Anchors Bring Ease: An Embarrassingly Simple Approach to Partial Multi-view Clustering

Jun Guo* **and Jiahui Ye***
Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen 518055, China
eeguojun@outlook.com, yejiahui079@gmail.com

Abstract

Clustering on multi-view data has attracted much more attention in the past decades. Most previous studies assume that each instance appears in all views, or there is at least one view containing all instances. However, real world data often suffers from missing some instances in each view, leading to the research problem of partial multi-view clustering. To address this issue, this paper proposes a simple yet effective **A**nchor-based **P**artial **M**ulti-view **C**lustering (APMC) method, which utilizes anchors to reconstruct instance-to-instance relationships for clustering. APMC is conceptually simple and easy to implement in practice, besides it has clear intuitions and non-trivial empirical guarantees. Specifically, APMC firstly integrates intra- and inter- view similarities through anchors. Then, spectral clustering is performed on the fused similarities to obtain a unified clustering result. Compared with existing partial multi-view clustering methods, APMC has three notable advantages: 1) it can capture more non-linear relations among instances with the help of kernel-based similarities; 2) it has a much lower time complexity in virtue of a non-iterative scheme; 3) it can inherently handle data with negative entries as well as be extended to more than two views. Finally, we extensively evaluate the proposed method on five benchmark datasets. Experimental results demonstrate the superiority of APMC over state-of-the-art approaches.

1 Introduction

Multi-view data has already been widely studied in the past few years (Lahat, Adali, and Jutten 2015; Bai et al. 2016; 2017). Various clustering models have been proposed to solve the problem that grouping unlabeled data from diverse domains into a unified partition (Kumar and Daume 2011; Kumar, Rai, and Daume 2011; Zhang et al. 2018a; 2018b). However, real-world data often suffers from incompleteness, which makes traditional multi-view clustering methods inevitably degenerate or even fail.

Incompleteness in multi-view data can be roughly divided into two cases. One is at feature level, in which certain features are missing from particular data points (Williams and Carin 2005; Williams et al. 2007; Dick, Haider, and Scheffer 2008). The other case is at instance level, where some instances are not available (Yuan et al. 2012; Shao, Shi, and Yu 2013). Specifically, the ratio of missing instances may be approximately 90% in industrial data (Little and Rubin 2014). Besides, each view may suffer from missing some instances (Xiang et al. 2013; Xu, Tao, and Xu 2015; Yang et al. 2018b). This situation typically refers to ***partial multi-view data***, which is common in practical applications (Cai et al. 2018; Zheng et al. 2018).

Previous complete multi-view methods such as Multi-NMF (Liu et al. 2013) cannot work well in this scenario, thus partial multi-view clustering has attracted increasing attention recently. Many efforts have been made to solve this problem. Representative works mainly resort to matrix factorization to exploit latent spaces for clustering. PVC (Li, Jiang, and Zhou 2014) is a pioneering work using nonnegative matrix factorization (NMF) and L_1-norm sparse regularizer to learn common and private latent spaces. IMG (Zhao, Liu, and Fu 2016) integrates PVC and manifold learning to adaptively capture the global structure of all instances. Meanwhile, MIC (Shao, He, and Yu 2015) extends Multi-NMF (Liu et al. 2013) via weighted NMF with $L_{2,1}$ regularization. DAIMC (Hu and Chen 2018) carries forward MIC through semi-NMF and $L_{2,1}$-norm regularized regression. However, they have several drawbacks to some extent.

- *Few non-linear relations among instances*. Almost all of these partial multi-view clustering approaches inherit the limitations of matrix factorization, *i.e.*, involving in linear operations. Therefore, they usually capture various linear correlations among instances, neglecting the potential and valuable non-linear relations.
- *Relatively high time cost*. There are seldom closed-form solutions for these multi-variable optimization problems. Hence, iterative algorithms are utilized for the optimal results. Most existing works have a quadratic or even cubic time complexity due to matrix eigen decomposition and inverse operation in updating variables. This will harmfully restrict their efficiency in large-scale datasets.
- *Limited generalization ability*. Some previous studies such as MIC cannot directly work in the situation where negative entries exist in features. Besides, most of the above methods such as PVC and IMG cannot be generalized to data with more than two views. The former is caused by the usage of NMF which is only applicable to nonnegative data. The latter is resulted from the specific design for two-view scenarios in their models.

To address these issues, this paper proposes an ***A**nchor-based **P**artial **M**ulti-view **C**lustering (**APMC**)* approach, which utilizes anchors to reconstruct instance-to-instance relationships for clustering. We summarize two characteristics in partial multi-view data as follows.

1) The common instances appearing in all views can help bridge the instances with non-overlapping partial views.

2) The instances with missing views cannot be removed since they still provide necessary information for clustering.

