Simultaneous Representation Learning and Clustering for Incomplete Multi-view Data

Wenzhang Zhuge[1], **Chenping Hou**[1*], **Xinwang Liu**[2], **Hong Tao**[1] and **Dongyun Yi**[1]
[1]College of Liberal Arts and Sciences, National University of Defense Technology, Changsha, China
[2]School of Computer, National University of Defense Technology, Changsha, China
hcpnudt@hotmail.com

Abstract

Incomplete multi-view clustering has attracted various attentions from diverse fields. Most existing methods factorize data to learn a unified representation linearly. Their performance may degrade when the relations between the unified representation and data of different views are nonlinear. Moreover, they need post-processing on the unified representations to extract the clustering indicators, which separates the consensus learning and subsequent clustering. To address these issues, in this paper, we propose a Simultaneous Representation Learning and Clustering (SRLC) method. Concretely, SRLC constructs similarity matrices to measure the relations between pair of instances, and learns low-dimensional representations of present instances on each view and a common probability label matrix simultaneously. Thus, the nonlinear information can be reflected by these representations and the clustering results can obtained from label matrix directly. An efficient iterative algorithm with guaranteed convergence is presented for optimization. Experiments on several datasets demonstrate the advantages of the proposed approach.

1 Introduction

Conventional multi-view learning assumes that each example of data appears in all views [Jing *et al.*, 2017; Tao *et al.*, 2017; Nie *et al.*, 2016a; Karasuyama and Mamitsuka, 2013; Sun, 2013; Hou *et al.*, 2010]. However, in real-world applications, it is often the case that every view suffers from some data missing, which results in incomplete multi-view data. For example, in cross-language document clustering, documents have been translated into different languages representing multiple views. However, not all documents are translated into each language. Another example is web image retrieval. Not all web images are associated with text descriptions and the image itself may be inaccessible due to deletion or invalid url. Such incompleteness makes it impossible to apply conventional methods on these data directly. Therefore, how to efficiently manipulate this kind of incomplete multi-view data becomes a practical and important problem.

Since clustering is a basic and common task, a few clustering methods have recently been proposed for incomplete multi-view data. As a pioneering work, [Li *et al.*, 2014] learns the representations of view-specific examples and complete examples simultaneously, thus in the learned latent subspace, all examples are homogeneously represented. Such strategy has also been adopted by [Yin *et al.*, 2015; Zhao *et al.*, 2016; Yin *et al.*, 2017; Cai *et al.*, 2018; Zhao *et al.*, 2018]. The limitation of this strategy is that it requires each data sample appears in all views or only one view. A more general case for a multi-view data with more than two views is that each data sample is present on any number of views. To deal with this case, another strategy has been used by [Xu *et al.*, 2015; Hu and Chen, 2018; Tao *et al.*, 2018]. By introducing weight matrices which distinguish present and missing instances on each view, these methods factorize data matrices of different views into a common latent subspace, and then perform post-processing on the common representation matrix to obtain the clustering results. Their performance can be further improved due to the following reasons: (1) Since these methods are based on Matrix Factorization (MF), they are essentially linear, and thus cannot well disclose non-linear structure hidden in data, which limits their learning ability. (2) The manner of separately performing representation learning and clustering may not be able to jointly obtain the optimal solution.

In this paper, we propose Simultaneous Representation Learning and Clustering (SRLC) to address the aforementioned issues. Specifically, to utilize the non-linear information, on each view, SRLC constructs a similarity matrix to measure the relationships of present instances. And based on these matrices, SRLC incorporates representation learning and clustering by learning low-dimensional representations of present instances and a common probability label matrix simultaneously. To enhance the reasonability of the model, rotation matrices are introduced to deliver clustering information, and class coding vectors are employed to establish a weight mechanism. The mechanism characterizes the contribution of each sample according to its clustering uncertainty, which improve the robustness of SRLC. An iterative optimization algorithm for SRLC with proved convergence is proposed. The effectiveness of SRLC is demonstrated by comparing with state-of-the-art methods on six datasets.

*Contact Author

2 Notations and Problem Setting

Throughout the paper, matrices and vectors are written as boldface uppercase letters and boldface lowercase letters, respectively. For a matrix $\mathbf{M}$, its i-th row and ij-th element are denoted by $\mathbf{m}_i$ and m_{ij}, respectively. The transpose, the trace and Frobenius norm of matrix $\mathbf{M}$ are denoted by $\mathbf{M}^T$, $tr(\mathbf{M})$ and $||\mathbf{M}||_F$, respectively. For a vector $\mathbf{m}_i$, the 2-norm of $\mathbf{m}_i$ is denoted by $||\mathbf{m}_i||$.

Given a dataset $\{\mathbf{x}_i|i = 1, ..., n\}$ with n samples sampled from V views, where $\mathbf{x}_i$ is the i-th sample. Each sample has V representations, i.e., $\mathbf{x}_i = [\mathbf{x}_i^{(1)}, ..., \mathbf{x}_i^{(V)}] \in \mathcal{R}^{1\times d}$, where $\mathbf{x}_i^{(v)} \in \mathcal{R}^{1\times d^{(v)}}$ is the i-th example on the v-th view and $d = \sum_{v=1}^{V} d^{(v)}$. The dataset can be denoted in a matrix form $\mathbf{X} = [\mathbf{X}^1, ..., \mathbf{X}^V] \in \mathcal{R}^{n\times d}$, where $\mathbf{X}^{(v)} = [\mathbf{x}_1^{(v)}; ...; \mathbf{x}_n^V] \in \mathcal{R}^{n\times d^{(v)}}$ collects the instances of the V-th view.

In the incomplete multi-view setting, each $\mathbf{x}_i^{(v)}$ can be missing. For each view, a diagonal indicator matrix $\mathbf{O}^{(v)} \in \{0, 1\}^{n\times n}$ is defined as:

$$o_{ii}^{(v)} = \begin{cases} 1, & \text{if } \mathbf{x}_i^{(v)} \text{ appears in the } v\text{-th view} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Incomplete multi-view clustering aims to cluster the n samples into C clusters by integrating all incomplete views.

