Multi-View Clustering on Topological Manifold

Shudong Huang[1], **Ivor Tsang**[2], **Zenglin Xu**[3], **Jiancheng Lv**[1], **Quan-Hui Liu**[1*]

[1] College of Computer Science, Sichuan University, Chengdu 610065, China
[2] Centre for Artificial Intelligence, FEIT, University of Technology Sydney, Sydney, NSW 2007, Australia
[3] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen, Shenzhen 518055, China
{huangsd,lvjiancheng,quanhuiliu}@scu.edu.cn, ivor.tsang@uts.edu.au, xuzenglin@hit.edu.cn

Abstract

Multi-view clustering has received a lot of attentions in data mining recently. Though plenty of works have been investigated on this topic, it is still a severe challenge due to the complex nature of the multiple heterogeneous features. Particularly, existing multi-view clustering algorithms fail to consider the topological structure in the data, which is essential for clustering data on manifold. In this paper, we propose to exploit the implied data manifold by learning the topological relationship between data points. Our method coalesces multiple view-wise graphs with the topological relevance considered, and learns the weights as well as the consensus graph interactively in a unified framework. Furthermore, we manipulate the consensus graph by a connectivity constraint such that the data points from the same cluster are precisely connected into the same component. Substantial experiments on benchmark datasets are conducted to validate the effectiveness of the proposed method, compared to the state-of-the-art algorithms over the clustering performance.

Introduction

In many real scenarios, data are often generated from different sources in diverse domains or described by various feature sets (i.e., views) (Bickel and Scheffer 2004; Huang, Kang, and Xu 2020). A prime example is the documents, which can be written in different languages; other prominent examples include images as well as web pages, the former is represented by different visual descriptors, and the latter is typically classified based on their content or citation links (Liu et al. 2018; Wang, Yang, and Liu 2019). These data are referred to as multi-view data. Usually, different views capture different aspects of information, any of which suffices for mining knowledge (Li, Chen, and Wang 2019). Multi-view clustering, which partitions the data points into distinct clusters according to their compatible and complementary information encoded in heterogeneous features, has attracted widespread attention in the domain of unsupervised learning during the past two decades (Huang et al. 2021b; Nie, Cai, and Li 2017).

Numerous multi-view clustering methods have been investigated up to now, among which the graph-oriented multi-view clustering methods make up a large proportion due their efficiency of learning relationships and underlying common structure shared by multiple views. (Liang, Huang, and Wang 2019) designed a new alternating optimization scheme such that the consistent and inconsistent parts of each single-view graph can be explicitly detected. (Huang et al. 2021a) proposed to simultaneously leverage the multi-view consistency and the multi-view diversity in a joint framework. Due to the efficiency of extracting similarities between multiple views, the kernel strategy is widely utilized to boost the learning performance of multi-view clustering methods. (Tzortzis and Likas 2012) expressed each view in terms of given kernel matrix, and learned a weighted combination of the kernels in parallel to the partitioning. (Houthuys, Langone, and Suykens 2018) formulated the multi-view kernel spectral clustering as a weighted kernel canonical correlation analysis in a primal-dual optimization setting, in which a coupling term is included to enforce the clustering scores corresponding to the different views to align. (Huang et al. 2019) further performed multi-view clustering task and learned similarity relationships in kernel spaces simultaneously. Moreover, multiple kernel learning is adopted such that the clustering performance is robust to the input kernel matrix. (Chen, Xiao, and Zhou 2019) presented a novel kernelized method to handle nonlinear data structures by jointly learning the representation tensor and affinity matrix.

There are also several works that seek for a joint graph compatible across multiple views by making use of the bipartite graph fusion method. For instance, (Li et al. 2015) used the local manifold fusion to integrate heterogeneous features, and approximated the similarity graphs using bipartite graphs for multi-view spectral clustering. Furthermore, this kind of multi-view spectral clustering method is able to handle the out-of-sample problem. Inspired by the idea of anchor graph, (Nie et al. 2020) and (Kang et al. 2020) involved a small number of the data points as anchors so that the bipartite relations of anchor points and the data points can be used to cover the affinity of the entire point cloud.

Although graph-oriented multi-view clustering methods achieve promising results, there still exist several drawbacks. For one thing, the similarity predefined in data graph is large only for the neighbors. Considering that real world data are typically sampled from a nonlinear manifold, the distant data points may keep high consistency if they are linked by con-

[*] Corresponding author.

secutive neighbors. Therefore, these methods cannot fully investigate the latent topological structure of data lying on manifold. For another, the clustering performance of these methods are critically relies on initial graphs as they did not involve the graph learning as a part of the optimization procedure, which could lead to a degradation of their performance.

Regrading the aforementioned deficiencies, we propose to exploit the implied data manifold by learning the topological relationship between data points. To be more specific, instead of the utilizing Euclidean structure, a more suitable manifold topological structure is explored to calculate the intrinsic similarities. We explicitly exploit the manifold structure of data by propagating the topological connectivities between data points from near to far. We further structurize the consensus graph by a connectivity constraint so that the data points from the same cluster are precisely connected into the same component. As a result, the proposed method coalesces multiple view-wise graphs with the topological relevance considered, and learns the weights as well as the structured consensus graph interactively in a unified framework. Substantial experiments on both toy data and real datasets are conducted to validate the effectiveness of exploring manifold topological structure, and demonstrate its superior performance compared to state-of-the-art competitors.

Preliminary

Previous studies have shown that real world data are typically sampled from a nonlinear low-dimensional manifold which is embedded in the high dimensional ambient space (Roweis and Saul 2000; Zhang, Wang, and Zha 2011; Minh, Bazzani, and Murino 2016). Thus it is obviously crucial to uncover the manifold structure implied within the original data matrix.

Recently, (Wang, Chen, and Li 2017) presented a propagation-based manifold learning method to reveal the intrinsic structures of crowds and calculate collectiveness by exploring the topological relationship between individuals. It is based on a simple yet intuitive assumption that the topological connectivities between individuals could be propagated from near to far. That is, the spatial similarity between two individuals may be low, but their topological relevance to each other would be high if they are linked by consecutive neighbors. Instead of the utilizing Euclidean structure, a more suitable manifold topological structure is explored to calculate the intrinsic similarities. According to the topological structure learning theory, if two data points keep high consistency, their topological relevance to any other point is assumed to be similar.

