Multi-view Clustering in Latent Embedding Space

Abstract

Previous multi-view clustering algorithms mostly partition the multi-view data in their original feature space, the efficacy of which heavily and implicitly relies on the quality of the original feature presentation. In light of this, this paper proposes a novel approach termed Multi-view Clustering in Latent Embedding Space (MCLES), which is able to cluster the multi-view data in a learned latent embedding space while simultaneously learning the global structure and the cluster indicator matrix in a unified optimization framework. Specifically, in our framework, a latent embedding representation is firstly discovered which can effectively exploit the complementary information from different views. The global structure learning is then performed based on the learned latent embedding representation. Further, the cluster indicator matrix can be acquired directly with the learned global structure. An alternating optimization scheme is introduced to solve the optimization problem. Extensive experiments conducted on several real-world multi-view datasets have demonstrated the superiority of our approach.

Introduction

In the past decade, multi-view clustering has become a hot research topic in data mining and machine learning, due to the rapid emergence of a great deal of multi-view data from different areas (Xu, Tao, and Xu 2013; Xu, Wang, and Lai 2016; Tao et al. 2018; Huang, Chao, and Wang 2019; Xing et al. 2019; Wang et al. 2019; Yao et al. 2019). In multi-view data, the same instance is represented by multiple views obtaining from multiple sources or different feature subsets. For instance, in a webpage, different types of data, such as texts, videos and images, can be taken into consideration as they are different aspects of the webpage. A text news can be translated into multiple languages. Considering the diversity of multiple views, it is essential to study how to integrate such kind of data efficiently and cluster them effectively.

Prior to the most multi-view learning methods, a direct way to deal with multi-view data is to concatenate all the features into a new feature vector, which is then fed into a single-view clustering method to obtain the final clustering results. However, this naive strategy neglects the different characteristics as well as the correlation among multiple views. Recently, a large number of multi-view clustering methods have been proposed to handle multi-view data by effectively considering the rich information from multiple views (Cai et al. 2011; Kumar and Daumé III 2011; Kumar, Rai, and Daumé III 2011; Xia et al. 2014; Zhang et al. 2017; Zhan et al. 2018; Zhang et al. 2018; Huang et al. 2019). For instance, to minimize the disagreement between each pair of views, a co-regularization technique was introduced in multi-view spectral clustering (Kumar, Rai, and Daumé III 2011). Similarly, inspired by the idea of co-training, Kumar et al. proposed to generate clusters that are consistent across the multiple views (Kumar and Daumé III 2011). However, these methods may be easily affected by the poor quality of the original views, thereby resulting in the degraded clustering performance. Besides, most of these methods directly compute on the original features from the dataset, in which there may be gross noise and corruption. To handle the possible noise, Zhang et al. performed data reconstruction based on the learned subspace (Zhang et al. 2017). Xia et al. developed a Markov chain method which takes as input a shared low-rank transition probability matrix associated with all views (Xia et al. 2014). Nevertheless, on the one hand, these methods generally rely on the original features in each view, but still lack the ability to discover a unified feature representation for multi-view data. On the other hand, in the spectral clustering phase, they mostly tend to consider the two components of spectral clustering (i.e., affinity matrix construction and cluster indicator matrix calculation) separately, but often lack the ability to formulate these two components simultaneously in an optimization framework.

Aiming to address the above limitations, in this paper, we propose a unified framework termed Multi-view Clustering in Latent Embedding Space (MCLES). The proposed method jointly learns the latent embedding representation, the similarity information and the cluster indicator matrix in a unified model. The latent embedding space learned from the multi-view features is able to explore the relationships among different samples and avoid the possible corruption

as well as the curse of dimensionality. With the idea of self-expression, the similarity matrix is constructed based on the learned latent embedding representation rather than the original features of data. Further, the cluster indicator matrix is directly learned without the additional procedure of spectral clustering. The main contributions of this paper are summarized as follows:

- We propose a novel multi-view clustering approach termed MCLES, which jointly learns a latent embedding space, a robust similarity matrix and an accurate cluster indicator matrix in a unified optimization framework.
- By leveraging the intrinsic interactions among them, our framework extracts the global structure based on the learned latent embedding representation, and further acquires the cluster indicator matrix based on the global structure.
- An alternating optimization scheme is developed to efficiently deal with the optimization problem.
- Extensive experiments on image and document datasets have demonstrated the superiority of the proposed method when compared with the state-of-the-art approaches.

Related Work

In the past few years, many multi-view clustering methods have been proposed, which can be classified into three main categories, i.e., the co-training based methods, the multiple kernel learning based methods, and the subspace learning based methods (Xu, Tao, and Xu 2013).

The co-training based methods try to maximize the agreement among different views of the data in an alternate training manner (Blum and Mitchell 1998; Ghani 2002; Brefeld and Scheffer 2004; Kumar, Rai, and Daumé III 2011; Kumar and Daumé III 2011). Based on the co-training technique, several unsupervised multi-view clustering methods were proposed (Kumar, Rai, and Daumé III 2011; Kumar and Daumé III 2011). These methods constructed the graph beforehand, and then separately performed the data clustering. The graph constructed on the original feature space of multiple views may not be suitable for the subsequent clustering, and may be negatively affected by some original single-view features.

The multiple kernel learning based methods linearly or non-linearly combine kernels corresponding to different views to improve their performance (Cortes, Mohri, and Rostamizadeh 2009; Gönen and Alpaydın 2011; Tzortzis and Likas 2012). In (Cortes, Mohri, and Rostamizadeh 2009), multiple kernels are directly and effectively combined for multi-view clustering. The study in (Tzortzis and Likas 2012) proposed to weight different kernels before the combination. These studies performed multiple kernel learning with each kernel corresponding to each original single view, which makes the clustering performance highly dependent on the quality of the original views.

According to the assumption that the multiple views are drawn from a latent subspace, the subspace learning based methods expect to seek for the common latent subspace shared by different views (Elhamifar and Vidal 2009; Chaudhuri et al. 2009; Liu et al. 2012; Patel, Van Nguyen, and Vidal 2013; Xia et al. 2014; Zhang et al. 2017; Li et al. 2019). In (Patel, Van Nguyen, and Vidal 2013), dimensionality reduction and sparse coding for sparse subspace clustering (SSC) were performed, in which the dimensionality reduction was performed on the original single-view features. Zhang et al. proposed to simultaneously recover the underlying multi-view subspace and the projections associated to different views (Zhang et al. 2017).