Then, our proposed APMC method makes full use of the two characteristics and integrates intra- and inter- view similarities through anchors. More specifically, we regard the common instances as anchors to bridge all the inter-view instances. This can solve the dilemma that instances sharing no common views cannot be directly used for computing cross-view similarities. After obtaining the fused similarities by anchors, spectral clustering is performed to obtain a unified clustering result. APMC potentially provides a simple yet effective partial multi-view clustering solution which is non-iterative. Experimental results well validate that our proposed APMC method performs remarkably and generally better than state-of-the-art approaches.

Compared with existing partial multi-view clustering approaches, our proposed APMC method is conceptually simple and easy to implement in practice, besides it has clear intuitions and non-trivial empirical guarantees. The major contributions of our paper are four-fold.

- We develop Gaussian kernel function based instance-to-anchor similarities, which helps bridge all the inter-view instances and capture more non-linear relations.
- We fuse intra- and inter- view similarities in one step. Without iterative steps and complicated matrix operations, APMC has a relatively low time complexity.
- We propose a strategy to directly extend APMC for more than two partial views. Meanwhile, our method can inherently handle data with negative entries.
- Experimental results on five benchmark datasets demonstrate the superiority of APMC over state-of-the-art partial multi-view clustering methods.

2 Related Work

Our paper is most related to *partial multi-view clustering*. Here follows a brief review of related methods.

The pioneering work PVC was proposed in (Li, Jiang, and Zhou 2014), which utilized the information of shared instances to learn a common latent representation. Meanwhile, it explored private latent spaces for unaligned instances via NMF. IMG (Zhao, Liu, and Fu 2016) extended PVC by adding a graph Laplacian term to learn the global structure over all instances across all views. GPMVC (Rai et al. 2016) also extended PVC to be a k partial-view algorithm with a view specific graph Laplacian regularization. In (Yin, Wu, and Wang 2015; 2017), unified latent representations and projection matrices were learned for incomplete multi-view data. $L_{2,1}$ regularization was used in MIC (Shao, He, and Yu 2015) together with a weighted NMF framework. After that, (Shao et al. 2016) designed an online version OMVC to deal with large-scale cases, which utilized a dynamic weight setting and a faster projected gradient descent algorithm. In (Zhao et al. 2016), a partial multi-modal sparse coding framework was proposed to exploit the similarity structure within the same modality and between different modalities. (Qian et al. 2016) developed a double constrained framework called DCNMF by incorporating the cluster similarity and manifold preserving constraints. (Gao, Peng, and Jian 2016) gave an IVC algorithm for clustering with more than two incomplete views, which was based on spectral graph theory and kernel alignment principle. Recently, (Xu et al. 2018) sought a latent space and then performed data reconstruction for partial multi-view subspace representation. (Yang et al. 2018b) leveraged the intrinsic and extrinsic information together to yield an inductive learner SLIM for semi-supervised scenarios. It can also be readily adopted to either classification or clustering tasks. In (Yang et al. 2018a), partial multi-view clustering problem was mathematically formulated as sparse low-rank representation and jointly measuring inter- and inter- view relations. (Hu and Chen 2018) proposed a DAIMC algorithm based on weighted semi-NMF. It combined the advantages of PVC and MIC while being able to handle data with negative entries. Moreover, it declared to be capable of more than two views.

3 The Proposed Framework

3.1 Notation and Problem Definition

Except in some specified cases, italic letters but not in bold $(k, K, \cdots)$ represent scalars. Bold lowercase letters $(\mathbf{x}, \cdots)$ denote vectors, while bold uppercase letters $(\mathbf{X}, \cdots)$ are matrices. $\mathbf{I}$ is an identity matrix with an appropriate size and $\mathbf{1}$ is an all-one vector with a compatible length.

For the ease of discussion and without loss of generality, we first take two-view case for illustration. For partial multi-view data, we follow (Li, Jiang, and Zhou 2014) to separate original data as $\left\{\mathbf{X}^{(1,2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right\}$, where $\mathbf{X}^{(1,2)} \in \mathbb{R}^{n_c \times (d_1+d_2)}$, $\mathbf{X}^{(1)} \in \mathbb{R}^{n_1 \times d_1}$ and $\mathbf{X}^{(2)} \in \mathbb{R}^{n_2 \times d_2}$ denote the instances present in both views, only view-1 and only view-2, respectively. The feature dimensions of view-1 and view-2 instances are d_1 and d_2, respectively. The number of common instances in two views is n_c. The number of instances only in view-1 is n_1; and n_2 has a similar meaning. The total number of instances is $N = n_c + n_1 + n_2$.

Partial multi-view clustering aims to group all the above-mentioned instances into K clusters, where K is assumed to be predefined by users.

3.2 Motivation and Framework

As stated in §1, partial multi-view data has the following two main characteristics. 1) On one hand, the common instances appearing in all views can help bridge the instances with non-overlapping partial views. 2) On the other hand, the instances with missing views cannot be removed since they still provide necessary information for clustering.

