering accuracy (ACC), normalized mutual information (NMI), and purity are chosen to evaluate the performance of different methods [2]. For the above datasets, we randomly select the percentage of 10, 30, 50, 70, and 90 samples as the paired samples with all views, and treat the remaining samples as incomplete samples, in which half of samples only have one of the views. All methods are repeatedly performed 5 times and their average values (%) are reported for comparison.

Compared methods: Following the experimental settings in [17, 28], we compare the proposed method with the following baselines. (1) **BSV** (Best Single View): BSV first fills in the missing views with the average of samples in the corresponding view, and then performs k-means on each view separately. Finally, the best clustering result of the two views is reported. (2) **Concat**: It first fills in all missing views with the average of samples of the corresponding view, and then concatenates all views of each sample into one feature vector, followed by performing k-means to obtain the clustering reuslt. (3) **PVC** [30]. PVC uses the non-negative matrix factorization technique to learn a common latent representation for incomplete multi-view clustering. (4) **IMG** [28]: IMG extends the PVC by embedding the adaptively learned Laplacian graph. (5) Double constrained NMF (**DCNMF**) [13]: DCNMF is an extension of PVC, which further introduces a Laplacian graph regularizer into PVC. (6) Graph regularized partial multi-view clustering (**GPMVC**) [16]: GPMVC can be viewed as an improved method to DCNMF, which exploits a scale normalization technique in the consensus representation learning term. *The code of the proposed method is available at: http://www.yongxu.org/lunwen.html.*

4.2 Experimental results and analyses

The clustering results of different methods on the above four datasets are enumerated in Table 2-Table 5 and Fig.2. It is obvious to see that the proposed method can significantly improve the ACC, NMI, and purity. In particular, the

Table 2: ACCs/NMIs (%) of different methods on the handwritten digit dataset.

Method/Rate	0.1	0.3	0.5	0.7	0.9
BSV	43.08/37.04	50.46/44.48	57.39/51.50	64.44/58.61	69.29/66.26
Concat	46.01/47.71	57.46/54.43	66.45/61.12	78.64/70.30	86.63/79.34
PVC	63.81/55.13	70.90/60.85	73.44/64.88	75.20/68.54	77.82/72.83
IMG	69.22/58.04	75.41/62.38	76.36/64.91	77.54/68.21	81.78/73.57
DCNMF	51.21/54.23	76.63/65.56	80.61/74.41	86.16/78.14	89.16/80.90
GPMVC	65.60/60.99	74.04/63.99	76.94/72.23	79.06/73.68	81.08/75.24
Ours	**72.70/66.48**	**79.67/71.28**	**86.22/77.27**	**88.98/80.48**	**90.77/83.55**

Table 3: ACCs/NMIs (%) of different methods on the BUAA dataset.

Method/Rate	0.1	0.3	0.5	0.7	0.9
BSV	48.33/43.10	56.96/53.03	64.26/61.78	70.81/69.91	80.16/82.56
Concat	45.62/51.22	46.61/51.95	47.46/52.43	52.34/56.51	57.58/62.66
PVC	57.41/61.35	66.46/67.07	70.01/71.97	75.92/78.70	80.73/84.22
IMG	53.95/54.72	67.39/67.53	76.14/76.74	79.36/82.83	80.78/85.90
DCNMF	58.36/61.78	67.58/68.75	72.15/72.05	76.58/79.66	82.42/86.42
GPMVC	58.98/62.12	68.75/70.25	74.28/74.33	78.28/81.63	84.24/86.78
Ours	**63.82/64.64**	**76.72/76.04**	**82.76/81.35**	**86.20/85.77**	**92.62/91.20**

proposed method achives nearly 8 percent higher than those of the related methods in terms of the ACC on the BUAA dataset. The good performance strongly validates the effectiveness of the proposed method in handling the IMC tasks. Besides, we can obtain the following observations from the experimental results.

(1) Generally, with the ratio of missing views decreases, the clustering performances of all methods improve obviously. This proves that the complementary information of different views is very useful in multi-view learning.

(2) In most cases, BSV and Concat perform much worse than the other methods. This proves that filling in the missing views with the average of samples of the corresponding view is not a useful approach.

(3) DCNMF, GPMVC and the proposed method perform better than PVC in most cases. Compared with PVC, the other two methods and the proposed method all exploit the local geometric structure of each view to guide the representation learning. Thus, the experimental results prove that the local information of each view contain very useful information, which is beneficial to learn a more compact and discriminative representation. Meanwhile, we can find that our method achieves better performance than DCNMF and GPMVC, which further proves the effectiveness of the proposed novel graph regularization term.