In this paper, different from the aforementioned methods which rely on the original feature space or consider the two components of spectral clustering seperately (i.e., affinity matrix construction and cluster indicator matrix calculation), our proposed method jointly learns a latent embedding space, a robust similarity matrix and an accurate cluster indicator matrix in a unified optimization framework, where each variable boosts each other in an interplay manner to achieve the optimal solution.

The Proposed Approach

In this section, the proposed MCLES approach will be described in detail. First, the preliminary knowledge of global similarity learning will be briefly introduced. Then we will describe the proposed model and its optimization scheme in detail. Finally, we will summarize the overall algorithm and provide the time complexity analysis.

In this paper, matrices are written as uppercase letters. The j-th column and the (i,j)-th entry of matrix $\mathbf{X}$ are denoted as $\mathbf{X}_{:,j}$ and x_{ij} respectively. The v-th view of the matrix $\mathbf{X}$ is expressed as $\mathbf{X}^{(v)}$. $Tr(\mathbf{X})$ stands for the trace of the matrix $\mathbf{X}$. The l_2-norm and the Frobenius norm of the matrix $\mathbf{X}$ are respectively $\|\mathbf{X}\|_2$ and $\|\mathbf{X}\|_F$. In addition, $\mathbf{1}$ denotes a column vector whose elements are all one.

Global Similarity Learning

Similar to Locally Linear Embedding (LLE) (Roweis and Saul 2000), it is assumed that data points lie on a manifold and each data point can be expressed as a linear combination of the other data points. According to the self-expressive property (Elhamifar and Vidal 2009), we have

$$\mathbf{X}_{:,i} \approx \sum_j \mathbf{X}_{:,j} s_{ij}, \ \text{ s.t. } \mathbf{S}_{:,i}^T \mathbf{1} = 1, 0 \leq \mathbf{S} \leq 1, \quad (1)$$

where s_{ij} is the weight for the i-th sample corresponding to the j-th sample. The weights should be larger for more similar data points, and less similar points should be assigned smaller weights. Therefore, $\mathbf{S}$ is called the learned similarity matrix, which reflects the global structure of data. Note that there is no neighborhood specified in Eq. (1), and it can be automatically determined during the learning procedure, which is different from LLE.

The Proposed Model

Inspired by (White et al. 2012; Guo 2013), it is assumed that the multiple views are drawn from one underlying latent representation, which describes the data intrinsically and discovers the shared latent structure among different views. According to this assumption, in this paper, we consider the latent embedding space in a lower dimension among multiple

views for data clustering, instead of directly using the original single-view feature space. Given a multi-view dataset $\mathbf{X} = \left\{ \left[\mathbf{x}_i^{(1)}; \mathbf{x}_i^{(2)}; ...; \mathbf{x}_i^{(V)} \right] \right\}_{i=1}^{N}$ consisting of N samples represented by V different views, our method aims to discover a shared latent embedding representation $\mathbf{h}_i$ for each data point such that all these different views are drawn from the latent embedding space $\mathbf{H}$. To be specific, the shared latent embedding space $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^{N}$ can be employed to obtain the observation from multiple views by using the corresponding mapping model $\mathbf{W} = \left\{\mathbf{W}^{(v)}\right\}_{v=1}^{V}$, which can be denoted as

$$\mathbf{x}_i^{(v)} = \mathbf{W}^{(v)}\mathbf{h}_i. \tag{2}$$

To obtain $\mathbf{H}$, the following problem needs to be solved:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2, \\ &\text{s.t. } \|\mathbf{W}_{:,j}\|_2^2 \leq 1, \\ &\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \cdots \\ \mathbf{X}^{(V)} \end{bmatrix} \text{and } \mathbf{W} = \begin{bmatrix} \mathbf{W}^{(1)} \\ \cdots \\ \mathbf{W}^{(V)} \end{bmatrix}, \end{aligned} \tag{3}$$

where $\mathbf{X}$ and $\mathbf{W}$ are respectively the multi-view observations and mapping models. The constraint of $\mathbf{W}$ is to prevent $\mathbf{H}$ from being too small while scaling. In general, the latent embedding space $\mathbf{H}$ will be more comprehensive by combining the complementary information from multiple views. Unlike most of the methods that construct the similarity matrix based on the original feature space of the data (Kumar, Rai, and Daumé III 2011; Kumar and Daumé III 2011; Zhan et al. 2018), the proposed method learns the similarity matrix from the latent embedding space, which effectively enhances the robustness and accuracy of the learned similarity matrix. According to the global affinity learning, we have

$$\begin{aligned} &\min_{\mathbf{S}} \|\mathbf{H} - \mathbf{H}\mathbf{S}\|_F^2 + \beta\|\mathbf{S}\|_F^2, \\ &\text{s.t. } \mathbf{S}_{:,i}^T\mathbf{1} = 1, 0 \leq \mathbf{S} \leq 1, \end{aligned} \tag{4}$$

where β is the trade-off parameter. Ideally, the number of connected components in $\mathbf{S}$ is expected to be the same as the cluster number of dataset $\mathbf{X}$, i.e. c. In other words, $\mathbf{S}$ is a block diagonal matrix with proper permutations. However, the solution $\mathbf{S}$ in Eq. (4) may not satisfy the desired property. To settle this problem, the rank constraint needs to be introduced based on the following theorem (Mohar et al. 1991).