In light of this analysis, we wish to make full use of the two characteristics and integrate intra- and inter- view sim-

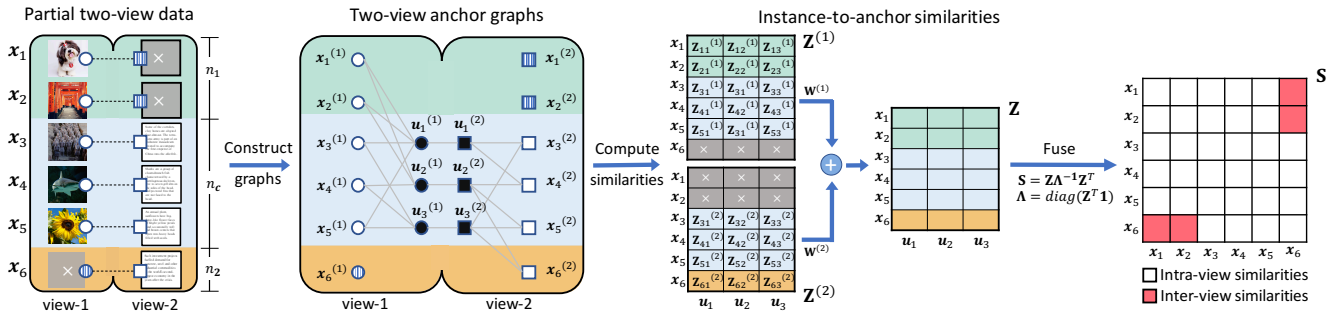


Figure 1: Anchor-based similarity reconstruction in our Anchor-based Partial Multi-view Clustering (APMC) method.

ilarities for clustering. Fortunately, anchor-based strategies (Sa et al. 2010; Liu et al. 2011) provide us an inspiring perspective. We can select the common instances as anchors to bridge all the inter-view instances (§3.3). This can solve the dilemma that instances sharing no common views cannot be directly used for computing cross-view similarities. Then, after obtaining the fused similarities by anchors, spectral clustering is performed to obtain a unified clustering result (§3.4). The detailed descriptions of our proposed ***Anchor-based Partial Multi-view Clustering (APMC)*** method are in the following two subsections.

3.3 Anchor-based Similarity Reconstruction

Anchor-based similarity reconstruction is consist of two main steps, *i.e.*, the generation of anchor sets and the construction of anchor-based similarity matrix. Figure 1 illustrates these key modules.

Generation of anchor sets. It is very challenging to directly estimate the instance-to-instance similarities in partial multi-view data, as some instances appear in one single view thus pairwise information may be unavailable. Inspired by the idea of anchor graph, we determine l pairs of anchor points by selecting the common instances that appear in both views[1], *i.e.*, $l = n_c$. As illustrated in the middle figure in Figure 1, the instances in the common area of view-1 and view-2 are selected as anchors, which can bridge the instances appearing in non-overlapping partial views.

Construction of anchor-based similarity matrix. After selecting anchor points, we build a bipartite graph to generate instance-to-anchor similarities, which is called truncated similarities in some works. Then, we construct a unified similarity matrix by fusing intra- and inter- view similarities.

- *Intra-view similarity*. Denote the set of all instances in the v-th view as $\{\mathbf{x}_i^{(v)}\}_{i=1}^{n_c+n_v}$, the anchor points set in the v-th view as $\{\mathbf{u}_i^{(v)}\}_{i=1}^{l}$. The similarity between the instance $\mathbf{x}_i^{(v)}$ and anchor point $\mathbf{u}_i^{(v)}$ is defined as

$$\mathbf{Z}_{ij}^{(v)} = \begin{cases} \frac{exp(-\mathcal{D}^2(\mathbf{x}_i^{(v)}, \mathbf{u}_j^{(v)})/\sigma^2)}{\sum_{j \in \langle i \rangle^v} exp(-\mathcal{D}^2(\mathbf{x}_i^{(v)}, \mathbf{u}_j^{(v)})/\sigma^2)} & , \forall j \in \langle i \rangle^v \\ 0 & , \text{otherwise}, \end{cases} \tag{1}$$

where $\langle i \rangle^v$ is an index set of m ($\ll l$) nearest anchors of $\mathbf{x}_i^{(v)}$ according to a distance function $\mathcal{D}(\mathbf{x}, \mathbf{u})$ such as l_2 distance. The truncated similarity is defined based on a kernel function $\mathcal{K}_\sigma(\cdot)$, which is usually a Gaussian kernel $\mathcal{K}_\sigma(\mathbf{x}, \mathbf{u}) = exp(-\mathcal{D}^2(\mathbf{x}, \mathbf{u})/\sigma^2)$. The parameter σ can be set to 1 without loss of generality. Note that the matrix $\mathbf{Z}^{(v)} \in \mathbb{R}^{(n_c+n_v)\times l}$ is highly sparse. Each row contains only m nonzero entries summing to 1.