Given a similarity graph $\mathbf{G} \in \mathbb{R}^{n\times n}$, where n is number of data points. Based on the assumption that data points with large similarity would share similar topological relevance to any other point, (Wang, Chen, and Li 2017) proposed to extract the topological relationship of data points by minimizing the following objective function

$$\min_{\mathbf{Z}} \frac{1}{2} \sum_{i,j,k=1}^{n} \mathbf{G}_{jk} \left(\mathbf{Z}_{ij} - \mathbf{Z}_{ik}\right)^2 + \alpha \left\|\mathbf{Z} - \mathbf{I}\right\|_F^2, \quad (1)$$

where i, j, and k are data points indexes, $\mathbf{Z}$ indicates the target topological relationship matrix, and $\mathbf{Z}_{ij}$ denotes the data point j's topological relevance to i. In Eq. (1), the first term is essentially a smoothness constraint that follows the above assumption, i.e., it guarantees that data points j and k share similar topological relationship with data point i if j and k are similar. The second term is actually a fitting constraint that prevents the trivial solution, where all the elements of $\mathbf{Z}$ to be identical. And the parameter α balances the two terms. Based on Eq. (1), the topological consistency is propagated through neighbors with high similarities, and the distant data points will keep close relationship if they are linked by consecutive neighbors. Finally, the optimization to the cost function defined in Eq. (1) is able to guide the search of the topological relationship matrix $\mathbf{Z}$.

Notwithstanding, the learned $\mathbf{Z}$ does not contain the explicit cluster structures, hence a subsequent postprocessing step is required to obtain the final discrete clustering results. Moreover, it is designed for the single-view setting, and thus cannot be directly applied for multi-view clustering tasks.

Our Proposed Methodology

In order to utilize the manifold topological structure for multi-view clustering, in this paper we extend the formulation introduced in Eq. (1) to the multi-view clustering domain. For multi-view data with m views, let $\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \ldots, \mathbf{G}^{(m)}$ be the corresponding input similarity graphs, and $\mathbf{G}^{(v)} \in \mathbb{R}^{n\times n} (1 \leq v \leq m)$. According to Eq. (1), we can search the topological relationship for each view by solving

$$\begin{aligned} \min_{\mathbf{Z}^{(v)}} \frac{1}{2} \sum_{v=1}^{m} \sum_{i,j,k=1}^{n} \mathbf{G}_{jk}^{(v)} \left(\mathbf{Z}_{ij}^{(v)} - \mathbf{Z}_{ik}^{(v)}\right)^2 + \alpha \left\|\mathbf{Z}^{(v)} - \mathbf{I}\right\|_F^2 \\ \text{s.t.} \quad \left(\mathbf{z}_i^{(v)}\right)^T \mathbf{1} = 1, z_{ij}^{(v)} \geq 0. \end{aligned} \quad (2)$$

Here we constrain that the sum of each row of $\mathbf{Z}^{(v)}$ is one, and all elements of $\mathbf{Z}^{(v)}$ are non-negative. It is clear that if point j is connected with many similar neighbors, it will largely affect the objective value. Hence we propose a normalized version of Eq. (2) so that each point is treated equally, which can be formulated as

$$\begin{aligned} \min_{\mathbf{Z}^{(v)}} \frac{1}{2} \sum_{v=1}^{m} \sum_{i,j,k=1}^{n} \mathbf{G}_{jk}^{(v)} \left(\frac{\mathbf{Z}_{ij}^{(v)}}{\sqrt{\mathbf{D}_{jj}^{(v)}}} - \frac{\mathbf{Z}_{ik}^{(v)}}{\sqrt{\mathbf{D}_{kk}^{(v)}}}\right)^2 + \alpha \left\|\mathbf{Z}^{(v)} - \mathbf{I}\right\|_F^2 \\ \text{s.t.} \quad \left(\mathbf{z}_i^{(v)}\right)^T \mathbf{1} = 1, z_{ij}^{(v)} \geq 0, \end{aligned} \quad (3)$$

where $\mathbf{D}^{(v)}$ is the degree matrix of $\mathbf{G}^{(v)}$.

Once the topological relationship matrices of all views are obtained, we need to adopt a set of suitable weights $\mu^{(v)} (1 \leq v \leq m)$ to reflect the importance of each view. In detail, we can approximate every $\mathbf{Z}^{(v)}$ with different confidences by learning a consensus graph $\mathbf{S}$. This thought can be modeled by minimizing the linear combination of $\left\|\mathbf{S} - \mu^{(v)}\mathbf{Z}^{(v)}\right\|_F^2$. Furthermore, as pointed out by (Mohar et al. 2001; Nie et al. 2020), we can manipulate the consensus graph to contain exactly c connected components by adding a connectivity constraint $rank\left(\mathbf{L}_S\right) = n - c$, where c is the number of clusters and $\mathbf{L}_S$ is the Laplacian matrix of $\mathbf{S}$. Thus we arrive at

$$\begin{aligned}
\min_{\mathbf{Z}^{(v)},\mathbf{S},\mu^{(v)}} & \frac{1}{2}\sum_{v=1}^{m}\sum_{i,j,k=1}^{n}\mathbf{G}_{jk}^{(v)}\left(\frac{\mathbf{Z}_{ij}^{(v)}}{\sqrt{\mathbf{D}_{jj}^{(v)}}}-\frac{\mathbf{Z}_{ik}^{(v)}}{\sqrt{\mathbf{D}_{kk}^{(v)}}}\right)^2 \\
& +\alpha\left\|\mathbf{Z}^{(v)}-\mathbf{I}\right\|_F^2+\beta\left\|\mathbf{S}-\mu^{(v)}\mathbf{Z}^{(v)}\right\|_F^2 \\
\text{s.t.} \quad & \left(\mathbf{z}_i^{(v)}\right)^T\mathbf{1}=1, z_{ij}^{(v)}\geq 0, \left(\mathbf{s}_i^{(v)}\right)^T\mathbf{1}=1, s_{ij}^{(v)}\geq 0, \\
& \mu^{(v)}\geq 0, \sum_{v=1}^{m}\mu^{(v)}=1, rank\left(\mathbf{L}_S\right)=n-c
\end{aligned} \tag{4}$$