Theorem 1. *The multiplicity c of the eigenvalue 0 of the Laplacian matrix $\mathbf{L_s}$ of $\mathbf{S}$ is equal to the number of connected components in the graph with the similarity matrix $\mathbf{S}$.*

Theorem 1 shows that if the similarity matrix $\mathbf{S}$ consists of exactly c connected components, we can have $rank(\mathbf{L_s}) = n - c$. Therefore, the problem in Eq. (4) can be reformulated as

$$\begin{aligned} &\min_{\mathbf{S}} \|\mathbf{H} - \mathbf{H}\mathbf{S}\|_F^2 + \beta\|\mathbf{S}\|_F^2, \\ &\text{s.t. } \mathbf{S}_{:,i}^T\mathbf{1} = 1, 0 \leq \mathbf{S} \leq 1, rank(\mathbf{L_s}) = n - c. \end{aligned} \tag{5}$$

By integrating the latent embedding learning in Eq. (3) and the global similarity learning in Eq. (5) into a unified framework, we have

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{H},\mathbf{S}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \alpha\|\mathbf{H} - \mathbf{H}\mathbf{S}\|_F^2 + \beta\|\mathbf{S}\|_F^2, \\ &\text{s.t. } \|\mathbf{W}_{:,j}\|_2^2 \leq 1, \mathbf{S}_{:,i}^T\mathbf{1} = 1, 0 \leq \mathbf{S} \leq 1, rank(\mathbf{L_s}) = n - c, \\ &\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \cdots \\ \mathbf{X}^{(V)} \end{bmatrix} \text{and } \mathbf{W} = \begin{bmatrix} \mathbf{W}^{(1)} \\ \cdots \\ \mathbf{W}^{(V)} \end{bmatrix}, \end{aligned} \tag{6}$$

where $\alpha > 0$ and $\beta > 0$ balance these three terms. As a matter of fact, it is not easy to tackle the problem in Eq. (6), since $\mathbf{L_s} = \mathbf{D} - \frac{\mathbf{S}^T + \mathbf{S}}{2}$, in which $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with the i-th diagonal element being $\sum_j \frac{s_{ij} + s_{ji}}{2}$.

Since $\mathbf{L_s}$ is positive semi-definite, we have $\sigma_i(\mathbf{L_s}) \geq 0$, in which $\sigma_i(\mathbf{L_s})$ represents the i-th smallest eigenvalue of $\mathbf{L_s}$. It is well-known that the optimization problem with rank constraint is of combinatorial complexity (Kang, Peng, and Cheng 2017). To solve this problem, it is suggested to incorporate the rank constraint into the objective function as a regularization term (Wang et al. 2015; Nie et al. 2016). According to (Mohar et al. 1991), $rank(\mathbf{L_s}) = n - c$ is equivalent to $\sum_{i=1}^{c} \sigma_i(\mathbf{L_s}) = 0$. Thus, the constraint is relaxed and the problem in Eq. (6) can be rewritten as

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{H},\mathbf{S}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \alpha\|\mathbf{H} - \mathbf{H}\mathbf{S}\|_F^2 \\ &\qquad + \beta\|\mathbf{S}\|_F^2 + \gamma\sum_{i=1}^{c} \sigma_i(\mathbf{L_s}), \\ &\text{s.t. } \|\mathbf{W}_{:,j}\|_2^2 \leq 1, \mathbf{S}_{:,i}^T\mathbf{1} = 1, 0 \leq \mathbf{S} \leq 1, \\ &\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \cdots \\ \mathbf{X}^{(V)} \end{bmatrix} \text{and } \mathbf{W} = \begin{bmatrix} \mathbf{W}^{(1)} \\ \cdots \\ \mathbf{W}^{(V)} \end{bmatrix}. \end{aligned} \tag{7}$$

If γ is large enough, the minimization above will make the regularization term $\sum_{i=1}^{c} \sigma_i(\mathbf{L_s}) \to 0$, which will satisfy the constraint $rank(\mathbf{L_s}) = n - c$. Despite of this, the optimization problem in Eq. (7) is still challenging due to the last term. To mitigate this problem, we introduce the Ky Fan's Theorem (Fan 1949). That is,

$$\sum_{i=1}^{c} \sigma_i(\mathbf{L_s}) = \min_{\mathbf{P}^T\mathbf{P}} Tr\left(\mathbf{P}^T\mathbf{L_s}\mathbf{P}\right), \tag{8}$$

where $\mathbf{P} \in \mathbb{R}^{n \times c}$ is the cluster indicator matrix. Therefore, the problem in Eq. (7) is finally reformulated as

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H},\mathbf{S}} & \underbrace{\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2}_{\text{latent embedding learning}} + \underbrace{\alpha\|\mathbf{H} - \mathbf{H}\mathbf{S}\|_F^2 + \beta\|\mathbf{S}\|_F^2}_{\text{global similarity learning}} \\ &+ \underbrace{\gamma Tr\left(\mathbf{P}^T\mathbf{L_s}\mathbf{P}\right)}_{\text{cluster indicator learning}}, \\ \text{s.t. } &\|\mathbf{W}_{:,j}\|_2^2 \leq 1, \mathbf{S}_{:,i}^T\mathbf{1} = 1, 0 \leq \mathbf{S} \leq 1, \mathbf{P}^T\mathbf{P} = \mathbf{I}, \\ &\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \cdots \\ \mathbf{X}^{(V)} \end{bmatrix} \text{and } \mathbf{W} = \begin{bmatrix} \mathbf{W}^{(1)} \\ \cdots \\ \mathbf{W}^{(V)} \end{bmatrix}, \end{aligned} \tag{9}$$

where there are three parameters $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ for balancing these four terms. The first term is to model the latent embedding space $\mathbf{H}$ and each mapping model $\mathbf{W}^{(v)}$ for reconstructing the observations. The second term is to penalize the construction error in similarity learning. The third term is used to avoid the trivial solution $\mathbf{S} = \mathbf{I}$. The last term is to guarantee the similarity matrix to meet the rank constraint, and directly obtain the cluster indicator matrix $\mathbf{P}$. In this triangle relationship, the latent embedding learning is guaranteed by the complementary information among multiple views and improved by the global similarity learning and the cluster indicator learning. The global similarity learning is guaranteed by the latent embedding learning and the cluster indicator learning. And the cluster indicator learning is guaranteed by the latent embedding learning and the global similarity learning. As a matter of fact, there is a mutual self-taught property in our unified framework because of the feedback among the latent embedding representation, the similarity matrix and the cluster indicator matrix.

Optimization

In this subsection, an alternating optimization scheme is introduced to solve the problem in Eq. (9), i.e., one variable is updated while fixing the others in one iteration.