- *Inter-view similarity*. We can derive the truncated similarity matrix $\mathbf{Z} = \left[\tilde{\mathbf{Z}}; \tilde{\mathbf{Z}}^{(1)}; \tilde{\mathbf{Z}}^{(2)}\right] \in \mathbb{R}^{N\times l}$ among all instances and anchors, where $\tilde{\mathbf{Z}}^{(v)} \in \mathbb{R}^{n_v \times l}$ $(v = 1, 2)$ is consist of the last n_v rows of $\mathbf{Z}^{(v)} \in \mathbb{R}^{(n_c+n_v)\times l}$. $\tilde{\mathbf{Z}} \in \mathbb{R}^{n_c \times l}$ indicates the similarities between the common instances and anchors, which could be computed in either view. To leverage the information from both views, we define each element of $\tilde{\mathbf{Z}}$ as $\tilde{\mathbf{Z}}_{ij} = \frac{1}{2}(\mathbf{Z}_{ij}^{(1)} + \mathbf{Z}_{ij}^{(2)})$. In Figure 1, matrices $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ help realize this inter-view fusion. $\mathbf{W}_{i\cdot}^{(1)} = \frac{1}{2}$ if the i-th instance appears in both views, $\mathbf{W}_{i\cdot}^{(1)} = 1$ if the i-th instance only appears in view-1, and $\mathbf{W}_{i\cdot}^{(1)} = 0$ if the i-th instance only appears in view-2. In a similar way, $\mathbf{W}^{(2)}$ can be computed.

In (Liu, He, and Chang 2010; Liu et al. 2011), *anchor graph* is a powerful low-rank approximation of the neighborhood graph. To this end, the similarity matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$ among all instances in partial two-view data can be approximated by anchor graph in a low-rank manner as $\mathbf{S} = \mathbf{Z}\mathbf{\Lambda}^{-1}\mathbf{Z}^T$, where $\mathbf{\Lambda} = diag(\mathbf{Z}^T\mathbf{1}) \in \mathbb{R}^{l\times l}$ is a diagonal matrix. The function $diag(\cdot)$ returns a diagonal matrix with the elements of input vector on the main diagonal.

Note that $\mathbf{S}$ has the following critical properties.

1) $\mathbf{S}_{ij} \geq 0, \forall i, \forall j$ since $\mathbf{Z}$ is nonnegative. The nonnegative similarity matrix is sufficient to make the resulting graph Laplacian matrix positive semi-definite.

2) Sparse $\mathbf{Z}$ leads to a sparse and low-rank $\mathbf{S}$, which has much less spurious connections among dissimilar instances and tends to exhibit high quality (Zhu 2006).

3) $\mathbf{S}$ is a doubly stochastic matrix, *i.e.*, it has unit row and column sums. Thus, the resulting graph Laplacian matrix is $\mathbf{L} = \mathbf{I} - \mathbf{S}$ (Liu, He, and Chang 2010).

[1]We can also utilize other ways such as k-means.

3.4 Spectral Clustering on Fused Similarities

After obtaining the fused similarity matrix $\mathbf{S}$, we can conduct spectral clustering. We firstly minimize the problem (2) via eigen decomposition on $\mathbf{L}$ to derive the corresponding K smallest eigen vectors[2], then perform k-means clustering to calculate the discrete cluster indicators.

$$\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}} tr(\mathbf{F}^T\mathbf{L}\mathbf{F}), \tag{2}$$

where $tr(\cdot)$ is the matrix trace operator, $\mathbf{F} \in \mathbb{R}^{N\times K}$ and K is the number of clusters. $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is called Laplacian matrix in graph theory, and $\mathbf{D} \in \mathbb{R}^{N\times N}$ is defined as a diagonal matrix with $\mathbf{D}_{ii} = \sum_{j=1}^{N} \mathbf{S}_{ij}$.

Remark 1: The computational complexity of eigen decomposition on the Laplacian matrix $\mathbf{L} \in \mathbb{R}^{N\times N}$ is $\mathcal{O}\left(N^3\right)$, which is not suitable for large-scale data.

Fortunately, $\mathbf{S}$ is nonnegative and doubly stochastic as mentioned in §3.3, *i.e.*,

$$\begin{aligned} diag(\mathbf{S1}) &= diag(\mathbf{Z}\mathbf{\Lambda}^{-1}\mathbf{Z}^T\mathbf{1}) \\ &= diag(\mathbf{Z}\mathbf{\Lambda}^{-1}\mathbf{\Lambda}\mathbf{1}) \\ &= diag(\mathbf{Z1}) \\ &= diag(\mathbf{1}) \\ &= \mathbf{I}. \end{aligned} \tag{3}$$

Therefore, $\mathbf{S}$ is automatically normalized making the degree matrix $\mathbf{D} = diag(\mathbf{S1})$ be an identity matrix $\mathbf{I}$.