3 Proposed Approach

In this section, we first present the motivation of this paper. Next, we introduce our model in two aspects and then give an unified objective function. Finally, an efficient iterative algorithm is presented for optimization.

3.1 Motivation

Most existing incomplete multi-view clustering methods are based on matrix factorization. There are mainly two separate steps of these methods:

Step 1. Factorizing each $\mathbf{X}^{(v)}$ into a common latent feature matrix $\mathbf{V} \in \mathbb{R}^{n\times C}$ and a basis matrix $\mathbf{U}^{(v)} \in \mathbb{R}^{C\times d^{(v)}}$ simultaneously by solving the following problem:

$$\begin{aligned} &\min_{\mathbf{V},\mathbf{U}^{(v)}} \sum_{v=1}^{V} \{||\mathbf{O}^{(v)}(\mathbf{X}^{(v)} - \mathbf{V}\mathbf{U}^{(v)})||_F^2 + \Psi(\mathbf{V}, \mathbf{U}^{(v)})\} \\ &s.t. \quad \mathbf{V} \in \mathcal{C}_1, \mathbf{U}^{(v)} \in \mathcal{C}_2^{(v)}, v = 1, ..., V \end{aligned} \tag{2}$$

where $\Psi(\mathbf{V}, \mathbf{U}^{(v)})$ is the regularization term on $\mathbf{V}$ and $\mathbf{U}^{(v)}$, and $\mathcal{C}_1$ and $\mathcal{C}_2^{(v)}$ are constraint sets. These methods distinguish each other by employing different constraints or regularization terms on $\mathbf{V}$ and $\mathbf{U}^{(v)}$.

Step 2. Applying a post-processing algorithm such as K-means on $\mathbf{V}$ to obtain the clustering indicators.

Although the incomplete multi-view data can be clustered by applying these two steps, the performance of these methods can be further improved due to the following reasons: (1) the learned unified representation matrix $\mathbf{V}$ may be of low quality if the relations between $\mathbf{V}$ and $\{\mathbf{X}^{(v)}\}_{v=1}^V$ are nonlinear; (2) successively and independently performing two steps are not guaranteed to yield globally optimal solution.

3.2 Partial Spectral Embedding

To disclose the non-linear structure and utilize the complementary information of different views, we construct an undirected weighted graph $\mathbf{S}^{(v)} \in \mathcal{R}^{n\times n}$ on each view according to pairwise similarity of $\{\mathbf{x}_i^{(v)}\}_{i=1}^n$. Since some instances can be missing, $s_{ij}^{(v)}$ is calculated by

$$s_{ij}^{(v)} = \begin{cases} f(\mathbf{x}_i^{(v)}, \mathbf{x}_j^{(v)}), & \text{if } o_{ii}^{(v)} o_{jj}^{(v)} = 1 \\ \Theta, & \text{otherwise} \end{cases} \tag{3}$$

where $f(\mathbf{x}_i^{(v)}, \mathbf{x}_j^{(v)})$ is a similarity calculation method such as [Zelnik-Manor and Perona, 2004; Nie *et al.*, 2016b], and Θ denotes the information of $s_{ij}^{(v)}$ is missing. According to Eq. (3), $s_{ij}^{(v)}$ can be estimated only if both $\mathbf{x}_i^{(v)}$ and $\mathbf{x}_j^{(v)}$ are present. To eliminate the influence of Θ in subsequent calculating, for convenience, we define $\Theta \cdot 0 = 0 \cdot \Theta = 0$.

Based on $\mathbf{S}^{(v)}$, we aim to learn a low-dimensional representation matrix $\mathbf{F}^{(v)} = [\mathbf{f}_1^{(v)}; ...; \mathbf{f}_n^{(v)}] \in \mathcal{R}^{n\times C}$ which contains the clustering information of $\mathbf{X}^{(v)}$. Since the information of $\mathbf{S}^{(v)}$ can be partial, the i-th row $\mathbf{f}_i^{(v)}$ of $\mathbf{F}^{(v)}$ can be computed only if $\mathbf{x}_i^{(v)}$ is present, i.e., $o_{ii}^{(v)} = 1$. As a result, the optimization problem can be written as

$$\begin{aligned} &\min \sum_{i,j=1}^{n} o_{ii}^{(v)} s_{ij}^{(v)} o_{jj}^{(v)} ||\mathbf{f}_i^{(v)} - \mathbf{f}_j^{(v)}||^2 \\ &s.t. \ (\mathbf{F}^{(v)})^T \mathbf{O}^{(v)} \mathbf{F}^{(v)} = \mathbf{I}_C \end{aligned} \tag{4}$$

where $\mathbf{I}_C \in \mathbb{R}^{C\times C}$ is the identity matrix. According to Eq. (4), if $o_{ii}^{(v)} = 0$, the elements of $\mathbf{f}_i^{(v)}$ can be assigned with arbitrary values.

3.3 Probability Spectral Rotation

To incorporate clustering into representation learning and facilitate consensus, we learn a common probability label matrix $\mathbf{Y} = [\mathbf{y}_1; ...; \mathbf{y}_n] \in \mathcal{R}^{n\times C}$ together with $\{\mathbf{F}^{(v)}\}_{v=1}^V$.