Update $\mathbf{W}$ By fixing all the variables except $\mathbf{W}$, the problem in Eq. (9) can be reduced to solve the following problem,

$$\begin{aligned} &\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2, \\ &\text{s.t. } \|\mathbf{W}_{:,j}\|_2^2 \leq 1, \mathbf{W} = \begin{bmatrix} \mathbf{W}^{(1)} \\ \cdots \\ \mathbf{W}^{(V)} \end{bmatrix}. \end{aligned} \tag{10}$$

Specifically, this problem can be optimized by introducing a variable $\mathbf{G}$,

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{G}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2, \\ &\text{s.t. } \mathbf{W} = \mathbf{G}, \|\mathbf{G}_{:,j}\|_2^2 \leq 1. \end{aligned} \tag{11}$$

The optimal solution of Eq. (11) can be obtained by the Alternating Direction Method of Multipliers (ADMM) algorithm (Gu et al. 2014):

$$\begin{cases} \mathbf{W}^{r+1} = \arg\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \rho\|\mathbf{W} - \mathbf{G}^r + \mathbf{T}^r\|_F^2, \\ \mathbf{G}^{r+1} = \arg\min_{\mathbf{G}} \rho\|\mathbf{W}^{r+1} - \mathbf{G} + \mathbf{T}^r\|_F^2, \text{s.t. } \|\mathbf{G}_{:,j}\|_2^2 \leq 1, \\ \mathbf{T}^{r+1} = \mathbf{T}^r + \mathbf{W}^{r+1} - \mathbf{G}^{r+1}, \quad \text{update } \rho \text{ if appropriate}, \end{cases} \tag{12}$$

where r stands for the steps of iterations. In each step of optimization, the ADMM based optimization of $\mathbf{W}$ converges rapidly due to the good convergent performance of the ADMM algorithm.

Update $\mathbf{H}$ By fixing all the variables except $\mathbf{H}$, it is equivalent to solving

$$\min_{\mathbf{H}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \alpha\|\mathbf{H} - \mathbf{H}\mathbf{S}\|_F^2. \tag{13}$$

The optimal solution $\mathbf{H}^\star$ can be obtained by differentiating Eq. (13) with respect to $\mathbf{H}$ and setting it to zero, which satisfies

$$\mathbf{W}^T\mathbf{W}\mathbf{H}^\star + \mathbf{H}^\star * \alpha\left(\mathbf{I} - \mathbf{S}\right)\left(\mathbf{I} - \mathbf{S}\right)^T = \mathbf{W}^T\mathbf{X}. \tag{14}$$

The above equation is a standard Sylvester equation which has a unique solution and can be solved by the Bartels-Stewart algorithm (Bartels and Stewart 1972). Therefore, the similar method to the smooth subspace clustering (Hu et al. 2014) can be used to optimize our latent embedding representation $\mathbf{H}$.

Update $\mathbf{S}$ By fixing all the variables except $\mathbf{S}$, the problem in Eq. (9) can be written as

$$\begin{aligned} &\min_{\mathbf{S}} \|\mathbf{H} - \mathbf{H}\mathbf{S}\|_F^2 + \frac{\beta}{\alpha}\|\mathbf{S}\|_F^2 + \frac{\gamma}{\alpha}Tr\left(\mathbf{P}^T\mathbf{L_s}\mathbf{P}\right), \\ &\text{s.t. } \mathbf{S}_{:,i}^T\mathbf{1} = 1, 0 \leq \mathbf{S} \leq 1. \end{aligned} \tag{15}$$

For convenience, we introduce a variable $\mathbf{K}$, which is equal to $\mathbf{H}^T\mathbf{H}$. Thus, the problem in Eq. (15) can be reformulated as

$$\begin{aligned} &\min_{\mathbf{S}} Tr\left(\mathbf{K} - 2\mathbf{K}\mathbf{S} + \mathbf{S}^T\mathbf{K}\mathbf{S}\right) + \frac{\beta}{\alpha}\|\mathbf{S}\|_F^2 + \frac{\gamma}{\alpha}Tr\left(\mathbf{P}^T\mathbf{L_s}\mathbf{P}\right), \\ &\text{s.t. } \mathbf{S}_{:,i}^T\mathbf{1} = 1, 0 \leq \mathbf{S} \leq 1. \end{aligned} \tag{16}$$

The problem in Eq. (16) can be rewritten in a column-wise manner as

$$\begin{aligned} &\min_{\mathbf{S}_{:,i}} \mathbf{K}_{ii} - 2\mathbf{K}_{i,:}\mathbf{S}_{:,i} + \mathbf{S}_{:,i}^T\mathbf{K}\mathbf{S}_{:,i} + \frac{\beta}{\alpha}\mathbf{S}_{:,i}^T\mathbf{S}_{:,i} + \frac{\gamma}{2\alpha}\mathbf{b}_i^T\mathbf{S}_{:,i}, \\ &\text{s.t. } \mathbf{S}_{:,i}^T\mathbf{1} = 1, 0 \leq \mathbf{S} \leq 1, \end{aligned} \tag{17}$$

where $\mathbf{b}_i \in \mathbb{R}^{n\times 1}$ is a column vector with the j-th element $\mathbf{b}_{ij}$ being $\mathbf{b}_{ij} = \|\mathbf{P}_{i,:} - \mathbf{P}_{j,:}\|^2$.

Specifically, the problem in Eq. (17) can be further simplified as follows,

$$\begin{aligned} &\min_{\mathbf{S}_{:,i}} \mathbf{S}_{:,i}^T\left(\frac{\beta}{\alpha}\mathbf{I} + \mathbf{K}\right)\mathbf{S}_{:,i} + \left(\frac{\gamma\mathbf{b}_i^T}{2\alpha} - 2\mathbf{K}_{i,:}\right)\mathbf{S}_{:,i}, \\ &\text{s.t. } \mathbf{S}_{:,i}^T\mathbf{1} = 1, 0 \leq \mathbf{S} \leq 1. \end{aligned} \tag{18}$$

Many existing quadratic programming packages (Kang, Peng, and Cheng 2017) can be utilized to solve the problem in Eq. (18).

Update $\mathbf{P}$ By fixing all the variables except $\mathbf{P}$, the problem in Eq. (9) becomes

$$\begin{aligned} &\min_{\mathbf{P}} Tr\left(\mathbf{P}^T\mathbf{L_s}\mathbf{P}\right), \\ &\text{s.t. } \mathbf{P}^T\mathbf{P} = \mathbf{I}. \end{aligned} \tag{19}$$

The optimal solution $\mathbf{P}$ can be obtained by the c eigenvectors of $\mathbf{L_s}$ corresponding to the c smallest eigenvalues. With the help of the alternating optimization scheme, the variables $\mathbf{W}$, $\mathbf{H}$, $\mathbf{S}$ and $\mathbf{P}$ can be updated iteratively in an interplay manner until convergence.