Remark 2: With the resulting Laplacian matrix $\mathbf{L} = \mathbf{I} - \mathbf{S}$, Eq.(2) is equivalent to Eq.(4).

$$\max_{\mathbf{F}^T\mathbf{F}=\mathbf{I}} tr(\mathbf{F}^T\mathbf{S}\mathbf{F}) \tag{4}$$

Note that $\mathbf{S} = \mathbf{Z}\mathbf{\Lambda}^{-1}\mathbf{Z}^T = \mathbf{Z}\mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}^{-1/2}\mathbf{Z}^T = \mathbf{A}\mathbf{A}^T$, where $\mathbf{A} = \mathbf{Z}\mathbf{\Lambda}^{-1/2}$. The Singular Value Decomposition (SVD) of $\mathbf{A}$ can be formulated as $\mathbf{A} = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^T$, where $\mathbf{\Sigma} \in \mathbb{R}^{N\times l}$, $\mathbf{P} \in \mathbb{R}^{N\times N}$, and $\mathbf{Q} \in \mathbb{R}^{l\times l}$ are the singular value matrix, left singular vector matrix, and right singular vector matrix, respectively. It is obvious that

$$\begin{aligned} \mathbf{S} &= \mathbf{A}\mathbf{A}^T \\ &= \left(\mathbf{P}\mathbf{\Sigma}\mathbf{Q}^T\right)\left(\mathbf{P}\mathbf{\Sigma}\mathbf{Q}^T\right)^T \\ &= \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^T\mathbf{Q}\mathbf{\Sigma}^T\mathbf{P}^T \\ &= \mathbf{P}\left(\mathbf{\Sigma}\mathbf{\Sigma}^T\right)\mathbf{P}^T. \end{aligned} \tag{5}$$

Recall that $\mathbf{\Sigma} \in \mathbb{R}^{N\times l}$ is the singular value matrix of $\mathbf{A}$, then $\mathbf{\Sigma}\mathbf{\Sigma}^T$ returns an N-by-N diagonal matrix storing all the eigen values of $\mathbf{S} = \mathbf{A}\mathbf{A}^T$. The column vectors of $\mathbf{P}$ are the eigen vectors of $\mathbf{S}$. This has been proved in Fast Spectral Clustering (Wang, Nie, and Yu 2017). To reduce the computational complexity, we can perform SVD on $\mathbf{A}$ to derive the desired $\mathbf{F}$ rather than eigen decomposition on $\mathbf{S}$.

Alternatively, we can skillfully follow (Liu et al. 2011) to perform eigen decomposition on a small $l \times l$ matrix $\mathbf{R} = \mathbf{\Lambda}^{-1/2}\mathbf{Z}^T\mathbf{Z}\mathbf{\Lambda}^{-1/2} = \mathbf{A}^T\mathbf{A}$, resulting in $K\,(< l)$ eigen vector-value pairs $\{(\mathbf{b}_i, \theta_i)\}_{i=1}^{K}$ where $1 > \theta_1 \geq \cdots \geq \theta_K > 0$. We denote $\mathbf{\Theta} \in \mathbb{R}^{K\times K}$ as a diagonal matrix storing the K eigen values on the main diagonal, and $\mathbf{B} = [\mathbf{b}_1, \cdots, \mathbf{b}_K] \in \mathbb{R}^{l\times K}$ as a column-orthonormal matrix containing the K eigen vectors. The desired solution of $\mathbf{F} \in \mathbb{R}^{N\times K}$ is derived as $\mathbf{F} = \mathbf{Z}\mathbf{\Lambda}^{-1/2}\mathbf{B}\mathbf{\Theta}^{-1/2}$.

[2]Some previous works ignore the smallest eigen value 0 of Laplacian matrix, since the corresponding eigen vector is $\mathbf{1}$ which is useless for the following k-means clustering.

Algorithm 1: APMC

Input: Partial multi-view data $\left\{\mathbf{X}^{(1,2)}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right\}$; the number of clusters K; the number of nearest anchors m;
Output: The cluster indicators;
1 Build anchor graphs and compute $\mathbf{Z}$;
2 Derive the fused similarity matrix by $\mathbf{S} = \mathbf{Z}\mathbf{\Lambda}^{-1}\mathbf{Z}^T$;
3 Obtain the cluster indicators by spectral clustering.

3.5 Computational Complexity Analysis

The whole procedure of APMC is summarized in Algorithm 1. We now analyze the computational cost of our proposed method, which is non-iterative with two main steps. The corresponding time costs are summarized as follows.

In the first stage of anchor-based similarity reconstruction, the time cost is $\mathcal{O}\left(Nl\sum_v d_v\right)$ to generate truncated similarity matrix $\mathbf{Z}$, where l is the total number of anchors and d_v is the feature dimension of the v-th view.

In the second stage of spectral clustering on fused similarities, the time complexity is $\mathcal{O}\left(N^3\right)$, which involves the eigen decomposition on $\mathbf{L} \in \mathbb{R}^{N\times N}$. Benefit from the properties of similarity matrix $\mathbf{S}$, the time complexity becomes $\mathcal{O}\left(min\{Nl^2, N^2l\}\right)$ by performing SVD on $\mathbf{A} \in \mathbb{R}^{N\times l}$, which can be reduced to $\mathcal{O}\left(NK^2\right)$ (Holmes, Gray, and Isbell 2007) if we only need the K largest singular values. Alternatively, the time complexity can be $\mathcal{O}\left(l^3\right)$ if we follow (Liu et al. 2011) to eigen decompose $\mathbf{R} \in \mathbb{R}^{l\times l}$ to obtain $\mathbf{F}$. Finally, we need $\mathcal{O}\left(tNK^2\right)$ to conduct k-means clustering for t iterations on $\mathbf{F} \in \mathbb{R}^{N\times K}$.