We see that Eq. (4) is difficult to solve due to the rank constraint as it depends on $\mathbf{S}$. Let $\sigma_i\left(\mathbf{L}_S\right)$ be the i-th smallest eigenvalue of $\mathbf{L}_S$. The constraint $rank\left(\mathbf{L}_S\right) = n - c$ would be satisfied if $\sum_{i=1}^{k}\sigma_i\left(\mathbf{L}_S\right) = 0$ as $\mathbf{L}_S$ is a positive semidefinite matrix. We incorporate the constraint term $\sum_{i=1}^{k}\sigma_i\left(\mathbf{L}_S\right)$ into the cost function, thus our model can be finally formulated as

$$\begin{aligned}
\min_{\substack{\mathbf{Z}^{(v)},\mathbf{S}\\ \mathbf{F},\mu^{(v)}}} & \sum_{v=1}^{m}\underbrace{\frac{1}{2}\sum_{i,j,k=1}^{n}\mathbf{G}_{jk}^{(v)}\left(\frac{\mathbf{Z}_{ij}^{(v)}}{\sqrt{\mathbf{D}_{jj}^{(v)}}}-\frac{\mathbf{Z}_{ik}^{(v)}}{\sqrt{\mathbf{D}_{kk}^{(v)}}}\right)^2+\alpha\left\|\mathbf{Z}^{(v)}-\mathbf{I}\right\|_F^2}_{\text{topological relevance learning}} \\
& \underbrace{+\beta\left\|\mathbf{S}-\mu^{(v)}\mathbf{Z}^{(v)}\right\|_F^2}_{\text{graph fusion}}+\underbrace{2\lambda\mathrm{Tr}\left(\mathbf{F}^T\mathbf{L}_S\mathbf{F}\right)}_{\text{partition label learning}} \\
\text{s.t.} \quad & \left(\mathbf{z}_i^{(v)}\right)^T\mathbf{1}=1, z_{ij}^{(v)}\geq 0, \left(\mathbf{s}_i^{(v)}\right)^T\mathbf{1}=1, s_{ij}^{(v)}\geq 0, \\
& \mu^{(v)}\geq 0, \sum_{v=1}^{m}\mu^{(v)}=1, \mathbf{F}^T\mathbf{F}=\mathbf{I},
\end{aligned} \tag{5}$$

where $\mathbf{F} \in \mathbb{R}^{n\times c}$ denotes the cluster indicator matrix, λ is a self-tuned parameter, and the following Ky Fan's Theorem (Fan 1949) is employed

$$\begin{aligned}
\sum_{i=1}^{c}\sigma_i\left(\mathbf{L}_S\right) &= \min_{\mathbf{F}}\mathrm{Tr}\left(\mathbf{F}^T\mathbf{L}_S\mathbf{F}\right) \\
\text{s.t.} \quad & \mathbf{F}\in\mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F}=\mathbf{I}.
\end{aligned} \tag{6}$$

It is noteworthy that our model formulated in Eq. (5) is distinct from other approaches in several aspects:

- Orthogonal to other multi-view clustering methods, our model explicitly exploit the implied data manifold by learning the topological relationship between data points. Considering the topological relevance of two data points could be high if they are linked by consecutive neighbors, it is critical to search a suitable manifold topological structure so that the intrinsic similarities can be better calculated.
- It is well-known that learning with multi-stage strategy usually leads to sub-optimal performance. Therefore, we propose a joint learning framework that seamlessly integrates subtasks including topological relevance learning, graph fusion, and partition label learning together.
- We manipulate the consensus graph $\mathbf{S}$ by a connectivity constraint so that it contains exactly c connected components. Thus $\mathbf{S}$ can be considered as an indicator matrix, where the points from the same cluster are connected into the same component. In this way, the discretization procedure is no longer required in our model. Hence it is an end-to-end single-stage learning paradigm.

Optimization

In this section, we design an iterative updating algorithm to solve the optimization problem in Eq. (5). Since it is not jointly convex in all variables, we propose to optimize the objective function with respect to one variable while fix other variables. And the procedure repeats until convergence.

Update $\mathbf{F}$

With other variables fixed, we solve $\mathbf{F}$ according to

$$\min_{\mathbf{F}\in\mathbb{R}^{n\times c},\mathbf{F}^T\mathbf{F}=\mathbf{I}}\mathrm{Tr}\left(\mathbf{F}^T\mathbf{L}_S\mathbf{F}\right), \tag{7}$$

which is a classical spectral problem and the corresponding solution can be obtained by calculating the c eigenvectors of $\mathbf{L}_S$ corresponding to the c smallest eigenvalues.

Update $\mathbf{Z}^{(v)}$ for Each View

For each $\mathbf{Z}^{(v)}$, we need to solve

$$\begin{aligned}
\min_{\mathbf{Z}^{(v)}} & \frac{1}{2}\sum_{v=1}^{m}\sum_{i,j,k=1}^{n}\mathbf{G}_{jk}^{(v)}\left(\frac{\mathbf{Z}_{ij}^{(v)}}{\sqrt{\mathbf{D}_{jj}^{(v)}}}-\frac{\mathbf{Z}_{ik}^{(v)}}{\sqrt{\mathbf{D}_{kk}^{(v)}}}\right)^2 \\
& +\alpha\left\|\mathbf{Z}^{(v)}-\mathbf{I}\right\|_F^2+\beta\left\|\mathbf{S}-\mu^{(v)}\mathbf{Z}^{(v)}\right\|_F^2 \\
\text{s.t.} \quad & \left(\mathbf{z}_i^{(v)}\right)^T\mathbf{1}=1, z_{ij}^{(v)}\geq 0.
\end{aligned} \tag{8}$$

Note that Eq. (8) is independent for different v, thus for the v-th view we have

$$\begin{aligned}
\min_{\mathbf{z}_i^{(v)}} & \frac{1}{2}\sum_{i=1}^{n}\left\{\sum_{j,k=1}^{n}\mathbf{G}_{jk}^{(v)}\left(\frac{\mathbf{Z}_{ij}^{(v)}}{\sqrt{\mathbf{D}_{jj}^{(v)}}}-\frac{\mathbf{Z}_{ik}^{(v)}}{\sqrt{\mathbf{D}_{kk}^{(v)}}}\right)^2\right. \\
& \left.+\alpha\sum_{j=1}^{n}\left\|\mathbf{Z}_{ij}^{(v)}-\mathbf{I}_{ij}\right\|_F^2+\beta\sum_{j=1}^{n}\left\|\mathbf{S}_{ij}-\mu^{(v)}\mathbf{Z}_{ij}^{(v)}\right\|_F^2\right\} \\
\text{s.t.} \quad & \left(\mathbf{z}_i^{(v)}\right)^T\mathbf{1}=1, z_{ij}^{(v)}\geq 0.
\end{aligned} \tag{9}$$