To establish reasonable interactions between $\mathbf{Y}$ and $\{\mathbf{F}^{(v)}\}_{v=1}^V$, for each $\mathbf{F}^{(v)}$, a rotation matrix $\mathbf{R}^{(v)} \in \mathbb{R}^{C\times C}$ is employed to extract the clustering information, and C coding vectors $\{\mathbf{t}_{(c)}\}_{c=1}^C$ are introduced to identify the C classes. For the c-th coding vector $\mathbf{t}_{(c)} \in \mathbb{R}^{1\times C}$, only its c-th element is equal to 1 and the other ones are 0 ($c = 1, ..., C$). The probability spectral rotation term is formulated as:

$$\begin{aligned} &\min \sum_{i=1}^{n} \sum_{c=1}^{C} (y_{ic})^\gamma \sum_{v=1}^{V} o_{ii}^{(v)} ||\mathbf{t}_{(c)} - \mathbf{f}_i^{(v)} \mathbf{R}^{(v)}||^2 \\ &s.t. \ \mathbf{Y} \geqslant 0, \mathbf{Y}\mathbf{1}_c = \mathbf{1}_n, (\mathbf{R}^{(v)})^T \mathbf{R}^{(v)} = \mathbf{I}_C \end{aligned} \tag{5}$$

where $\gamma \geq 1$ is an adaptive parameter. From Eq. (5), it can be observed that $\mathbf{f}_i^{(v)}$ affects $\mathbf{Y}$ and $\mathbf{R}^{(v)}$ only if $o_{ii}^{(v)} = 1$. Different from the classical procrustes average technique [Hastie *et al.*, 2010] which rotates fixed $\{\mathbf{F}^{(v)}\}_{v=1}^V$ to form a unified binary indicator matrix, Eq. (5) generates a probability matrix $\mathbf{Y}$ according to the rotation loss of $\{\mathbf{f}_i^{(v)}\}_{i,v}$ to

$\{\mathbf{t}_{(c)}\}_{c=1}^{C}$. Thus, each sample can be weighted by Eq. (5) automatically according to the degree of clustering uncertainty $\sum_{c=1}^{C}(y_{ic})^{\gamma}$, which enables clearly clustered samples to play more important roles in the learning stage.

3.4 Unified Objective Function

By combining Eq. (4) and Eq. (5), we propose our SRLC model as follows:

$$\begin{aligned}
&\min \mathcal{J}(\{\mathbf{R}^{(v)},\mathbf{F}^{(v)}\}_{v=1}^{V},\mathbf{Y}) \\
&=\sum_{v=1}^{V}\{\sum_{i=1}^{n}o_{ii}^{(v)}\sum_{c=1}^{C}(y_{ic})^{\gamma}||\mathbf{t}_{(c)}-\mathbf{f}_{i}^{(v)}\mathbf{R}^{(v)}||^{2} \\
&\qquad +\lambda\sum_{i,j=1}^{n}o_{ii}^{(v)}s_{ij}^{(v)}o_{jj}^{(v)}||\mathbf{f}_{i}^{(v)}-\mathbf{f}_{j}^{(v)}||^{2}\} \\
&s.t.\ (\mathbf{R}^{(v)})^{T}\mathbf{R}^{(v)}=\mathbf{I}_{C},(\mathbf{F}^{(v)})^{T}\mathbf{O}^{(v)}\mathbf{F}^{(v)}=\mathbf{I}_{C}, \\
&\quad \mathbf{Y}\geqslant 0,\mathbf{Y}\mathbf{1}_{c}=\mathbf{1}_{n}
\end{aligned} \tag{6}$$

where $\lambda>0$ is a balanced-parameter that controls the trade-off between the aforementioned two objectives. With the increase of λ, SRLC learns $\{\mathbf{F}^{(v)}\}_{v=1}^{V}$ more based on $\mathbf{Y}$ to utilize more consensus information. With the decrease λ, SRLC learns $\{\mathbf{F}^{(v)}\}_{v=1}^{V}$ more based on $\{\mathbf{S}^{(v)}\}_{v=1}^{V}$ to utilize more complementarity information.

3.5 Optimization

The problem (6) is not convex over all three groups of variables $\{\mathbf{R}^{(v)}\}_{v=1}^{V},\{\mathbf{F}^{(v)}\}_{v=1}^{V},\mathbf{Y}$, simultaneously. To solve this problem, we adopt alternating minimization strategy.

Update $\mathbf{R}^{(v)}$. With $\mathbf{Y}$ and $\{\mathbf{F}^{(v)}\}_{v=1}^{V}$ fixed, for each $\mathbf{R}^{(v)}$, we need to solve the following problem

$$\min_{(\mathbf{R}^{(v)})^{T}\mathbf{R}^{(v)}=\mathbf{I}_{C}}\sum_{i=1}^{n}o_{ii}^{(v)}\sum_{c=1}^{C}(y_{ic})^{\gamma}||\mathbf{t}_{(c)}-\mathbf{f}_{i}^{(v)}\mathbf{R}^{(v)}||^{2} \tag{7}$$

To update $\mathbf{R}^{(v)}$, we introduce the following propositions.

Proposition 1. *The minimum problem (7) is equivalent to*

$$\max_{(\mathbf{R}^{(v)})^{T}\mathbf{R}^{(v)}=\mathbf{I}_{C}} tr\big((\mathbf{R}^{(v)})^{T}(\mathbf{F}^{(v)})^{T}\mathbf{O}^{(v)}\mathbf{G}\big) \tag{8}$$

where $\mathbf{G}\in\mathbb{R}^{n\times C}$ and its i-th row $\mathbf{g}_i=\sum_{c=1}^{C}(y_{ic})^{\gamma}\mathbf{t}_{(c)}$.

The detailed proofs of all propositions of this paper can be found in the supplementary material.