4 Extension for Multiple Views

Our proposed APMC method can not only deal with partial two-view clustering, but also be easily extended to more than two partial views. We propose a "divide-and-conquer" strategy for multiple partial views. The whole procedure is in Figure 2 taking three views as an example, which verifies that APMC in §3 can be straightforwardly extended to the scenarios of more than two views.

4.1 Divide

Figure 2 considers all possible cases in partial three-view data, which contains three types of instances, *i.e.*, missing no view, missing one view, and missing two views. Similar to the problem definition in §3.1, we still assume the partial three-view data contains N instances. n_c is the number of instances present in all three views. n_{12} denotes the number of instances shared by view-1 and view-2; n_{13} and n_{23} have

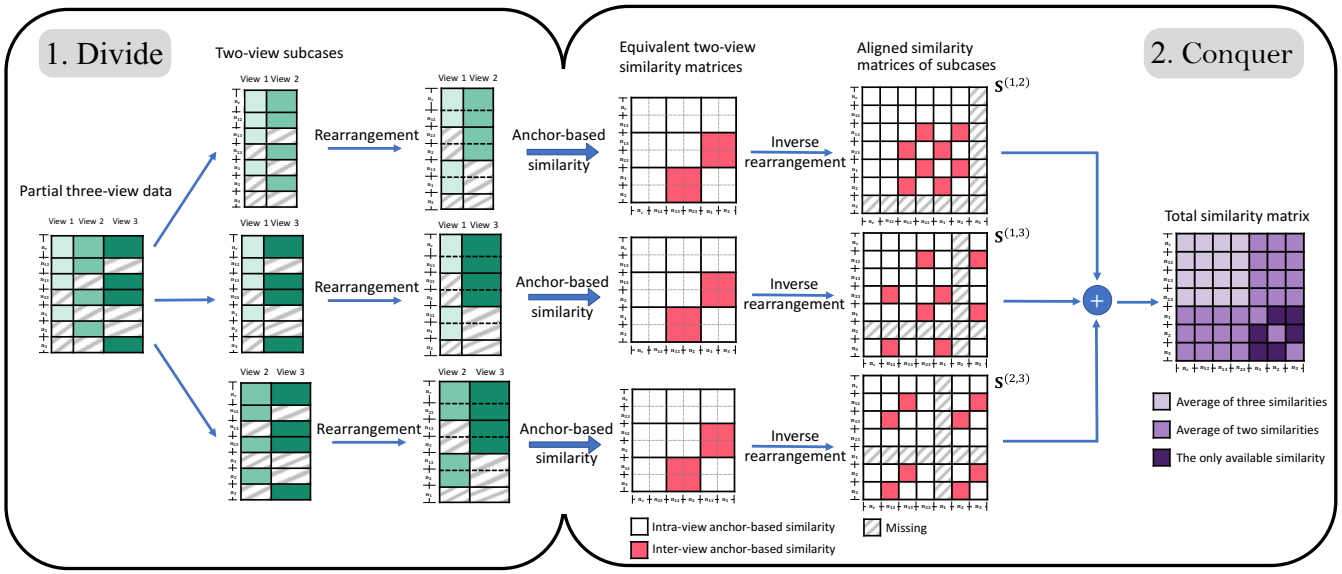


Figure 2: Anchor-based similarity reconstruction for partial three-view data.

similar meanings. $n_v (v = 1, 2, 3)$ stands for the number of instances only existing in the v-th view.

We first divide this partial three-view case into three two-view subcases. To adjust each subcase so that we can directly conduct two-view anchor-based similarity construction, we rearrange the instances according to their types. Seen from the third column in Figure 2, each subcase can be represented by an equivalent partial two-view form together with a group of common missing instances. Taking the first subcase as an example, there are $n_c + n_{12}$ instances shown in both views, $n_{23} + n_2$ only appear in view-2 while $n_{13} + n_1$ only appear in view-1. Besides, n_3 instances misses in this subcase. Similarly, we can analyze the other subcases.

4.2 Conquer

After dividing the partial three-view case, we then construct a similarity matrix for each subcase. The anchor-based similarity reconstruction method described in §3.3 can be parallelly applied here. For each partial two-view subcase, we select the common instances present in both views as anchors. Next, we compute a truncated similarity matrix and the corresponding similarity matrix for each subcase.

To further fuse the above similarity matrices in three partial two-view subcases, we rearrange them into aligned similarity matrices whose rows and columns follow the original order of instances. To alleviate the impact of missing instances in each subcase, we obtain the final similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ by a weighted combining scheme as Figure 2 shows. The light purple ones indicate that they are average of three aligned similarities, deep purple blocks are average of two aligned similarities, while dark purple stands for the only available aligned similarity.