For each i, Eq. (9) can be further rewritten in a vector form as

$$\begin{aligned}
\min_{\mathbf{z}_i^{(v)}} & \left(\mathbf{z}_i^{(v)}\right)^T\left(\mathbf{I}-\mathbf{D}^{-\frac{1}{2}}\mathbf{G}^{(v)}\mathbf{D}^{-\frac{1}{2}}\right)\mathbf{z}_i^{(v)} \\
& +\alpha\left\|\mathbf{z}_i^{(v)}-\mathbf{e}_i\right\|_2^2+\beta\left\|\mathbf{s}_i-\mu^{(v)}\mathbf{z}_i^{(v)}\right\|_2^2 \\
\text{s.t.} \quad & \left(\mathbf{z}_i^{(v)}\right)^T\mathbf{1}=1, z_{ij}^{(v)}\geq 0.
\end{aligned} \tag{10}$$

Algorithm 1: Algorithm to solve Eq. (13)

```
Input: a nonzero matrix $\mathbf{A}$ and a nonzero vector $\mathbf{b}$.
    Set $1 < \rho < 2$, initialize $\eta > 0$, $\mathbf{q}$.
Output: $\mathbf{Z}^{(v)}$.
repeat
    Update $\mathbf{p}$ according to (16).
    Update $\mathbf{z}_i^{(v)}$ according to (17).
    Update $\eta \leftarrow \rho\eta$.
    Update $\mathbf{q} \leftarrow \mathbf{q} + \eta\left(\mathbf{z}_i^{(v)} - \mathbf{p}\right)$.
until converge
```

Denote $\mathbf{A} = \left(1 + \alpha + \beta\left(\mu^{(v)}\right)^2\right)\mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{G}^{(v)}\mathbf{D}^{-\frac{1}{2}}$ and $\mathbf{b} = 2\alpha\mathbf{e}_i + 2\beta\mu^{(v)}\mathbf{s}_i$, Eq. (10) can be stated as

$$\min_{\left(\mathbf{z}_i^{(v)}\right)^T\mathbf{1}=1, z_{ij}^{(v)}\geq 0} \left(\mathbf{z}_i^{(v)}\right)^T\mathbf{A}\mathbf{z}_i^{(v)} - \left(\mathbf{z}_i^{(v)}\right)^T\mathbf{b}. \tag{11}$$

It is clear that Eq. (11) is a quadratic convex optimization problem, and we can solve it with the classical augmented Lagrangian multiplier (ALM) method (Bertsekas 1997). In detail, Eq. (11) can be solved by tackling its counterpart

$$\min_{\left(\mathbf{z}_i^{(v)}\right)^T\mathbf{1}=1, z_{ij}^{(v)}\geq 0, \mathbf{p}=\mathbf{z}_i^{(v)}} \left(\mathbf{z}_i^{(v)}\right)^T\mathbf{A}\mathbf{p} - \left(\mathbf{z}_i^{(v)}\right)^T\mathbf{b}. \tag{12}$$

Via ALM, the augmented Lagrangian function of Eq. (12) can be defined as

$$\begin{aligned}\min_{\left(\mathbf{z}_i^{(v)}\right)^T\mathbf{1}=1, z_{ij}^{(v)}\geq 0, \mathbf{p}} & \left(\mathbf{z}_i^{(v)}\right)^T\mathbf{A}\mathbf{p} - \left(\mathbf{z}_i^{(v)}\right)^T\mathbf{b} \\ & + \frac{\eta}{2}\left\|\mathbf{z}_i^{(v)} - \mathbf{p} + \frac{1}{\eta}\mathbf{q}\right\|_2^2,\end{aligned} \tag{13}$$

where the second term in Eq. (13) is a penalty function term which guarantees that $\mathbf{p} = \mathbf{z}_i^{(v)}$, η and $\mathbf{q}$ are the corresponding penalty coefficient and parameter, respectively.

Note that $\mathbf{p}$ and $\mathbf{z}_i^{(v)}$ can be iteratively optimized:

1) Update $\mathbf{p}$ with fixed $\mathbf{z}_i^{(v)}$. The Lagrange function of Eq. (13) w.r.t. $\mathbf{p}$ is

$$\mathcal{L}_{\mathbf{p}} = \left(\mathbf{z}_i^{(v)}\right)^T\mathbf{A}\mathbf{p} + \frac{\eta}{2}\left\|\mathbf{z}_i^{(v)} - \mathbf{p} + \frac{1}{\eta}\mathbf{q}\right\|_2^2, \tag{14}$$

Taking the derivative of $\mathcal{L}_{\mathbf{p}}$ w.r.t $\mathbf{p}$ and setting the derivative to zero, i.e.,

$$\frac{\partial\mathcal{L}_{\mathbf{p}}}{\partial\mathbf{p}} = 0, \tag{15}$$

thus we have

$$\mathbf{p} = \mathbf{z}_i^{(v)} - \frac{1}{\eta}\left(\mathbf{A}^T\mathbf{z}_i^{(v)} + \mathbf{q}\right). \tag{16}$$

2) Update $\mathbf{z}_i^{(v)}$ with fixed $\mathbf{p}$. The Lagrange function of Eq. (13) w.r.t. $\mathbf{z}_i^{(v)}$ can be written as

$$\min_{\left(\mathbf{z}_i^{(v)}\right)^T\mathbf{1}=1, z_{ij}^{(v)}\geq 0} \left\|\mathbf{z}_i^{(v)} - \mathbf{p} + \frac{1}{\eta}\mathbf{q} + \frac{\mathbf{A}\mathbf{p} - \mathbf{b}}{\eta}\right\|_2^2, \tag{17}$$

which has a closed-form solution and can be readily obtained by the optimization algorithm proposed in (Huang, Nie, and Huang 2015).

Algorithm 2: The Algorithm for Eq. (5)

```
Input: Initial graphs $\{\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \ldots, \mathbf{G}^{(m)}\}$ for the $m$
    views, cluster number $c$, parameters $\alpha$ and $\beta$.
    Initialize the weight of each view $\mu^{(v)} = \frac{1}{m}$.
    Initialize the consensus graph $\mathbf{S} = \sum_{v=1}^{m}\mu^{(v)}\mathbf{G}^{(v)}$.
Output: The indicator matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ with exactly $c$
    connected components.
repeat
    Update $\mathbf{F}$ according to Eq. (7).
    Update $\mathbf{Z}^{(v)}$ by Algorithm 1.
    Update $\mathbf{S}$ according to Eq. (21).
    Update $\mu^{(v)}$ according to Eq. (24).
until converge
```

According to the ALM principles (Bertsekas 1997), η can be exaggerated increasingly during each iteration, and $\mathbf{q}$ is updated by $\mathbf{q} \leftarrow \mathbf{q} + \eta\left(\mathbf{z}_i^{(v)} - \mathbf{p}\right)$. The detailed algorithm to solve Eq. (13) is summarized in Algorithm 1.