Algorithm 1 Multi-view Clustering in Latent Embedding Space

Input: Multi-view matrices $\mathbf{X} = \left\{\mathbf{X}^{(1)}, ..., \mathbf{X}^{(V)}\right\}$, cluster number c, parameters α, β and γ, and the embedding dimension d of latent representation.

Initialize: $\mathbf{W} = 0$, $\mathbf{S} = 0$, $\mathbf{P} = 0$, and $r = 0$; Initialize $\mathbf{H}$ with random values.

repeat
  repeat
    $r \leftarrow r + 1$;
    Update $\mathbf{W}$ according to Eq. (12).
  until convergence
  Update $\mathbf{H}$ according to Eq. (14).
  repeat
    For each i, update the i-th column of $\mathbf{S}$ by solving the problem in Eq. (18).
  until convergence
  Update $\mathbf{P}$, which is formed by the c eigenvectors of $\mathbf{L_s} = \mathbf{D} - \frac{\mathbf{S}^T + \mathbf{S}}{2}$ corresponding to the c smallest eigenvalues.
until convergence

Output: $\mathbf{W}$, $\mathbf{H}$, $\mathbf{S}$ and $\mathbf{P}$.

Algorithm Summary and Complexity Analysis

For clarity, the overall algorithm of the proposed MCLES method is outlined in Algorithm 1. In what follows, the time complexity analysis will be provided. With the alternating optimization scheme, the ADMM algorithm is utilized for updating $\mathbf{W}$, the complexity of which is $O\Big(\big(\sum_{v=1}^{V} d^{(v)}\big)^2 d\Big)$ where $d^{(v)}$ is the dimension of the v-th view of the dataset and d is the dimension of the latent embedding representation. The computation of $\mathbf{H}$ requires $O\left(d^3\right)$. To update $\mathbf{S}$, the quadratic programming takes $O\left(n^2\right)$. The complexity for $\mathbf{P}$ is $O\left(cn^2\right)$. Therefore, for each iteration, the overall computational complexity is $O\Big(\big(\sum_{v=1}^{V} d^{(v)}\big)^2 d + d^3 + (1+c)n^2\Big)$.

Experiments

In this section, extensive experiments are conducted on several real-world datasets including three image datasets and one document dataset to validate the superiority of the proposed method. We compare the performance of the proposed MCLES method with six state-of-the-art multi-view clustering methods and two single-view baseline methods in terms of three evaluation metrics.

Datasets Description

Yale[1]: It is a widely used face image dataset consisting of 165 gray-scale images belonging to 15 distinct subjects, with each subject consisting of 11 images. Variations of the dataset are composed of left light, center light, right light, with glasses or not, happy or sad, normal, sleepy, wink and surprised. In our experiments, three views are used, whose dimensions are respectively 4096, 3304 and 6750.

Table 1: The default values of the four parameters.

Parameter	Yale	MSRCv1	ORL	BBCSport
d	30	70	50	40
α	0.8	0.8	0.8	0.8
β	0.4	0.4	0.5	0.4
γ	0.004	0.004	0.004	0.004

MSRCv1 (Winn and Jojic 2005): It is an image dataset consisting of 210 objects belonging to seven classes. The seven classes are composed of tree, building, airplane, cow, face, car, and bicycle. In our experiments, the MSRCv1 dataset consists of four views, which are the CM feature (view 1), the GIST feature (view 2), the LBP feature (view 3) and the GENT feature (view 4).

ORL[2]: It is a widely used face image dataset consisting of 400 face images belonging to 40 distinct subjects with 10 images for each subject. For each subject, images were taken at different times, lights, facial expression (open or closed eyes, smiling or not smiling) and facial details (with glasses or not). In our experiments, three kinds of features, namely, intensity feature (view 1), LBP feature (view 2), and Gabor feature (view 3) are used to represent the images.

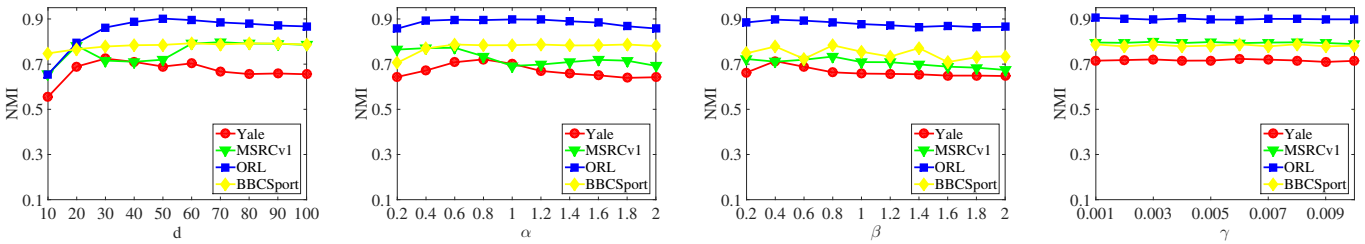
BBCSport (Xia et al. 2014): It is a document dataset consisting of 544 documents from the BBC Sport website of the sports news in five topical areas in 2004-2005. The five topical areas are composed of business, entertainment, politics, sport and tech. In our experiments, the BBCSport dataset contains two views, whose dimensions are respectively 3183 and 3203.

Baselines and Evaluation Metrics

In comparison experiments, we compare our method with two classical single-view methods and six state-of-the-art multi-view clustering methods.

1. Spectral Clustering (SC) (Ng, Jordan, and Weiss 2002): The classical single-view SC method is conducted on each view.
2. ConcatePCA-SC: It is an extended SC method which firstly concatenates the features of all views and then applies the PCA method to extract the low-dimensional representation, and finally feeds the low-dimensional representation into the spectral clustering algorithm to obtain the final clustering results.
3. Co-Regularized Spectral Clustering (Co-Reg) (Kumar, Rai, and Daumé III 2011): It co-regularizes the clustering hypotheses to enforce the same cluster membership among views.
4. Co-Training Multi-view Clustering (Co-Training) (Kumar and Daumé III 2011): It assumes that a point would be assigned by the true underlying clustering to the same cluster regardless of the views.
5. Spectral Clustering with two views (Min-Disagreement) (De Sa 2005): Based on the spectral clustering algorithm, it constructs a bipartite graph with the "minimizing-disagreement" strategy.