Proposition 2. *Denote $(\mathbf{F}^{(v)})^{T}\mathbf{O}^{(v)}\mathbf{G}$ as $\mathbf{M}^{(v)}$. Suppose the SVD of $\mathbf{M}^{(v)}$ is $\mathbf{M}^{(v)}=\mathbf{U}^{(v)}\mathbf{\Sigma}^{(v)}(\mathbf{V}^{(v)})^{T}$, then the optimal solution $\mathbf{R}^{(v)}$ to the problem (7) is*

$$\mathbf{R}^{(v)}=\mathbf{U}^{(v)}(\mathbf{V}^{(v)})^{T} \tag{9}$$

Update $\mathbf{F}^{(v)}$. With $\{\mathbf{R}^{(v)}\}_{v=1}^{V}$ and $\mathbf{Y}$ fixed, for each $\mathbf{R}^{(v)}$, we need to solve the following problem

$$\begin{aligned}
\min_{(\mathbf{F}^{(v)})^{T}\mathbf{O}^{(v)}\mathbf{F}^{(v)}=\mathbf{I}_{C}}&\{\sum_{i=1}^{n}o_{ii}^{(v)}\sum_{c=1}^{C}(y_{ic})^{\gamma}||\mathbf{t}_{(c)}-\mathbf{f}_{i}^{(v)}\mathbf{R}^{(v)}||^{2} \\
&+\lambda\sum_{i,j=1}^{n}o_{ii}^{(v)}s_{ij}^{(v)}o_{jj}^{(v)}||\mathbf{f}_{i}^{(v)}-\mathbf{f}_{j}^{(v)}||^{2}\}
\end{aligned} \tag{10}$$

By analyzing Eq. (10), it can be observed that $\mathbf{f}_i^{(v)}$ needs to be optimized only if $o_{ii}^{(v)}=1$. Suppose $\sum_{i=1}^{n}o_{ii}^{(v)}=n^{(v)}$ data points appears on the v-th view, the corresponding rows of $\mathbf{F}^{(v)}$ and $\mathbf{G}$ are collected by $\mathbf{F}_{\Omega}^{(v)}\in\mathbb{R}^{n^{(v)}\times C}$ and $\mathbf{G}_{\Omega}^{(v)}\in\mathbb{R}^{n^{(v)}\times C}$, respectively. $\mathbf{S}_{\Omega}^{(v)}\in\mathbb{R}^{n^{(v)}\times n^{(v)}}$ collects the elements $f(\mathbf{x}_i^{(v)},\mathbf{x}_j^{(v)})$ of $\mathbf{S}^{(v)}$. To optimize $\mathbf{F}_{\Omega}^{(v)}$, according to Proposition 1, the problem (10) can be transformed into the following matrix form

$$\min_{(\mathbf{F}_{\Omega}^{(v)})^{T}\mathbf{F}_{\Omega}^{(v)}=\mathbf{I}_{C}} tr\big((\mathbf{F}_{\Omega}^{(v)})^{T}(\lambda\mathbf{L}_{\Omega}^{(v)}\mathbf{F}_{\Omega}^{(v)}-\mathbf{G}_{\Omega}^{(v)}(\mathbf{R}^{(v)})^{T})\big) \tag{11}$$

where $\mathbf{L}_{\Omega}^{(v)}$ is the Laplacian matrix of $\mathbf{S}_{\Omega}^{(v)}$. Since the problem (11) is difficult to solve directly, we consider the following relaxed problem:

$$\max_{(\mathbf{F}_{\Omega}^{(v)})^{T}\mathbf{F}_{\Omega}^{(v)}=\mathbf{I}_{C}} tr\big((\mathbf{F}_{\Omega}^{(v)})^{T}(\mathbf{A}^{(v)}\mathbf{F}_{\Omega}^{(v)}+\mathbf{B}^{(v)})\big) \tag{12}$$

where $\mathbf{A}^{(v)}=\alpha^{(v)}\mathbf{I}-\mathbf{L}_{\Omega}^{(v)}$ and $\mathbf{B}^{(v)}=\mathbf{G}_{\Omega}^{(v)}(\mathbf{R}^{(v)})^{T}/\lambda$. $\alpha^{(v)}$ is an arbitrary constant which ensures that $\mathbf{A}^{(v)}$ is a positive definite matrix. Motivated by [Nie *et al.*, 2017], the problem (12) can be solved by the following iteration strategy:

1. Update $\mathbf{C}^{(v)}=\mathbf{A}^{(v)}\mathbf{F}_{\Omega}^{(v)}+\mathbf{B}^{(v)}$.
2. Update $\mathbf{F}_{\Omega}^{(v)}$ with fixed $\mathbf{C}^{(v)}$. According to Proposition 2, suppose the SVD of $\mathbf{C}^{(v)}$ is $\mathbf{C}^{(v)}=\overline{\mathbf{U}}^{(v)}\overline{\mathbf{\Sigma}}^{(v)}(\overline{\mathbf{V}}^{(v)})^{T}$, then $\mathbf{F}_{\Omega}^{(v)}$ is updated by

$$\mathbf{F}_{\Omega}^{(v)}=\overline{\mathbf{U}}^{(v)}(\overline{\mathbf{V}}^{(v)})^{T} \tag{13}$$

After obtaining the $\mathbf{F}_{\Omega}^{(v)}$, $\mathbf{F}^{(v)}$ can be updated accordingly. The solution procedure of (12) is listed in Algorithm 1.

Algorithm 1 Algorithm to solve the problem (12)

Input: The matrices $\mathbf{F}_{\Omega}^{(v)}$, $\mathbf{L}_{\Omega}^{(v)}$, $\mathbf{G}_{\Omega}^{(v)}$ and $\mathbf{R}^{(v)}$, the parameters λ and $\alpha^{(v)}$, the maximum number of iteration T_1.
Output: $\mathbf{F}_{\Omega}^{(v)}$.
Initialization:
1. Calculate $\mathbf{A}^{(v)}=\alpha^{(v)}\mathbf{I}-\mathbf{L}_{\Omega}^{(v)}$.
2. Compute $\mathbf{B}^{(v)}=\mathbf{G}_{\Omega}^{(v)}(\mathbf{R}^{(v)})^{T}/\lambda$.
while *not converged* and *number of iteration*$<T_1$ **do**
1: Update $\mathbf{C}^{(v)}=\mathbf{A}^{(v)}\mathbf{F}_{\Omega}^{(v)}+\mathbf{B}^{(v)}$
2: Calculate the SVD of $\mathbf{C}^{(v)}=\overline{\mathbf{U}}^{(v)}\overline{\mathbf{\Sigma}}^{(v)}(\overline{\mathbf{V}}^{(v)})^{T}$
3: Update $\mathbf{F}_{\Omega}^{(v)}=\overline{\mathbf{U}}^{(v)}(\overline{\mathbf{V}}^{(v)})^{T}$.
end while