Update S

Drop all unrelated terms of Eq. (5) w.r.t. $\mathbf{S}$, thus we have

$$\begin{aligned}\min_{\mathbf{S}} & \sum_{v=1}^{m}\beta\left\|\mathbf{S} - \mu^{(v)}\mathbf{Z}^{(v)}\right\|_F^2 + 2\lambda\text{Tr}\left(\mathbf{F}^T\mathbf{L}_S\mathbf{F}\right) \\ \text{s.t.} & \quad \left(\mathbf{s}_i^{(v)}\right)^T\mathbf{1} = 1, s_{ij}^{(v)} \geq 0.\end{aligned} \tag{18}$$

Since Eq. (18) is independent for different i, thus we obtain

$$\begin{aligned}\min_{\mathbf{s}_i} & \sum_{v=1}^{m}\sum_{i,j=1}^{n}\left(s_{ij} - \mu^{(v)}z_{ij}^{(v)}\right)^2 + \frac{\lambda}{\beta}\left\|\mathbf{f}_i - \mathbf{f}_j\right\|_2^2 s_{ij} \\ \text{s.t.} & \quad \left(\mathbf{s}_i^{(v)}\right)^T\mathbf{1} = 1, s_{ij}^{(v)} \geq 0,\end{aligned} \tag{19}$$

Eq. (19) can be further written as

$$\begin{aligned}\min_{\mathbf{s}_i} & \sum_{i=1}^{n}\left(\sum_{v=1}^{m}\left(s_{ij} - \mu^{(v)}z_{ij}^{(v)}\right)^2 + \frac{\lambda}{\beta}\left\|\mathbf{f}_i - \mathbf{f}_j\right\|_2^2 s_{ij}\right) \\ \text{s.t.} & \quad \left(\mathbf{s}_i^{(v)}\right)^T\mathbf{1} = 1, s_{ij}^{(v)} \geq 0,\end{aligned} \tag{20}$$

For each i, we get the following compact formulation

$$\min_{\mathbf{s}_i^T\mathbf{1}=1, s_{ij}\geq 0}\left\|\mathbf{s}_i - \left(\sum_{v=1}^{m}\mu^{(v)}\mathbf{z}_i^{(v)} - \frac{\lambda}{2\beta}\mathbf{h}_i\right)\right\|_2^2, \tag{21}$$

which can be effectively solved by the optimization algorithm proposed in (Huang, Nie, and Huang 2015).

Update $\mu^{(v)}$ for Each View

Optimizing Eq. (5) w.r.t. $\mu^{(v)}$ is equivalent to solving

$$\min_{\sum_{v=1}^{m}\mu^{(v)}=1, \mu^{(v)}\geq 0}\sum_{v=1}^{m}\left\|\mathbf{S} - \mu^{(v)}\mathbf{Z}^{(v)}\right\|_F^2. \tag{22}$$

Method	ACC	NMI	Purity	F-score	Precision	Recall	ARI
SC(AllFea)	56.09±3.55	52.66±3.90	72.54±3.59	51.15±5.38	55.64±6.70	47.38±4.59	37.83±7.01
Co-train	56.45±4.77	56.55±3.61	76.15±2.03	53.87±4.44	62.43±4.87	47.38±4.10	42.31±5.48
Co-reg	55.86±4.00	55.38±1.77	74.56±2.18	54.08±4.68	58.20±5.92	50.55±3.94	41.38±6.24
DiMSC	76.15±0.56	63.47±0.82	80.30±0.56	68.59±0.35	64.28±0.84	78.29±0.60	56.01±0.57
WMSC	57.75±0.75	49.45±1.05	71.72±0.78	50.72±0.87	54.76±1.08	47.24±0.76	37.21±1.14
AWP	54.44±0.00	45.88±0.00	63.31±0.00	42.46±0.00	38.19±0.00	47.80±0.00	22.42±0.00
MCGC	56.80±0.00	34.21±0.00	65.09±0.00	51.58±0.00	41.21±0.00	68.93±0.00	31.72±0.00
mPAC	60.95±0.00	57.58±0.00	71.01±0.00	54.74±0.00	55.61±0.00	53.90±0.00	41.32±0.00
LMSC	61.48±3.44	49.72±2.31	70.89±1.33	58.52±2.65	60.87±2.88	56.46±3.79	46.58±3.24
GMC	69.23±0.00	54.80±0.00	74.56±0.00	60.47±0.00	48.44±0.00	80.45±0.00	44.31±0.00
CDG	71.78±6.04	**69.88±5.42**	81.54±3.17	67.43±5.38	66.27±2.82	66.05±8.36	57.99±6.29
Ours	**78.70±0.00**	65.93±0.00	**82.25±0.00**	**70.71±0.00**	**69.97±0.00**	**82.33±0.00**	**60.14±0.00**

Table 1: Clustering performance (mean±standard deviation) on dataset 3sources (%).

Method	ACC	NMI	Purity	F-score	Precision	Recall	ARI
SC(AllFea)	71.22±3.24	68.29±0.90	73.31±2.30	63.46±1.91	60.78±2.28	66.43±2.20	59.22±2.15
Co-train	76.97±1.27	70.30±0.59	77.65±0.37	67.09±1.02	66.20±1.04	68.05±2.16	63.40±1.09
Co-reg	66.64±4.79	62.75±1.96	67.83±3.49	56.64±2.45	55.51±3.24	57.86±1.74	51.73±2.81
DiMSC	87.47±0.27	75.64±0.12	87.99±0.17	83.00±0.13	82.34±0.11	83.73±0.16	72.81±0.13
WMSC	90.55±0.02	84.41±0.04	90.55±0.02	83.00±0.05	82.45±0.05	83.56±0.04	81.10±0.05
AWP	74.85±0.00	73.94±0.00	74.85±0.00	72.24±0.00	64.66±0.00	81.83±0.00	68.77±0.00
MCGC	82.40±0.00	83.27±0.00	84.70±0.00	79.04±0.00	72.99±0.00	86.18±0.00	76.51±0.00
mPAC	61.45±0.00	60.39±0.00	61.70±0.00	56.79±0.00	51.98±0.00	62.57±0.00	51.51±0.00
LMSC	81.41±4.59	80.71±1.75	84.98±2.77	77.44±2.75	73.97±4.11	81.36±2.22	74.81±3.11
GMC	88.20±0.00	89.32±0.00	88.20±0.00	86.53±0.00	82.60±0.00	90.85±0.00	84.96±0.00
CDG	86.00±0.00	85.88±0.00	86.00±0.00	81.94±0.00	81.59±0.00	82.28±0.00	79.93±0.00
Ours	**96.85±0.00**	**92.86±0.00**	**96.85±0.00**	**93.80±0.00**	**93.73±0.00**	**93.87±0.00**	**93.12±0.00**

Table 2: Clustering performance (mean±standard deviation) on dataset HW (%).