[1] http://cvc.yale.edu/projects/yalefaces/yalefaces.html
[2] http://www.cl.cam.ac.uk/research/dtg/



(a) Parameter d (b) Parameter α (c) Parameter β (d) Parameter γ

Figure 1: Parameters analysis on d, α, β and γ in terms of NMI on all the four benchmark datasets.

Table 2: Comparison results in terms of ACC on all datasets. The best results are highlighted in bold.

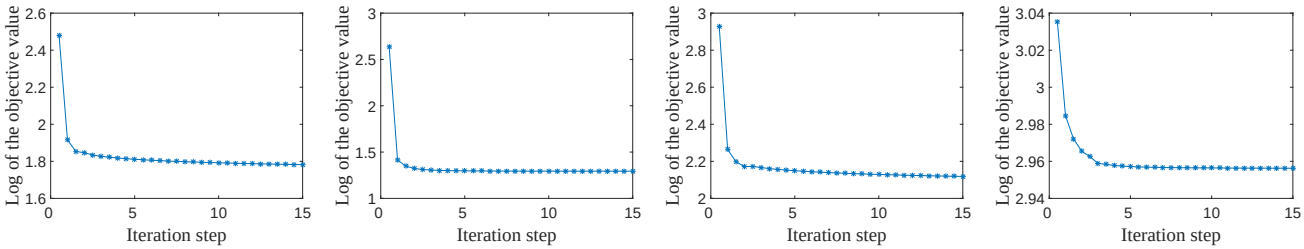
Method	Yale	MSRCv1	ORL	BBCSport
SC1	$0.5497_{\pm 0.0351}$	$0.4105_{\pm 0.0151}$	$0.6571_{\pm 0.0242}$	$0.8453_{\pm 0.0012}$
SC2	$0.5630_{\pm 0.0346}$	$0.6845_{\pm 0.0446}$	$0.7735_{\pm 0.0261}$	$0.5114_{\pm 0.0011}$
SC3	$0.6318_{\pm 0.0346}$	$0.6167_{\pm 0.0045}$	$0.6973_{\pm 0.0319}$	—
SC4	—	$0.6945_{\pm 0.0186}$	—	—
ConcatePCA-SC	$0.5618_{\pm 0.0404}$	$0.6155_{\pm 0.0067}$	$0.6607_{\pm 0.0215}$	—
Co-Reg	$0.5956_{\pm 0.0055}$	$0.6233_{\pm 0.0057}$	$0.6921_{\pm 0.0037}$	$0.6928_{\pm 0.0070}$
Co-training	$0.6223_{\pm 0.0039}$	$0.6918_{\pm 0.0099}$	$0.7539_{\pm 0.0058}$	$0.6979_{\pm 0.0039}$
Min-Disagreement	$0.5974_{\pm 0.0066}$	$0.5923_{\pm 0.0071}$	$0.7259_{\pm 0.0062}$	$0.8507_{\pm 0.0087}$
RMSC	$0.5625_{\pm 0.0426}$	$0.2998_{\pm 0.0189}$	$0.7603_{\pm 0.0259}$	$0.7737_{\pm 0.0098}$
LMSC	$0.6673_{\pm 0.0176}$	$0.6743_{\pm 0.0591}$	$\mathbf{0.8013}_{\pm 0.0333}$	$0.8512_{\pm 0.1203}$
MVGL	$0.6303_{\pm 0.0000}$	$0.6714_{\pm 0.0000}$	$0.7350_{\pm 0.0000}$	$0.4191_{\pm 0.0000}$
MCLES	$\mathbf{0.6945}_{\pm 0.0212}$	$\mathbf{0.8814}_{\pm 0.0055}$	$0.7973_{\pm 0.0230}$	$\mathbf{0.8733}_{\pm 0.0046}$

Table 3: Comparison results in terms of NMI on all datasets. The best results are highlighted in bold.

Method	Yale	MSRCv1	ORL	BBCSport
SC1	$0.5885_{\pm 0.0281}$	$0.3229_{\pm 0.0222}$	$0.8053_{\pm 0.0109}$	$0.6717_{\pm 0.0018}$
SC2	$0.5968_{\pm 0.0223}$	$0.5961_{\pm 0.0344}$	$0.8910_{\pm 0.0102}$	$0.2345_{\pm 0.0004}$
SC3	$0.6507_{\pm 0.0256}$	$0.5103_{\pm 0.0103}$	$0.8407_{\pm 0.0169}$	—
SC4	—	$0.5355_{\pm 0.0151}$	—	—
ConcatePCA-SC	$0.6076_{\pm 0.0269}$	$0.5087_{\pm 0.0092}$	$0.8069_{\pm 0.0113}$	—
Co-Reg	$0.6362_{\pm 0.0041}$	$0.5104_{\pm 0.0036}$	$0.8376_{\pm 0.0017}$	$0.5375_{\pm 0.0021}$
Co-training	$0.6561_{\pm 0.0049}$	$0.6156_{\pm 0.0064}$	$0.8813_{\pm 0.0031}$	$0.5657_{\pm 0.0017}$
Min-Disagreement	$0.6303_{\pm 0.0048}$	$0.5174_{\pm 0.0046}$	$0.8617_{\pm 0.0030}$	$\mathbf{0.7843}_{\pm 0.0055}$
RMSC	$0.5242_{\pm 0.0373}$	$0.2819_{\pm 0.0138}$	$0.7200_{\pm 0.0209}$	$0.7645_{\pm 0.0117}$
LMSC	$0.6896_{\pm 0.0155}$	$0.5776_{\pm 0.0606}$	$\mathbf{0.9066}_{\pm 0.0204}$	$0.7448_{\pm 0.1356}$
MVGL	$0.6381_{\pm 0.0000}$	$0.5775_{\pm 0.0000}$	$0.8651_{\pm 0.0000}$	$0.0880_{\pm 0.0000}$
MCLES	$\mathbf{0.7154}_{\pm 0.0199}$	$\mathbf{0.7934}_{\pm 0.0109}$	$0.9022_{\pm 0.0120}$	$0.7788_{\pm 0.0190}$

Table 4: Comparison results in terms of PUR on all datasets. The best results are highlighted in bold.