Update $\mathbf{Y}$. With $\{\mathbf{F}^{(v)}\}_{v=1}^{V}$ and $\{\mathbf{R}^{(v)}\}_{v=1}^{V}$ fixed, each row $\mathbf{y}_i$ can be updated by solving the following problem:

$$\min_{\mathbf{y}_i\geqslant\mathbf{0},\mathbf{y}_i\mathbf{1}_c=1}\sum_{c=1}^{C}(y_{ic})^{\gamma}\sum_{v=1}^{V}o_{ii}^{(v)}||\mathbf{t}_{(c)}-\mathbf{f}_{i}^{(v)}\mathbf{R}^{(v)}||^{2} \tag{14}$$

Denote $q_{ic} = \sum_{v=1}^{V} o_{ii}^{(v)} ||\mathbf{t}_{(c)} - \mathbf{f}_i^{(v)} \mathbf{R}^{(v)}||^2$, which is the (i, c)-th element of matrix $\mathbf{Q} \in \mathbb{R}^{n \times C}$. When $\gamma = 1$, the optimal solution of (14) is

$$y_{ij} =< j = arg \min_{c \in [1,C]} q_{ic} > \tag{15}$$

where $< \cdot >$ is 1 if the argument is true or 0 otherwise.

When $\gamma > 1$, the Lagrangian function of the problem (14) is $\mathcal{L}_\mu = \sum_{c=1}^{C} (y_{ic})^\gamma q_{ic} - \mu(\sum_{c=1}^{C} y_{ic} - 1)$, where μ is the Lagrange multiplier. Setting the derivative of $\mathcal{L}_\mu$ with respect to y_{ic} to zero and combining the constraint $\sum_{c=1}^{C} y_{ic} = 1$, we arrive at the closed-form solution of the problem (14):

$$y_{ic} = \frac{\left(q_{ic}\right)^{\frac{1}{1-\gamma}}}{\sum_{c=1}^{C} \left(q_{ic}\right)^{\frac{1}{1-\gamma}}} \tag{16}$$

Since the proposed (6) is solved in an alternative way, we initialize $\mathbf{F}^{(v)}$ and $\mathbf{Y}$ such that $(\mathbf{F}^{(v)})^T \mathbf{O}^{(v)} \mathbf{F}^{(v)} = \mathbf{I}_C$ and $\mathbf{Y} \in \mathbf{Ind}$, where $\mathbf{Ind}$ is a set of binary clustering indicator matrices. Additionally, $\alpha^{(v)}$ can be calculated by power method [Nie *et al.*, 2017]. At last, we resort to a decision function to assign the single class label for each $\mathbf{y}_i$,

$$y_{ij} =< j = arg \max_{c \in [1,C]} y_{ic} > \tag{17}$$

In summary, the procedure of SRLC is listed in Algorithm 2.

Algorithm 2 Optimization of SRLC

Input: The data matrices $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$, the indicator matrices $\{\mathbf{O}^{(v)}\}_{v=1}^{V}$, hyper-parameters λ and γ, the maximum number of iteration T_2.
Output: The discrete indicator matrix $\mathbf{Y}$ with Eq. (17).
Initialization:
1. Construct similarity matrices $\{\mathbf{S}^{(v)}\}_{v=1}^{V}$ with Eq. (3).
2. Initialize $\mathbf{F}^{(v)}$ such that $(\mathbf{F}^{(v)})^T \mathbf{O}^{(v)} \mathbf{F}^{(v)} = \mathbf{I}_C$.
3. Initialize $\mathbf{Y} \in \mathbf{Ind}$.
4. Compute $\alpha^{(v)}$ such that $\alpha^{(v)} \mathbf{I} - \mathbf{L}_\Omega^{(v)}$ is positive definite.
while *not converged* and *number of iteration*$< T_2$ **do**
1: Update $\mathbf{R}^{(v)}$ with Eq. (9), $\forall v \in [1, V]$.
2: Update $\mathbf{F}_\Omega^{(v)}$ of $\mathbf{F}^{(v)}$ according to Alg. 1, $\forall v \in [1, V]$.
3: Update $\mathbf{Y}$ with Eq. (15) or (16).
end while

4 Theoretical Analysis

In this section, we provide the convergence guarantee and computational complexity analysis of Algorithm 2.

4.1 Convergence Guarantee

In this subsection, we first present the convergence proof of Algorithm 1 based on [Nie *et al.*, 2017]. Next, we prove that Algorithm 2 converges to a stationary point of Eq. (6).

Proposition 3. *The objective function (12) is monotonically increased with Algorithm 1.*

Proposition 4. *Algorithm 2 decreases the objective value of (6) in each iteration monotonically until it converges to a stationary point.*

Dataset	samples	views	classes
MSRC	210	6	7
Caltech7	441	6	7
Dights	2000	6	10
ORL	400	3	40
Yale	165	2	15
WebKB	1051	2	2

Table 1: Statistic of Six Real-word Datasets.

4.2 Computational Complexity

In the following, we analyze the computational complexity of SRLC. In the initialization, the construction of $\{\mathbf{S}_\Omega^{(v)}\}_{v=1}^{V}$ costs $O(\sum_v (n^{(v)})^2 d^{(v)})$. In each iteration, SRLC has three alternating steps. To update $\{\mathbf{R}^{(v)}\}_{v=1}^{V}$, since $(\mathbf{F}^{(v)})^T \mathbf{O}^{(v)} \mathbf{G} = (\mathbf{F}_\Omega^{(v)})^T \mathbf{G}_\Omega^{(v)}$, we need $O(\sum_v n^{(v)} C^2)$ for matrix multiplication and $O(C^3 V)$ for SVD. To update $\mathbf{Y}$, we pay $O(\sum_v n^{(v)} C^2)$ to compute $\mathbf{Q}$ and $O(nC)$ to calculate $\mathbf{Y}$. To update $\{\mathbf{F}_\Omega^{(v)}\}_{v=1}^{V}$ of $\{\mathbf{F}^{(v)}\}_{v=1}^{V}$, the computational complexity of the proposed Algorithm 1 is $O(\sum_v n^{(v)} C^2 t_1 + \sum_v (n^{(v)})^2 C t_1)$, where t_1 is the number of iteration of Algorithm 1. Overall, since $V \ll n^{(v)}$ and $C \ll n^{(v)}$, the computational complexity of SRLC is $O(\sum_v (n^{(v)})^2 (d^{(v)} + t_1 t_2 C))$, where t_2 is the number of iteration of Algorithm 2.