Following §3.4, performing spectral clustering on the obtained total similarity matrix is unhindered as a final step to acquire a unified clustering result.

5 Experiments

In this section, we compare our proposed APMC approach with several state-of-the-art methods on a synthetic dataset and four real-world datasets.

5.1 Datasets

Synthetic Dataset (Guo and Zhu 2018) is composed of two views. For each view, we randomly select 200 data points from a two-component Gaussian mixture model as instances. There are two clusters (*i.e.*, cluster 1 and 2). Specifically, the cluster means are $\mu_1^{(1)} = [1, 1]$ and $\mu_2^{(1)} = [4, 2]$ in view-1, $\mu_1^{(2)} = [1, 3]$ and $\mu_2^{(2)} = [3, 1]$ in view-2. The corresponding covariances are

$$\mathbf{\Sigma}_1^{(1)} = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.4 \end{bmatrix}, \mathbf{\Sigma}_2^{(1)} = \begin{bmatrix} 0.2 & 0.15 \\ 0.15 & 0.35 \end{bmatrix};$$

$$\mathbf{\Sigma}_1^{(2)} = \begin{bmatrix} 0.25 & -0.05 \\ -0.05 & 0.2 \end{bmatrix}, \mathbf{\Sigma}_2^{(2)} = \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}.$$

Real-world datasets are described as follows.

- **USPS-MNIST Dataset** merges two famous handwritten datasets: USPS (Hull 1994) and MNIST (LeCun et al. 1998). USPS includes $9,298$ digit images with the size of 16×16 in ten classes, while MNIST contains $70,000$ digit images with the size of 28×28. The same digits in two datasets can be regarded as described in two different views. We follow (Guo and Zhu 2018) and randomly select 50 images per digit class from each dataset. Consequently, each view comprises 500 instances.
- **Oxford Flowers Dataset (Flowers)** (Nilsback and Zisserman 2006) is composed of 17 flower classes, each has 80 images described by color, shape and textures. Following (Shao, He, and Yu 2015), we adopt the χ^2 distance matrices of color and shape features as two views.

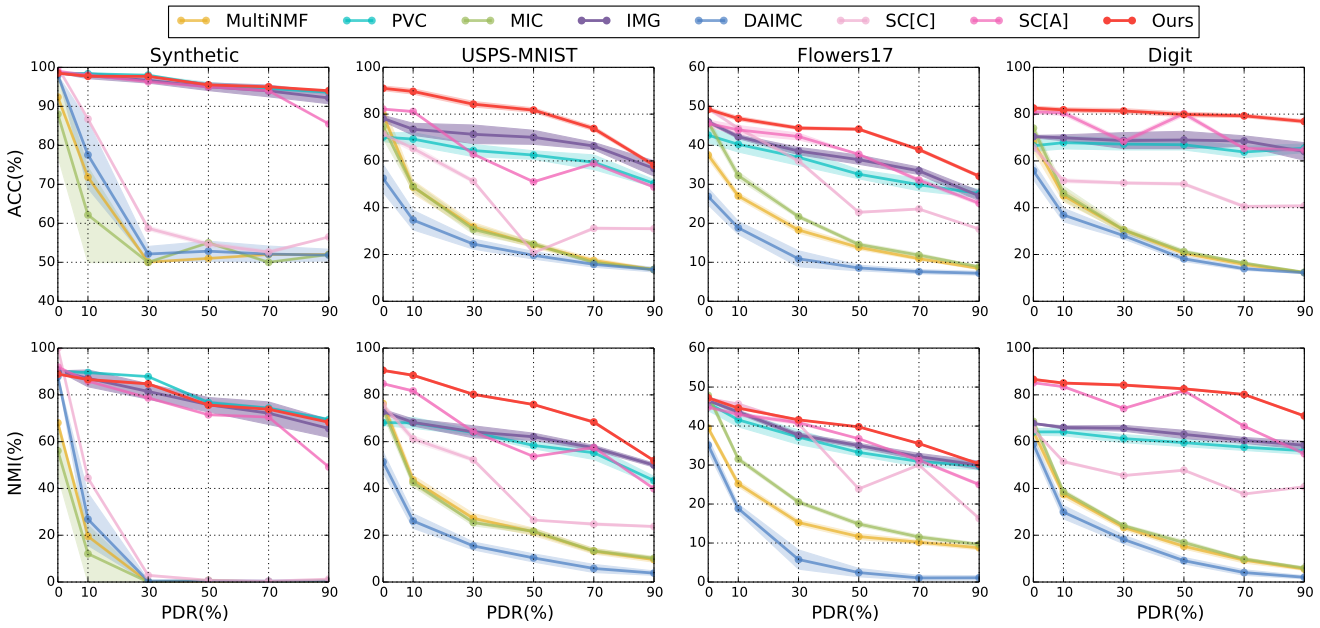


Figure 3: Clustering results of different methods on the synthetic dataset and three real-world datasets under different PDRs.