For each view, the Lagrange function of Eq. (22) is

$$\mathcal{L} = \sum_{v=1}^{m} \left\| \mathbf{S} - \mu^{(v)} \mathbf{Z}^{(v)} \right\|_F^2 + \gamma \left(\sum_{v=1}^{m} \mu^{(v)} - 1 \right) \tag{23}$$

where γ is the Lagrange multiplier for the v-th view.

Setting the derivative of $\mathcal{L}$ w.r.t $\mu^{(v)}$ to zero, we have

$$\mu^{(v)} = \frac{2\mathrm{Tr}\left(\mathbf{S}\mathbf{Z}^{(v)^T}\right) - \gamma}{2\mathrm{Tr}\left(\mathbf{Z}^{(v)}\mathbf{Z}^{(v)^T}\right)}. \tag{24}$$

Considering the constraint $\sum_{v=1}^{m} \mu^{(v)} = 1$, we can compute γ and further get each $\mu^{(v)}$.

The detailed algorithm to solve the objective in Eq. (5) is summarized in Algorithm 2.

Experiments

We validate the proposed method by comparing it with following state-of-the-art competitors: Co-training multi-view spectral clustering (Co-train) (Kumar and Daumé 2011), Co-regularized multi-view spectral clustering (Co-reg) (Kumar, Rai, and Daume 2012), Diversity-induced multiview subspace clustering (DiMSC) (Cao et al. 2015), Weighted multi-view spectral clustering (WMSC) (Zong et al. 2018), Multi-view clustering via adaptively weighted procrustes (AWP) (Nie, Tian, and Li 2018), Multi-view consensus graph clustering (MCGC) (Zhan et al. 2019), Multiple Partitions Aligned Clustering (mPAC) (Kang et al. 2019), Latent multi-view subspace clustering ((LMSC) (Zhang et al. 2020), Graph-based multi-view clustering (GMC) (Wang, Yang, and Liu 2020), and Multi-view clustering via cross-view graph diffusion (CGD) (Tang et al. 2020). The standard spectral clustering (SC) (Ng, Jordan, and Weiss 2002) is included as baseline. We perform SC on the concatenated features of all views (denoted by SC(AllFea)). Several benchmark multi-view data sets are used in this paper: 3source, Handwritten numerals, Caltech7 and Caltech20. Recall that

Datasets	n	m	c	d_v
3S	169	3	6	3560/3631/3068
HW	2000	6	10	216/76/64/6/240/47
Cal7	1474	6	7	48/40/254/1984/512/928
Cal20	2386	6	20	48/40/254/1984/512/928

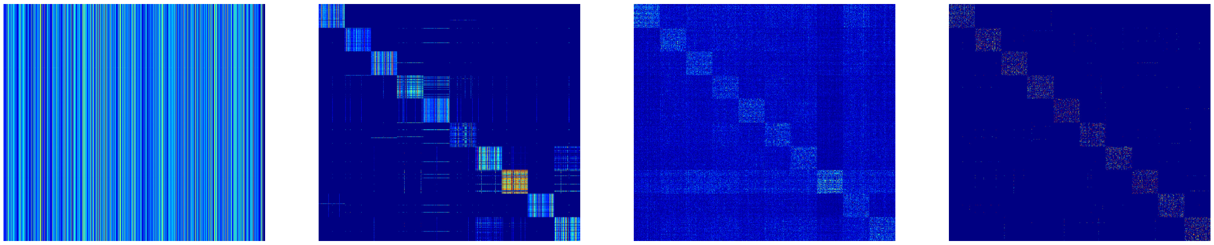
Table 3: Characteristics of the data sets.

Method	ACC	NMI	Purity	F-score	Precision	Recall	ARI
SC(AllFea)	40.55±3.00	29.32±1.00	80.01±0.28	39.34±2.14	68.15±0.29	27.70±2.19	22.00±1.45
Co-train	40.78±3.76	33.24±2.76	79.86±2.36	42.73±3.35	75.53±3.44	29.82±2.77	26.82±3.60
Co-reg	46.09±5.43	38.86±2.57	81.91±1.16	48.54±3.91	79.53±4.80	34.97±3.29	32.76±4.44
DiMSC	41.72±0.80	32.21±0.59	76.19±0.68	42.42±0.56	71.84±0.95	30.10±0.70	25.45±0.18
WMSC	38.92±0.11	28.07±0.01	79.58±0.00	37.78±0.05	67.72±0.02	26.19±0.04	20.73±0.04
AWP	58.96±0.00	46.25±0.00	83.04±0.00	61.83±0.00	87.56±0.00	47.79±0.00	47.56±0.00
MCGC	55.22±0.00	47.00±0.00	82.97±0.00	58.78±0.00	74.21±0.00	48.66±0.00	40.67±0.00
mPAC	54.41±0.00	46.24±0.00	85.14±0.00	57.51±0.00	88.68±0.00	42.55±0.00	43.35±0.00
LMSC	53.05±2.90	47.01±3.20	85.77±1.62	55.18±3.26	85.28±2.82	40.83±3.05	40.33±3.64
GMC	69.20±0.00	58.56±0.00	88.47±0.00	72.17±0.00	88.58±0.00	60.88±0.00	59.43±0.00
CDG	57.15±3.95	51.46±2.55	86.17±1.06	58.88±3.16	86.84±1.91	44.61±3.49	44.37±3.36
Ours	**70.22±0.00**	**60.66±0.00**	**88.81±0.00**	**72.52±0.00**	**89.05±0.00**	**61.16±0.00**	**59.94±0.00**

Table 4: Clustering performance (mean±standard deviation) on dataset Caltech7 (%).