5 Experiment

In this section, we conduct experiments to verify the proposed SRLC. Firstly, we compare SRLC with five state-of-the-art methods on partial multi-view clustering task. Then we study the impact of hyper-parameters and finally present the results about convergence behavior.

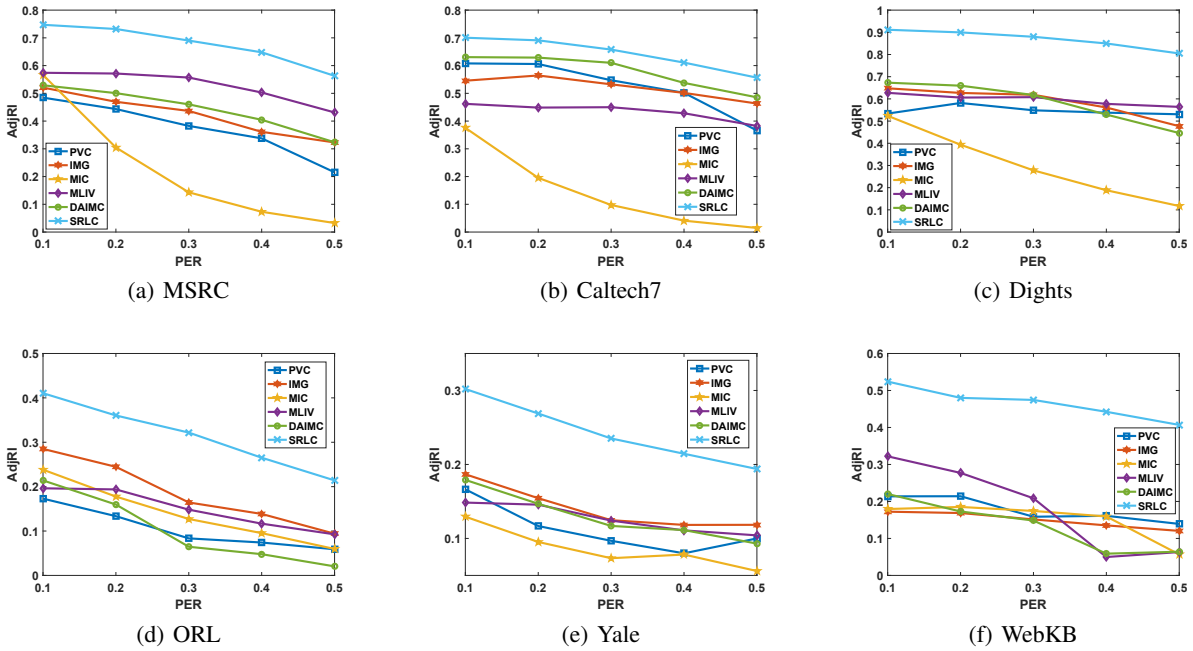
5.1 Experiments Setup

The experiments are conducted on six real-world datasets: Microsoft Research Cambridge Volume 1 (MSRC)[1], Caltech7[2], Handwritten digits (Dights)[3], ORL[4], Yale[5], WebKB[6]. A detailed summarization of these datasets is in Table 1.

Since all these datasets are originally complete, to mimic the incomplete multi-view setting, we randomly remove some examples from each sample. Concretely, for each $\mathbf{x}_i^{(v)}$, there is a probability to remove it. In the experiments, the probability is tuned form 10% to 50% with a step 10%. And for each sample $\mathbf{x}_i$, we ensure that it has at least one $\mathbf{x}_i^{(v)}$ remaining. The probability can also be regarded as the partial example ratio (PER) of the dataset.

In the experiments, we compare the proposed SRLC with several state-of-the-art methods: Partial multi-View Clustering (PVC) [Li *et al.*, 2014], Incomplete Multi-modality

[1]https://www.microsoft.com/en-us/research/project/.
[2]http://www.vision.caltech.edu/Image Datasets/Caltech101/.
[3]http://archive.ics.uci.edu/ml/datasets/Multiple+Features.
[4]http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html.
[5]http://vision.ucsd.edu/content/yale-face-database.
[6]http://www.cs.cmu.edu/afs/cs/project/theo-11/www/wwkb/.



(a) MSRC (b) Caltech7 (c) Dights

(d) ORL (e) Yale (f) WebKB

Figure 1: Mean AdjRI comparison on six datasets with different PERs.

Grouping (IMG) [Zhao *et al.*, 2016], Multiple Incomplete views Clustering (MIC) [Shao *et al.*, 2015], Multi-view Learning with Incomplete Views (MLIV) [Xu *et al.*, 2015], and Doubly Aligned Incomplete Multi-view Clustering (DAIMC) [Hu and Chen, 2018]. Since the original PVC and IMG can only deal with two incomplete views, we extend them according to Eq. (2), so that the extended versions can be applied on multi-view data with any number of incomplete views. All hyper-parameters of the compared methods are determined by grid-search.

Since all the compared methods except SRLC need post-processing to extract the clustering indicators, we apply K-means [Ding *et al.*, 2005] on the unified representation matrices of these methods to obtain the clustering the results. To make fair comparison, K-means and our proposed SRLC share the same stable initialization clustering indicator matrix which is calculated according to [Xu *et al.*, 2016].

The clustering results are evaluated by two metrics, the adjusted rand index (AdjRI) and the normalized mutual information (NMI). On each dataset, we repeat 10 independent times to create incomplete multi-view data for each PER, and the average performance is reported.