Method	Yale	MSRCv1	ORL	BBCSport
SC1	$0.5606_{\pm 0.0348}$	$0.4595_{\pm 0.0164}$	$0.6931_{\pm 0.0200}$	$0.8453_{\pm 0.0012}$
SC2	$0.5721_{\pm 0.0295}$	$0.7262_{\pm 0.0297}$	$0.8025_{\pm 0.0205}$	$0.5717_{\pm 0.0000}$
SC3	$0.6367_{\pm 0.0341}$	$0.6500_{\pm 0.0045}$	$0.7326_{\pm 0.0263}$	—
SC4	—	$0.6945_{\pm 0.0186}$	—	—
ConcatePCA-SC	$0.5764_{\pm 0.0348}$	$0.6490_{\pm 0.0069}$	$0.6932_{\pm 0.0180}$	—
Co-Reg	$0.6065_{\pm 0.0048}$	$0.6448_{\pm 0.0048}$	$0.7294_{\pm 0.0028}$	$0.7348_{\pm 0.0033}$
Co-training	$0.6287_{\pm 0.0051}$	$0.7179_{\pm 0.0071}$	$0.7879_{\pm 0.0050}$	$0.7601_{\pm 0.0020}$
Min-Disagreement	$0.6026_{\pm 0.0070}$	$0.6075_{\pm 0.0067}$	$0.7625_{\pm 0.0051}$	$0.8715_{\pm 0.0046}$
RMSC	$0.5511_{\pm 0.0355}$	$0.2826_{\pm 0.0163}$	$0.7387_{\pm 0.0169}$	$0.7597_{\pm 0.0106}$
LMSC	$0.6706_{\pm 0.0169}$	$0.6900_{\pm 0.0624}$	$0.8379_{\pm 0.0293}$	$0.8560_{\pm 0.1053}$
MVGL	$0.6424_{\pm 0.0000}$	$0.7048_{\pm 0.0000}$	$0.7950_{\pm 0.0000}$	$0.4228_{\pm 0.0000}$
MCLES	$\mathbf{0.6961}_{\pm 0.0203}$	$\mathbf{0.8814}_{\pm 0.0055}$	$\mathbf{0.8402}_{\pm 0.0181}$	$\mathbf{0.8733}_{\pm 0.0046}$



(a) Yale (b) MSRCv1 (c) ORL (d) BBCSport

Figure 2: Convergence analysis: the log of the objective value as a function of the iteration step.

6. Robust Multi-view Spectral Clustering (RMSC) (Xia et al. 2014): It is a robust multi-view spectral clustering method which uses the standard Markov chain for clustering.
7. Latent Multi-view Subspace Clustering (LMSC) (Zhang et al. 2017): It discovers a subspace representation based on the common latent structure of multiple views, and then feeds it into the spectral clustering algorithm.
8. Graph Learning for Multi-view Clustering (MVGL) (Zhan et al. 2018): It integrates a globally optimized graph based on the optimized graph of each view.

For the above eight methods, the parameters are tuned as suggested in the original papers so as to generate the best results.

For evaluation metrics, three widely used metrics, namely ACC (accuracy), NMI (normalized mutual information) and PUR (purity) are adopted to comprehensively evaluate the performance (Wang, Lai, and Yu 2016). For each metric, higher values indicate better performance. In the experiments, we run 20 times for each experiment and report the average performance and standard deviations.

Parameter Analysis

In this subsection, we conduct parameter analysis on the four parameters d, α, β and γ by varying the four parameters in the ranges [10 100], [0.2 2], [0.2 2] and [0.001 0.01] respectively. When analyzing one parameter, the other three parameters are set as the default values. The default values of the four parameters are listed in Table 1, which are also used in the comparison experiments and convergence analysis. Notice that, since the view dimension of the original multi-view datasets varies from one dataset to another, the default values of d differ significantly on different datasets. Figure 1 plots the results in terms of NMI when different parameters are used on the four datasets. It can be observed that for all the datasets our model is relatively insensitive to the four parameters over the corresponding ranges of values. In addition, there exists a wide range for each parameter in which relatively stable and good results can be obtained.

Comparison Results

The experimental results obtained by different clustering methods on the four benchmark datasets are reported in terms of ACC, NMI, and PUR in Table 2, 3, 4 respectively. As shown in the three tables, we can see that the proposed method achieves the best clustering results on most of the testing datasets. Specifically, the proposed method significantly outperforms other state-of-the-art methods on the Yale and MSRCv1 datasets. On the MSRCv1 dataset, the performance improvements over the second-best method are 18.69%, 17.18% and 15.52% respectively in terms of ACC, NMI and PUR. In addition, our method performs much better than the two single-view baseline methods, namely SC and ConcatePCA-SC, which demonstrates the effectiveness of the latent embedding learning and the cluster indicator learning in our model. Note that we can not perform ConcatePCA-SC on the BBCSport dataset, since the features on the BBCSport dataset are too sparse to run SVD. In conclusion, the proposed method obtains more robust and accurate clustering results by means of directly learning the similarity matrix and the cluster indicator matrix based on the latent embedding representation.

Convergence Analysis

To verify the convergence property of the proposed method, convergence analysis is conducted in this subsection. The optimization algorithm can be guaranteed to converge ultimately, since the objective function in Eq. (9) is non-increasing with the iterations. Figure 2 plots the log of the objective value as a function of the iteration step. From the subfigures, we find that the log of the objective value decreases rapidly during the iterations on all the four benchmark datasets. It can be obviously observed that, the convergence can be reached within the 10 steps of iterations.

Conclusion

In this paper, a novel Multi-view Clustering in Latent Embedding Space (MCLES) is proposed to jointly learn a latent embedding space, a robust similarity matrix and an accurate cluster indicator matrix in a unified optimization framework. Within this unified framework, a latent embedding representation from multiple views is discovered to better explore the multi-view data, and simultaneously the global structure and the cluster indicator matrix can be obtained. In addition, each variable can be boosted to be optimal in an interplay manner. An alternating optimization scheme is developed to solve the optimization problem. Experimental results on both image and document datasets have demonstrated the superiority of the proposed method when compared with the state-of-the-art approaches.