Method	ACC	NMI	Purity	F-score	Precision	Recall	ARI
SC(AllFea)	29.43±1.33	33.96±0.66	60.00±0.66	23.97±0.93	47.92±1.51	15.98±0.69	17.24±0.93
Co-train	38.26±2.18	47.79±1.18	70.83±1.30	33.28±2.02	66.18±2.82	22.23±1.49	27.34±2.07
Co-reg	43.58±4.18	54.91±1.62	76.66±1.48	39.19±3.25	72.24±3.58	26.91±2.61	33.32±3.32
DiMSC	28.69±0.75	27.20±0.42	54.33±0.45	19.52±0.31	39.71±0.65	12.94±0.22	12.52±0.33
WMSC	33.26±1.42	41.50±0.58	67.08±0.52	29.83±1.61	57.61±3.01	20.13±1.12	23.38±1.73
AWP	51.55±0.00	55.90±0.00	73.39±0.00	50.43±0.00	71.62±0.00	42.61±0.00	47.00±0.00
MCGC	47.53±0.00	54.57±0.00	68.65±0.00	40.17±0.00	41.74±0.00	38.71±0.00	29.06±0.00
mPAC	55.11±0.00	56.90±0.00	75.82±0.00	51.27±0.00	**78.68±0.00**	38.02±0.00	45.49±0.00
LMSC	44.46±3.42	55.32±0.75	**76.69±0.98**	37.49±3.14	68.73±3.46	25.80±2.56	31.42±3.21
GMC	45.64±0.00	38.46±0.00	55.49±0.00	34.03±0.00	22.78±0.00	67.28±0.00	12.84±0.00
CDG	53.35±2.56	58.75±1.17	76.54±0.61	48.41±2.51	75.85±3.79	35.57±2.14	42.43±2.74
Ours	**66.81±0.00**	**57.44±0.00**	76.03±0.00	**51.62±0.00**	70.96±0.00	**69.76±0.00**	**49.14±0.00**

Table 5: Clustering performance (mean±standard deviation) on dataset Caltech20 (%).



(a) DiMSC (b) MCGC (c) LMSC (d) Ours

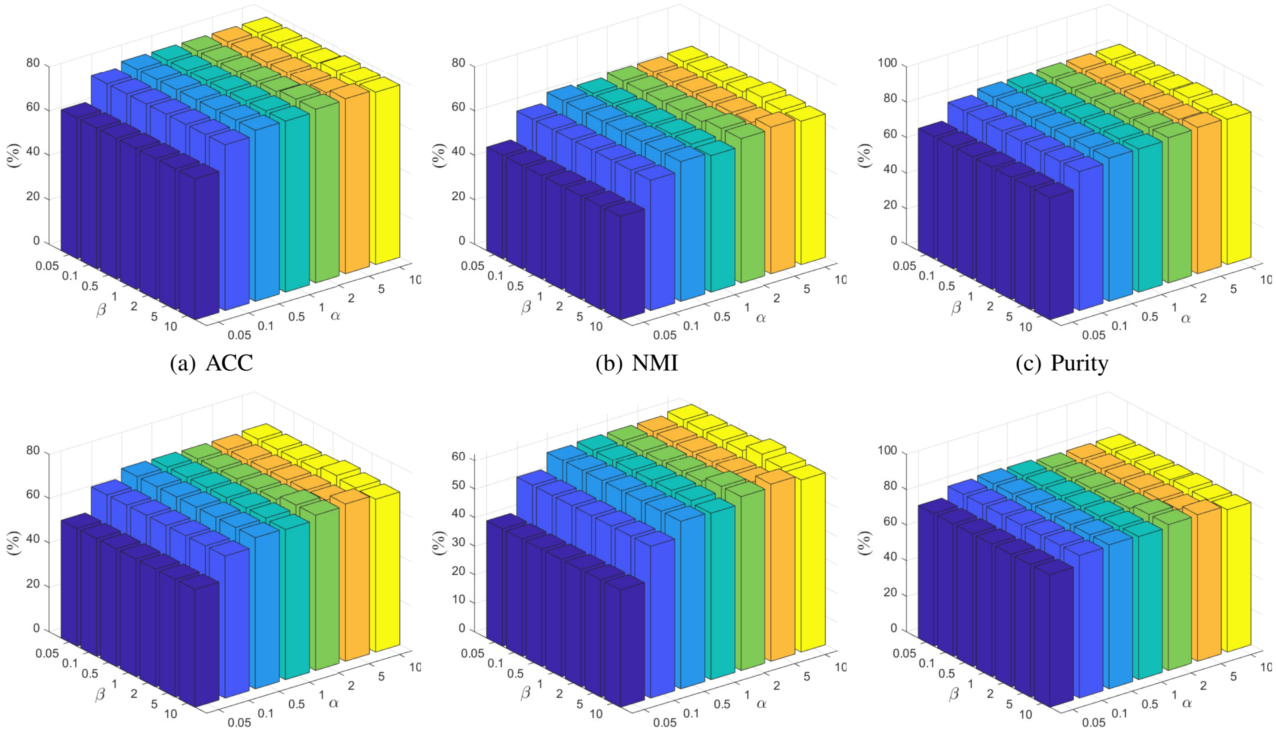
Figure 1: The consensus graph of data set HW learned by different methods. Best viewed in color.

n, m, and c denote the number of samples, views, and clusters, respectively. d_v denotes the dimensionality of the features in the v-th view. The specific characteristics of these data sets are given in Table 3.

Seven widely-used metrics are adopted to achieve a comprehensive evaluation: clustering accuracy (ACC), Normalized Mutual Information (NMI), Purity, Precision, Recall, F-score, and Adjusted Rand Index (ARI). Motivated by (Nie, Cai, and Li 2017), we initialize the initial graphs $\mathbf{G}^{(v)}$ by selecting 20-nearest neighbors among raw data.

Clustering Results

We repeat each experiment 10 times, and their mean values as well as standard deviations are reported for comparison. Note that the best clustering performance is bolded. As shown in Tables 1–5, it is clear that our approach achieves the best performance in majority cases, which verifies the effectiveness of our method. As mentioned before, $\mathbf{S}$ can be considered as an indicator matrix, where the points from the same cluster are connected into the same component. Once we obtain the consensus graph, the cluster label of each data



(a) ACC (b) NMI (c) Purity

(d) F-score (e) Precision (f) Recall

Figure 2: The clustering performance with respect to different parameter settings.

point can be directly assigned without any postprocessing. Hence our method is very stable and that's why the standard deviations of our method are always 0.00. To visualize the effect of connectivity constraint, we plot the consensus graph learned by different methods. Taking the dataset HW as an example, as shown in Figure 1, we see that compared method DiMSC cannot even find the block diagonal structure of the consensus graph. MCGC is able to search the block diagonal structure, but the number of diagonal blocks is not correct. LMSC can find the correct number of diagonal blocks, but it is seriously corrupted. It is clear that our method almost achieves a pure structured consensus graph with a much more clear clustering structure, which properly approximates the ground truth.

Parameter Analysis

This section investigates the clustering performance with respect to different parameter settings. Note that the parameter λ can be tuned in a heuristic way. That is, we initialize λ to a random positive value (e.g., $\lambda = 1$), then our model is able to automatically halve or double it when the number of connected components of $\mathbf{S}$ is greater or smaller than the cluster number c during each iteration. Thus we only need to search the parameters α and β. For simplicity, we search both α and β in the range [0.05, 0.1, 0.5, 1, 2, 5, 10]. Taking the dataset 3sources as an example, in Figure 2 we see that the clustering performance of our method is relatively stable under different parameter settings, which demonstrates the robustness of our model. For simplicity, we can achieve decent results by setting $\alpha = \beta = 1$ in practical applications.