5.2 Clustering Results

Figure 1 and Table 2 report the AdjRI and NMI results on six data sets with different partial example ratios, respectively. From these figures and table, we have the following observations. (1) With the increase of PER, in terms of both AdjRI and NMI, the performance of all the compared methods becomes worse in most cases, which is consistent with intuition. (2) On datasets MSRC, Caltech7 and Dights, as the PER increases, the performance of MIC is degenerated fast than the performance of other methods. The possible reason is that MIC simply fills the missing instances with the global feature average, which may lead to a deviation when PER is large. (3) Each of the MF-based methods achieves good performance on certain datasets, but performs worse on other datasets. The possible reason is that they adopt different regularization terms or constraints, which makes them good at grouping certain kind of data and poor at clustering the others. (4) The proposed SRLC consistently outperforms the other compared methods over all datasets as the PER varies from 10% to 50%. This may be because SRLC incorporates representation learning and clustering, and is capable to explore the non-linear information hidden in the data.

5.3 Hyper-parameter Study

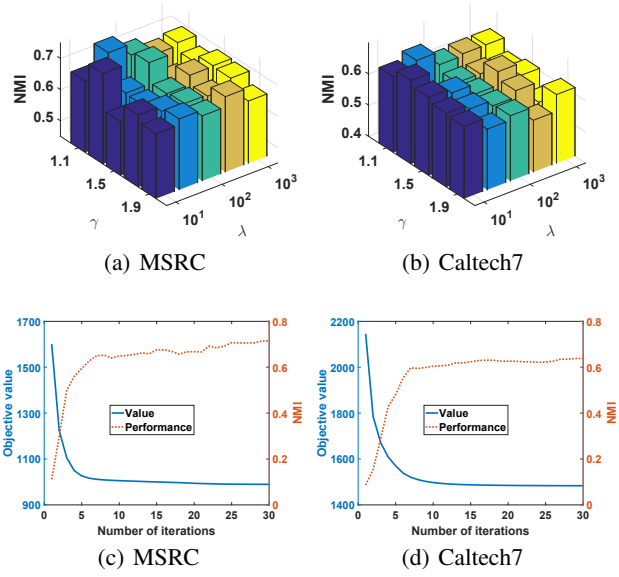
The proposed SRLC has two hyper-parameters $\{\lambda, \gamma\}$. λ is tuned in the range of $\{10^1, 10^{1.5}, 10^2, 10^{2.5}, 10^3\}$ while γ is tuned in the range of $\{1.1, 1.3, 1.5, 1.7, 1.9\}$. The experiments are conducted on datasets MSRC-v1 and Caltech7 with PER=30%. As shown in Figure 2(a) and Figure 2(b), SRLC achieves acceptable with different combinations of $\{\lambda, \gamma\}$. However, how to identify the optimal parameters is data dependent. Two datasets have different optimal parameters because their data characteristics are different.

5.4 Convergence Study

To verify the convergence property of the proposed SRLC, we conduct experiments on datasets MSRC and Caltech7 with PER=30% and set the hyper-parameters $\{\lambda, \gamma\}$ as $\{10^2, 1.1\}$ respectively. In Figure 2(c) and Figure 2(d), the blue curves show the objective function value of (6) and the red dashed lines presents the NMI of SRLC in each iteration. It can be

Data set	PER	PVC	IMG	MIC	MLIV	DAIMC	SRLC
MSRC	10%	.5837(.0525)	.6299(.0292)	.6565(.0416)	.6668(.0689)	.6332(.0597)	**.7839(.0136)**
	20%	.5506(.0478)	.5800(.0382)	.4401(.0561)	.6573(.0692)	.5956(.0238)	**.7731(.0258)**
	30%	.5024(.0430)	.5564(.0309)	.2859(.0337)	.6470(.0314)	.5639(.0433)	**.7289(.0161)**
	40%	.4502(.0641)	.5009(.0390)	.2006(.0327)	.6009(.0665)	.5011(.0514)	**.6918(.0395)**
	50%	.3491(.0451)	.4676(.0352)	.1288(.0260)	.5226(.0516)	.4361(.0824)	**.6197(.0379)**
Caltech7	10%	.6362(.0243)	.6016(.0099)	.4540(.0514)	.5234(.0568)	.6667(.0375)	**.7442(.0119)**
	20%	.6329(.0246)	.5996(.0191)	.3054(.0148)	.5140(.0531)	.6532(.0311)	**.7274(.0174)**
	30%	.5800(.0289)	.5766(.0212)	.1924(.0209)	.5105(.0437)	.6203(.0172)	**.6936(.0261)**
	40%	.5255(.0289)	.5416(.0199)	.1182(.0222)	.4727(.0454)	.5547(.0267)	**.6399(.0232)**
	50%	.3987(.0529)	.5185(.0298)	.0776(.0121)	.4382(.0472)	.5065(.0428)	**.5772(.0341)**
Dights	10%	.6450(.0177)	.7281(.0157)	.6466(.0169)	.7084(.0396)	.7348(.0249)	**.9124(.0076)**
	20%	.6730(.0203)	.7088(.0095)	.5831(.0141)	.6929(.0223)	.7258(.0385)	**.9017(.0062)**
	30%	.6470(.0320)	.7015(.0245)	.5214(.0125)	.6907(.0322)	.6902(.0216)	**.8839(.0094)**
	40%	.6224(.0394)	.6601(.0318)	.4447(.0144)	.6571(.0162)	.6220(.0481)	**.8564(.0047)**
	50%	.6085(.0229)	.6134(.0209)	.3480(.0101)	.6402(.0466)	.5437(.0669)	**.8175(.0082)**
ORL	10%	.5819(.0323)	.6763(.0046)	.6374(.0120)	.6069(.0182)	.6051(.0258)	**.7443(.0044)**
	20%	.5370(.0368)	.6409(.0127)	.5980(.0157)	.5971(.0215)	.5435(.0391)	**.7142(.0121)**
	30%	.4989(.0165)	.5883(.0222)	.5642(.0145)	.5584(.0148)	.4040(.0431)	**.6888(.0132)**
	40%	.4908(.0240)	.5680(.0204)	.5358(.0094)	.5346(.0120)	.3724(.0521)	**.6527(.0093)**
	50%	.4710(.0108)	.5292(.0143)	.5075(.0100)	.5073(.0156)	.3361(.0378)	**.6160(.0096)**
Yale	10%	.4425(.0170)	.4641(.0073)	.4127(.0116)	.4465(.0055)	.4618(.0087)	**.5538(.0050)**
	20%	.3912(.0174)	.4391(.0208)	.3816(.0133)	.4280(.0354)	.4247(.0254)	**.5322(.0235)**
	30%	.3775(.0174)	.4171(.0200)	.3697(.0230)	.4101(.0290)	.4015(.0212)	**.5046(.0115)**
	40%	.3635(.0286)	.4081(.0282)	.3760(.0295)	.3929(.0222)	.4022(.0261)	**.4913(.0177)**
	50%	.3847(.0315)	.4093(.0240)	.3635(.0255)	.3842(.0238)	.3849(.0196)	**.4706(.0265)**
WebKB	10%	.0977(.0305)	.0644(.0152)	.1106(.0117)	.2291(.1743)	.1770(.0518)	**.3898(.0849)**
	20%	.1207(.0408)	.0629(.0272)	.1116(.0241)	.1664(.1381)	.1181(.0699)	**.3647(.0295)**
	30%	.0763(.0415)	.0535(.0150)	.0913(.0286)	.1332(.1278)	.0769(.0384)	**.3233(.0689)**
	40%	.0718(.0245)	.0429(.0088)	.0770(.0181)	.0445(.0233)	.0264(.0356)	**.2856(.0940)**
	50%	.0446(.0062)	.0353(.0044)	.0143(.0104)	.0158(.0208)	.0168(.0036)	**.2736(.0418)**