4.3 Parameter analysis

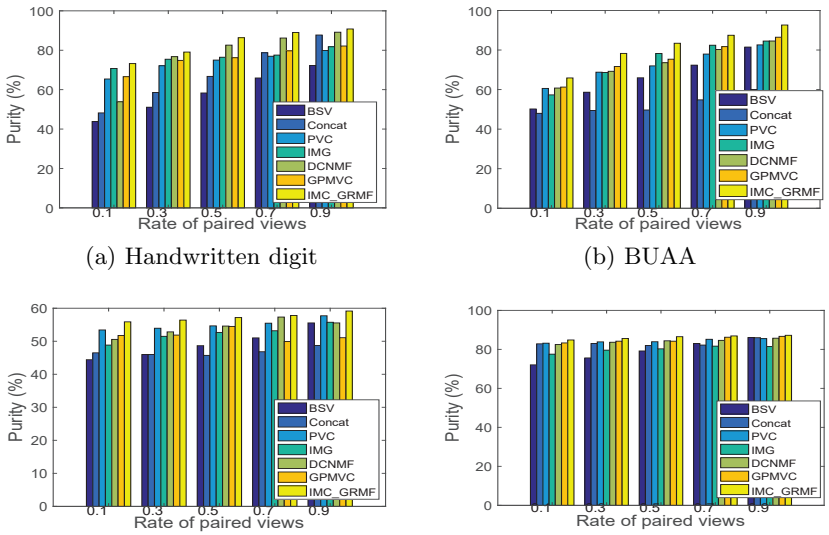
Fig.3 shows the ACC versus the parameters λ_1 and λ_2 on the handwritten digit and BUAA datasets with 70% paired samples. It is obvious that the ACC of the proposed method is relatively stable in some local areas, which indicates

Table 4: ACCs/NMIs (%) of different methods on the cornell dataset.

Method/Rate	0.1	0.3	0.5	0.7	0.9
BSV	42.41/8.66	43.93/8.19	44.84/8.89	46.32/12.69	47.66/19.34
Concat	38.80/8.07	38.06/7.56	36.96/8.30	36.79/10.21	38.48/13.47
PVC	42.56/15.76	42.56/16.00	43.79/18.21	42.56/19.76	43.03/21.03
IMG	45.13/12.56	45.79/16.62	47.08/19.24	45.51/20.89	44.76/22.98
DCNMF	39.94/13.59	43.29/17.72	43.18/19.17	45.74/21.69	45.52/23.98
GPMVC	40.39/13.90	43.86/16.07	46.53/18.99	44.56/15.03	44.35/17.07
Ours	**46.99/17.23**	**47.40/18.36**	**49.03/21.01**	**48.78/22.11**	**49.20/25.02**

Table 5: ACCs/NMIs (%) of different methods on the Caltech7 dataset.

Method/Rate	0.1	0.3	0.5	0.7	0.9
BSV	42.66/29.04	40.97/32.11	39.83/35.13	42.92/38.83	46.99/44.16
Concat	36.83/33.82	31.74/34.44	36.36/34.56	43.38/38.15	47.08/45.44
PVC	43.46/38.99	43.96/40.26	44.46/40.17	44.76/41.60	44.34/41.94
IMG	42.05/32.38	42.36/33.29	42.23/35.05	41.17/35.96	43.23/37.64
DCNMF	40.63/33.86	44.53/38.19	45.62/41.40	48.50/41.24	50.74/44.04
GPMVC	45.57/40.05	47.19/40.96	46.99/41.83	46.99/42.61	49.10/46.02
Ours	**50.88/41.01**	**51.15/42.74**	**51.40/45.69**	**51.79/46.78**	**51.88/48.28**

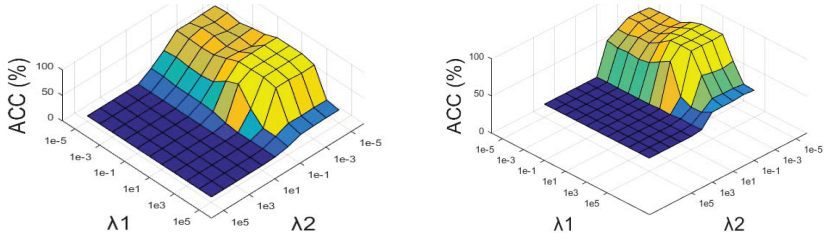


(a) Handwritten digit (b) BUAA

(c) Cornell (d) Caltech7

Fig. 2: Purity (%) of different methods on the above four datasets.

that the proposed method is insensitive to the selection of parameters to some extent. Moreover, we can find that when the two parameters are selected with