Conclusion

In this paper, we propose to exploit the implied data manifold by learning the topological relationship between data points. Our method coalesces multiple view-wise graphs with the topological relevance considered, and learns the weights as well as the consensus graph interactively in a unified framework. Furthermore, we manipulate the consensus graph by a connectivity constraint so that the data points from the same cluster are precisely connected into the same component. To solve the optimization problem of our model, an efficient iterative updating algorithm is proposed. Substantial experiments on real datasets are conducted to validate the effectiveness of the proposed method, compared to the state-of-the-art algorithms.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grants 62106164 and 62003230, Sichuan Science and Technology Program under Grant 2021ZDZX0011, Sichuan Key Project of Research and Development Program under Grant 2022YFG0188, and the 111 Project under Grant B21044.

References

Bertsekas, D. P. 1997. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334.

Bickel, S.; and Scheffer, T. 2004. Multi-view clustering. In *ICDM*, 2004, 19–26.

Cao, X.; Zhang, C.; Fu, H.; Liu, S.; and Zhang, H. 2015. Diversity-induced multi-view subspace clustering. In *CVPR*, 586–594.

Chen, Y.; Xiao, X.; and Zhou, Y. 2019. Jointly learning kernel representation tensor and affinity matrix for multi-view clustering. *IEEE Transactions on Multimedia*, 22(8): 1985–1997.

Fan, K. 1949. On a theorem of Weyl concerning eigenvalues of linear transformations I. *Proceedings of the National Academy of Sciences of the United States of America*, 35(11): 652–655.

Houthuys, L.; Langone, R.; and Suykens, J. A. 2018. Multi-view kernel spectral clustering. *Information Fusion*, 44: 46–56.

Huang, J.; Nie, F.; and Huang, H. 2015. A new simplex sparse learning model to measure data similarity for clustering. In *IJCAI*, 3569–3575.

Huang, S.; Kang, Z.; Tsang, I. W.; and Xu, Z. 2019. Auto-weighted multi-view clustering via kernelized graph learning. *Pattern Recognition*, 88: 174–184.

Huang, S.; Kang, Z.; and Xu, Z. 2020. Auto-weighted multi-view clustering via deep matrix decomposition. *Pattern Recognition*, 97: 1–11.

Huang, S.; Tsang, I.; Xu, Z.; and Lv, J. C. 2021a. Measuring Diversity in Graph Learning: A Unified Framework for Structured Multi-view Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 1–15.

Huang, S.; Tsang, I. W.; Xu, Z.; Lv, J.; and Liu, Q. 2021b. CDD: Multi-view Subspace Clustering via Cross-view Diversity Detection. In *ACM MM*, 2308–2316.

Kang, Z.; Guo, Z.; Huang, S.; Wang, S.; Chen, W.; Su, Y.; and Xu, Z. 2019. Multiple Partitions Aligned Clustering. In *IJCAI*, 2701–2707.

Kang, Z.; Zhou, W.; Zhao, Z.; Shao, J.; Han, M.; and Xu, Z. 2020. Large-scale multi-view subspace clustering in linear time. In *AAAI*, 4412–4419.

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

Kumar, A.; Rai, P.; and Daume, H. 2012. Co-regularized multi-view spectral clustering. In *NIPS*, 1413–1421.

Li, X.; Chen, M.; and Wang, Q. 2019. Adaptive consistency propagation method for graph clustering. *IEEE Transactions on Knowledge and Data Engineering*, 32(4): 797–802.

Li, Y.; Nie, F.; Huang, H.; and Huang, J. 2015. Large-scale multi-view spectral clustering via bipartite graph. In *AAAI*, 2750–2756.

Liang, Y.; Huang, D.; and Wang, C.-D. 2019. Consistency meets inconsistency: A unified graph learning framework for multi-view clustering. In *ICDM*, 1204–1209.

Liu, X.; Zhu, X.; Li, M.; Wang, L.; Tang, C.; Yin, J.; Shen, D.; Wang, H.; and Gao, W. 2018. Late fusion incomplete multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10): 2410–2423.

Minh, H. Q.; Bazzani, L.; and Murino, V. 2016. A unifying framework in vector-valued reproducing kernel hilbert spaces for manifold regularization and co-regularized multi-view learning. *Journal of Machine Learning Research*, 17: 1–72.

Mohar, B.; Alavi, Y.; Chartrand, G.; Oellermann, O. R.; and Schwenk, A. J. 2001. The laplacian spectrum of graphs. In *Graph Theory, Combinatorics, and Applications*, 871–898.

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

Nie, F.; Cai, G.; and Li, X. 2017. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI*, 2408–2414.

Nie, F.; Tian, L.; and Li, X. 2018. Multiview clustering via adaptively weighted procrustes. In *KDD*, 2022–2030.

Nie, F.; Zhang, H.; Wang, R.; and Li, X. 2020. Multi-view clustering: A scalable and parameter-free bipartite graph fusion method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–15.

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

Tang, C.; Liu, X.; Zhu, X.; Zhu, E.; Luo, Z.; Wang, L.; and Gao, W. 2020. CGD: Multi-view clustering via cross-view graph diffusion. In *AAAI*, 5924–5931.

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

Wang, H.; Yang, Y.; and Liu, B. 2019. GMC: Graph-based multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 32(6): 1116–1129.

Wang, H.; Yang, Y.; and Liu, B. 2020. GMC: Graph-based Multi-view Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 32(6): 1116–1129.

Wang, Q.; Chen, M.; and Li, X. 2017. Quantifying and detecting collective motion by manifold learning. In *AAAI*, 4292–4298.

Zhan, K.; Nie, F.; Wang, J.; and Yang, Y. 2019. Multiview Consensus Graph Clustering. *IEEE Transactions on Image Processing*, 28(3): 1261–1270.

Zhang, C.; Fu, H.; Hu, Q.; Cao, X.; Xie, Y.; Tao, D.; and Xu, D. 2020. Generalized Latent Multi-View Subspace Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1): 86–99.

Zhang, Z.; Wang, J.; and Zha, H. 2011. Adaptive manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2): 253–265.

Zong, L.; Zhang, X.; Liu, X.; and Yu, H. 2018. Weighted multi-view spectral clustering based on spectral perturbation. In *AAAI*, 4621–4628.