Table 2: Comparison results w.r.t NMI with different PERs. (mean(std).



(a) MSRC (b) Caltech7

(c) MSRC (d) Caltech7

Figure 2: Parameter and convergence studies of SRLC.

observed that as the iteration round increases, the objective function value decreases fast and the performance increases rapidly, indicating SRLC has fast convergence propety.

6 Conclusion

In this paper, we propose a spectral-based method to deal with incomplete multi-view problem by incorporating representation learning and clustering. Based on the similarity matrices independently constructed on each view, the proposed SRLC algorithm learns a common probability label matrix together with low-dimensional representations of present instances, which enables SRLC to utilize non-linear information and facilitates the optimization procedure to meet the demand of clustering. Experimental results on six datasets validate the effectiveness of SRLC. In the future, we will study how to extend SRLC to the semi-supervised classification task.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61473302, 61503396), and the NSF for Distinguished Young Scholars of Hunan Province under Grant 2019JJ20020.

References

[Cai *et al.*, 2018] Yang Cai, Yuanyuan Jiao, Wenzhang Zhuge, Tao Hong, and Chenping Hou. Partial multi-view spectral clustering. *Neurocomputing*, 311:316–324, 2018.

[Ding *et al.*, 2005] Chris Ding, Xiaofeng He, and Horst D. Simon. Nonnegative lagrangian relaxation of k-means and spectral clustering. In *ECML*, pages 530–538, 2005.

[Hastie *et al.*, 2010] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. The elements of statistical learning. *Technometrics*, 45(3):267–268, 2010.

[Hou *et al.*, 2010] Chenping Hou, Changshui Zhang, Yi Wu, and Feiping Nie. Multiple view semi-supervised dimensionality reduction. *Pattern Recognition*, 43(3):720–730, 2010.

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

[Jing *et al.*, 2017] Zhao Jing, Xijiong Xie, Xu Xin, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[Karasuyama and Mamitsuka, 2013] Masayuki Karasuyama and Hiroshi Mamitsuka. Multiple graph label propagation by sparse integration. *IEEE Transactions on Neural Networks and Learning Systems*, 24(12):1999–2012, 2013.

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

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

[Nie *et al.*, 2016b] Feiping Nie, Xiaoqian Wang, Michael I. Jordan, and Heng Huang. The constrained laplacian rank algorithm for graph-based clustering. In *AAAI*, pages 1969–1976, 2016.

[Nie *et al.*, 2017] Feiping Nie, Rui Zhang, and Xuelong Li. A generalized power iteration method for solving quadratic problem on the stiefel manifold. *Science China Information Sciences*, 60(11):146–155, 2017.

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

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

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

[Tao *et al.*, 2018] Hong Tao, Chenping Hou, Dongyun Yi, and Jubo Zhu. Unsupervised maximum margin incomplete multi-view clustering. In *ICAI*, pages 13–25, 2018.

[Xu *et al.*, 2015] Chang. Xu, Dacheng Tao, and Chao Xu. Multi-view learning with incomplete views. *IEEE Transactions on Image Processing*, 24(12):5812–5825, 2015.

[Xu *et al.*, 2016] Jinglin Xu, Junwei Han, and Feiping Nie. Discriminatively embedded k-means for multi-view clustering. In *CVPR*, pages 5356–5364, 2016.

[Yin *et al.*, 2015] Qiyue Yin, Shu Wu, and Liang Wang. Incomplete multi-view clustering via subspace learning. In *CIKM*, pages 383–392, 2015.

[Yin *et al.*, 2017] Qiyue Yin, Shu Wu, and Liang Wang. Unified subspace learning for incomplete and unlabeled multi-view data. *Pattern Recognition*, 67(67):313–327, 2017.

[Zelnik-Manor and Perona, 2004] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. *Advances in Neural Information Processing Systems*, 17:1601–1608, 2004.

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

[Zhao *et al.*, 2018] Liang Zhao, Zhikui Chen, Yi Yang, Z Jane Wang, and Victor CM Leung. Incomplete multi-view clustering via deep semantic mapping. *Neurocomputing*, 275:1053–1062, 2018.