(a) Handwritten digit (b) BUAA

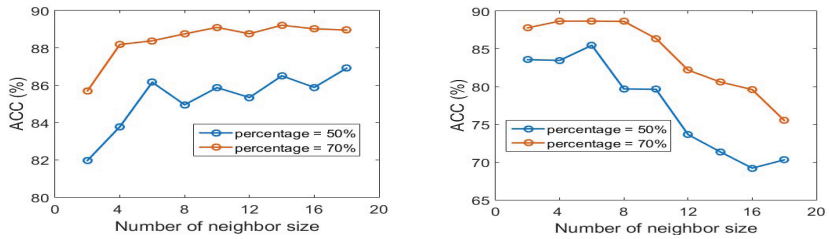
Fig. 3: ACC (%) versus parameters λ_1 and λ_2 of the proposed method on (a) handwritten digit and (b) BUAA datasets with 70% paired samples.

proper values from the candidate range of ($[10^0, 10^2]$, $[10^{-5}, 10^{-1}]$), the proposed method can achieve the satisfactory performance. This indicates that a relative larger value of parameter λ_1 encourages a better performance. In our work, we use the grid searching approach to find the optimal combinations of the two parameters from the two dimensional grid formed by ($[10^0, 10^2]$, $[10^{-5}, 10^{-1}]$) [24].

Fig.4 plots the relationships of ACC and the number of nearest neighbors of the proposed method on the handwritten digit and BUAA datasets. From the figures, we have the following conclusions: (1) The clustering performance is insensitive to the selection of nearest neighbor number to some extent when the nearest neighbor number is located in the proper range, such as $[8, 18]$ for the handwritten digit dataset and $[2, 6]$ for the BUAA dataset. (2) Generally, the number of nearest neighbors should better be less than the number of sample of each class. For example, from Fig.3(b), we can find that when the number of nearest neighbors is larger than the number of sample per class, *i.e.*, $N > 10$, the ACC decreases dramatically. However, in the real world applications, it is impossible to obtain the real number of sample per class. In this work, we use the following criterion to select the number of the nearest neighbors. Suppose we try to partition the available multi-view data with n samples into c groups, $m = n/c$. If $m \gg 10$, then we empirically select 10 as the number of nearest neighbors, otherwise we select $min(m - 4, 2)$ as the nearest neighbor number.

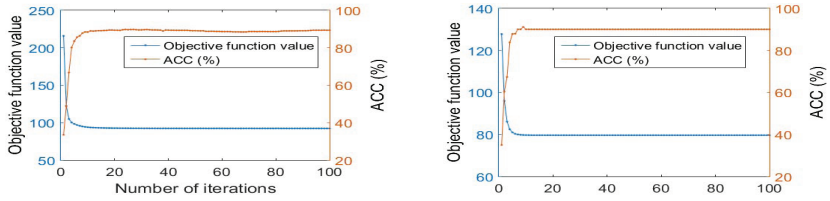
4.4 Experimental convergence study

Fig.5 shows the objective function value and ACC at each iteration step on the handwritten digit and BUAA datasets with 70% paired samples. From the figures, it is obvious to see that the objective function value decreases dramatically in the first few iteration steps (within 20 iterations). The experimental results plotted in the two figures prove the good convergence property of our method.



(a) Handwritten digit (b) BUAA

Fig. 4: ACC (%) versus the number of nearest neighbors of our method on (a) handwritten digit and (b) BUAA datasets with 50% and 70% paired samples.



(a) Handwritten digit (b) BUAA

Fig. 5: The objective function value and ACC (%) versus the iteration step of the proposed method on (a) handwritten digit and (b) BUAA datasets with 70% paired samples.

5 Conclusions

In this paper, we propose a novel framework for multi-view learning, which not only can handle the incomplete and complete multi-view clustering, but also is able to deal with the out-of-sample. Moreover, the proposed method has the potential to complete the missing views for any sample. Besides, we provide a novel approach to exploit the local information of data without introducing any extra regularization term and penalty parameter, which does not increase the complexity and computational burden. Extensive experimental results prove the effectiveness of the proposed method.

6 Acknowledgments

This work is supported in part by Economic, Trade and information Commission of Shenzhen Municipality (Grant no. 20170504160426188).