References

Bartels, R. H., and Stewart, G. W. 1972. Solution of the matrix equation ax+ xb= c [f4]. *Communications of the ACM* 15(9):820–826.

Blum, A., and Mitchell, T. M. 1998. Combining labeled and unlabeled data with co-training. In *COLT*, 92–100.

Brefeld, U., and Scheffer, T. 2004. Co-em support vector learning. In *ICML*, 16. ACM.

Cai, X.; Nie, F.; Huang, H.; and Kamangar, F. 2011. Heterogeneous image feature integration via multi-modal spectral clustering. In *CVPR*, 1977–1984.

Chaudhuri, K.; Kakade, S. M.; Livescu, K.; and Sridharan, K. 2009. Multi-view clustering via canonical correlation analysis. In *ICML*, 129–136. ACM.

Cortes, C.; Mohri, M.; and Rostamizadeh, A. 2009. Learning non-linear combinations of kernels. In *NIPS*, 396–404.

De Sa, V. R. 2005. Spectral clustering with two views. In *ICML*, 20–27.

Elhamifar, E., and Vidal, R. 2009. Sparse subspace clustering. In *CVPR*, 2790–2797.

Fan, K. 1949. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences* 35(11):652–655.

Ghani, R. 2002. Combining labeled and unlabeled data for multiclass text categorization. In *ICML*, volume 2, 8–12.

Gönen, M., and Alpaydın, E. 2011. Multiple kernel learning algorithms. *Journal of machine learning research* 12(Jul):2211–2268.

Gu, S.; Zhang, L.; Zuo, W.; and Feng, X. 2014. Projective dictionary pair learning for pattern classification. In *NIPS*, 793–801.

Guo, Y. 2013. Convex subspace representation learning from multi-view data. In *AAAI*.

Hu, H.; Lin, Z.; Feng, J.; and Zhou, J. 2014. Smooth representation clustering. In *CVPR*, 3834–3841.

Huang, Z.; Zhou, J. T.; Peng, X.; Zhang, C.; Zhu, H.; and Lv, J. 2019. Multi-view spectral clustering network. In *IJCAI*, 2563–2569.

Huang, L.; Chao, H.-Y.; and Wang, C.-D. 2019. Multi-view intact space clustering. *Pattern Recognition* 86:344–353.

Kang, Z.; Peng, C.; and Cheng, Q. 2017. Twin learning for similarity and clustering: A unified kernel approach. In *AAAI*.

Kumar, A., and Daumé III, H. 2011. A co-training approach for multi-view spectral clustering. In *ICML*, 393–400.

Kumar, A.; Rai, P.; and Daumé III, H. 2011. Co-regularized multi-view spectral clustering. In *NIPS*, 1413–1421.

Li, R.; Zhang, C.; Hu, Q.; Zhu, P.; and Wang, Z. 2019. Flexible multi-view representation learning for subspace clustering. In *IJCAI*, 2916–2922.

Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; and Ma, Y. 2012. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(1):171–184.

Mohar, B.; Alavi, Y.; Chartrand, G.; and Oellermann, O. 1991. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications* 2(871-898):12.

Ng, A. Y.; Jordan, M. I.; and Weiss, Y. 2002. On spectral clustering: Analysis and an algorithm. In *NIPS*, 849–856.

Nie, F.; Wang, X.; Jordan, M. I.; and Huang, H. 2016. The constrained laplacian rank algorithm for graph-based clustering. In *AAAI*, 1969–1976.

Patel, V. M.; Van Nguyen, H.; and Vidal, R. 2013. Latent space sparse subspace clustering. In *ICCV*, 225–232.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.

Tao, H.; Hou, C.; Liu, X.; Liu, T.; Yi, D.; and Zhu, J. 2018. Reliable multi-view clustering. In *AAAI*, 4123–4130.

Tzortzis, G., and Likas, A. 2012. Kernel-based weighted multi-view clustering. In *ICDM*, 675–684. IEEE.

Wang, X.; Liu, Y.; Nie, F.; and Huang, H. 2015. Discriminative unsupervised dimensionality reduction. In *IJCAI*, 3925–3931.

Wang, S.; Liu, X.; Zhu, E.; Tang, C.; Liu, J.; Hu, J.; Xia, J.; and Yin, J. 2019. Multi-view clustering via late fusion alignment maximization. In *IJCAI*, 3778–3784.

Wang, C.-D.; Lai, J.-H.; and Yu, P. S. 2016. Multi-view clustering based on belief propagation. *IEEE Trans. Knowl. Data Eng.* 28(4):1007–1021.

White, M.; Yu, Y.; Zhang, X.; and Schuurmans, D. 2012. Convex multi-view subspace learning. In *NIPS*, 1682–1690.

Winn, J. M., and Jojic, N. 2005. LOCUS: learning object classes with unsupervised segmentation. In *ICCV*, 756–763.

Xia, R.; Pan, Y.; Du, L.; and Yin, J. 2014. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *AAAI*, 2149–2155.

Xing, Y.; Yu, G.; Domeniconi, C.; Wang, J.; Zhang, Z.; and Guo, M. 2019. Multi-view multi-instance multi-label learning based on collaborative matrix factorization. In *AAAI*, 5508–5515.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *CoRR* abs/1304.5634.

Xu, Y.-M.; Wang, C.-D.; and Lai, J.-H. 2016. Weighted multi-view clustering with feature selection. *Pattern Recognition* 53:25–35.

Yao, S.; Yu, G.; Wang, J.; Domeniconi, C.; and Zhang, X. 2019. Multi-view multiple clustering. In *IJCAI*, 4121–4127.

Zhan, K.; Zhang, C.; Guan, J.; and Wang, J. 2018. Graph learning for multiview clustering. *IEEE Trans. Cybernetics* 48(10):2887–2895.

Zhang, C.; Hu, Q.; Fu, H.; Zhu, P.; and Cao, X. 2017. Latent multi-view subspace clustering. In *CVPR*, 4279–4287.

Zhang, G.-Y.; Wang, C.-D.; Huang, D.; Zheng, W.-S.; and Zhou, Y.-R. 2018. Tw-co-k-means: Two-level weighted collaborative k-means for multi-view clustering. *Knowl.-Based Syst.* 150:127–138.