References

1. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT. pp. 92–100. ACM (1998)
2. Cai, X., Nie, F., Huang, H.: Multi-view k-means clustering on big data. In: IJCAI. pp. 2598–2604 (2013)
3. Fei, L., Xu, Y., Zhang, B., Fang, X., Wen, J.: Low-rank representation integrated with principal line distance for contactless palmprint recognition. Neurocomputing **218**, 264–275 (2016)
4. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Comput. Vis. Image Und. **106**(1), 59–70 (2007)
5. Gao, H., Peng, Y., Jian, S.: Incomplete multi-view clustering. In: ICIIP. pp. 245–255. Springer (2016)
6. Guo, Y.: Convex subspace representation learning from multi-view data. In: AAAI. vol. 1, pp. 387–393 (2013)
7. Huang, D., Sun, J., Wang, Y.: The buaa-visnir face database instructions. School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001 (2012)
8. Kumar, A., Rai, P., Daume, H.: Co-regularized multi-view spectral clustering. In: NIPS. pp. 1413–1421 (2011)
9. Li, M., Xue, X.B., Zhou, Z.H.: Exploiting multi-modal interactions: A unified framework. In: IJCAI. pp. 1120–1125 (2009)
10. Li, Y., Nie, F., Huang, H., Huang, J.: Large-scale multi-view spectral clustering via bipartite graph. In: AAAI. pp. 2750–2756 (2015)
11. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. IEEE TPAMI **35**(1), 171–184 (2013)
12. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML. pp. 689–696 (2011)
13. Qian, B., Shen, X., Gu, Y., Tang, Z., Ding, Y.: Double constrained nmf for partial multi-view clustering. In: DICTA. pp. 1–7. IEEE (2016)
14. Qin, J., Liu, L., Shao, L., Ni, B., Chen, C., Shen, F., Wang, Y.: Binary coding for partial action analysis with limited observation ratios. In: CVPR. pp. 146–155 (2017)
15. Qin, J., Liu, L., Shao, L., Shen, F., Ni, B., Chen, J., Wang, Y.: Zero-shot action recognition with error-correcting output codes. In: CVPR. pp. 2833–2842 (2017)
16. Rai, N., Negi, S., Chaudhury, S., Deshmukh, O.: Partial multi-view clustering using graph regularized nmf. In: ICPR. pp. 2192–2197. IEEE (2016)
17. Shao, W., He, L., Philip, S.Y.: Multiple incomplete views clustering via weighted nonnegative matrix factorization with l_ {2, 1} regularization. In: ECML PKDD. pp. 318–334. Springer (2015)
18. Trivedi, A., Rai, P., Daumé III, H., DuVall, S.L.: Multiview clustering with incomplete views. In: NIPSW. pp. 1–7 (2010)
19. Wen, J., Fang, X., Cui, J., Fei, L., Yan, K., Chen, Y., Xu, Y.: Robust sparse linear discriminant analysis. IEEE TCSVT (2018). https://doi.org/10.1109/TCSVT.2018.2799214
20. Wen, J., Fang, X., Xu, Y., Tian, C., Fei, L.: Low-rank representation with adaptive graph regularization. Neural Networks **108**, 83–96 (2018)
21. Wen, J., Han, N., Fang, X., Fei, L., Yan, K., Zhan, S.: Low-rank preserving projection via graph regularized reconstruction. IEEE TCYB **PP**(99), 1–13 (2018). https://doi.org/10.1109/TCYB.2018.2799862

22. Zhang, Y., Zhang, Z., Qin, J., Zhang, L., Li, B., Li, F.: Semi-supervised local multi-manifold isomap by linear embedding for feature extraction. Pattern Recogn. **76**, 662–678 (2018)
23. Zhang, Z., Zhao, M., Chow, T.W.: Binary-and multi-class group sparse canonical correlation analysis for feature extraction and classification. IEEE TKDE **25**(10), 2192–2205 (2013)
24. Zhang, Z., Lai, Z., Xu, Y., Shao, L., Wu, J., Xie, G.S.: Discriminative elastic-net regularized linear regression. IEEE TIP **26**(3), 1466–1481 (2017)
25. Zhang, Z., Liu, L., Qin, J., Zhu, F., Shen, F., Xu, Y., Shao, L., Shen, H.T.: Highly-economized multi-view binary compression for scalable image clustering. In: ECCV (2018)
26. Zhang, Z., Shao, L., Xu, Y., Liu, L., Yang, J.: Marginal representation learning with graph structure self-adaptation. IEEE TNNLS (2017). https://doi.org/10.1109/TNNLS.2017.2772264
27. Zhang, Z., Xu, Y., Shao, L., Yang, J.: Discriminative block-diagonal representation learning for image recognition. IEEE TNNLS pp. 3111–3125 (2018)
28. Zhao, H., Liu, H., Fu, Y.: Incomplete multi-modal visual data grouping. In: IJCAI. pp. 2392–2398 (2016)
29. Zhao, J., Xie, X., Xu, X., Sun, S.: Multi-view learning overview: Recent progress and new challenges. Inform. Fusion **38**, 43–54 (2017)
30. Zhi, S.Y., Zhou, H.: Partial multi-view clustering. In: AAAI. pp. 1968–1974 (2014)
31. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. J. Comput. Graph. Stat. **15**(2), 265–286 (2006)