Table 1: Clustering results (ACC%±STD / NMI%±STD) of different methods on the 3Sources dataset. '-' represents the corresponding result is not available since the method cannot work on three-view scenarios.

	BBC-Guardian	BBC-Reuters	Guardian-Reuters	3Sources (All views)
MultiNMF	42.59±1.72 / 37.70±1.44	27.00±1.03 / 13.36±0.71	40.61±3.76 / 32.34±2.59	66.67±2.52 / 52.96±3.59
PVC	61.38±9.28 / 60.64±5.14	61.05±5.03 / 49.75±2.75	59.65±5.30 / 56.96±2.55	-
MIC	70.95±4.89 / 55.19±2.69	42.97±2.74 / 29.93±2.60	66.54±3.19 / 57.04±3.01	74.64±6.08 / 58.68±6.75
IMG	54.73±1.05 / 44.97±1.24	44.77±2.85 / 34.42±2.35	54.36±0.65 / 48.96±1.38	-
DAIMC	56.63±7.25 / 50.20±6.06	53.25±5.45 / 41.17±3.98	55.42±5.11 / 49.32±4.02	57.98±8.00 / 46.84±5.64
SC[C]	41.51±0.30 / 23.66±0.16	36.89±0.42 / 20.23±0.41	35.68±0.48 / 16.97±0.53	45.37±0.80 / 24.54±0.59
SC[A]	53.48±0.95 / 33.02±0.27	48.29±1.16 / 30.82±0.99	50.86±0.63 / 32.92±0.36	44.78±0.38 / 27.83±0.61
Ours	**78.28±1.57 / 63.86±0.61**	**75.27±1.42 / 65.93±0.68**	**74.37±1.49 / 66.20±0.59**	**80.31±1.47 / 68.12±0.65**

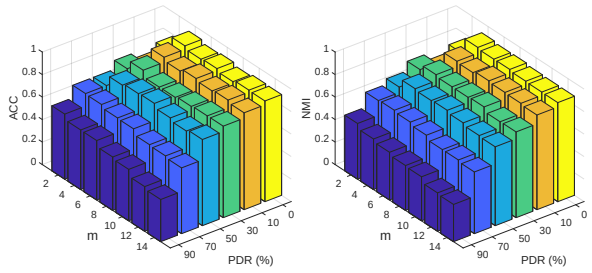
- **Multiple Features Handwritten Dataset (Digit)** (Jain, Duin, and Mao 2000) has six feature sets of ten classes of digits and each class holds 200 instances, summing up to $2,000$ instances. Following (Kumar and Daume 2011) and (Yin, Wu, and Wang 2015), we set view-1 as 76 Fourier coefficients of the character shapes, and view-2 as 216 profile correlations.
- **3Sources Dataset** (Greene and Cunningham 2009) is collected from three online news sources: BBC, Reuters, and The Guardian. In total, there are 948 news articles covering 416 distinct news stories of six topic classes from the period February to April 2009. Among these distinct stories, 169 are reported in all three sources, 194 are in two sources, and 53 appear in a single news source.

5.2 Comparison Methods

- **MultiNMF** (Liu et al. 2013) learns a common latent space based on joint NMF.
- **PVC** (Li, Jiang, and Zhou 2014) seeks a common latent space for the aligned instances and a private latent space for the unaligned instances.
- **MIC** (Shao, He, and Yu 2015) extends MultiNMF via weighted NMF with $L_{2,1}$ regularization.
- **IMG** (Zhao, Liu, and Fu 2016) integrates PVC and manifold learning to adaptively capture global structures.
- **DAIMC** (Hu and Chen 2018) applies weighted semi-NMF with the help of $L_{2,1}$-norm regularizer.

We also adopt two simple baselines that perform spectral clustering on different similarity matrices.

- **SC[C]**. After preprocessing, we concatenate each instance's features from different views into a single feature vector. Then, we obtain an instance-to-instance similarity matrix and perform spectral clustering.
- **SC[A]**. After preprocessing, we first compute an instance-to-instance similarity matrix for each view. Then, we fuse these similarity matrices by equal-weighted average and perform spectral clustering.



(a) ACC (b) NMI

Figure 4: Influence of the number of nearest anchors m on the Digit dataset with different PDR settings.

5.3 Settings

3Sources is naturally a partial multi-view dataset, while the other datasets are complete. To simplify the partial multi-view case, we follow (Shao, He, and Yu 2015) to delete the same number of instances for all views in the four complete datasets. We set Partial Data Ratio (PDR) from 10% to 90% with 20% as interval. 0% means all views are complete. The missing instances are distributed evenly in all views, and each instance is available in at least one view.

Since MultiNMF cannot directly deal with partial multi-view data, we first fill the missing instances with average feature values in each view. MultiNMF and MIC cannot work over the synthetic, USPS-MNIST and Digit datasets due to a few negative entries. Thus, we rescale the input data into the range of $[0, 1]$, then conduct normalization before we run all these clustering methods.

As for evaluation, we adopt two metrics Accuracy (ACC) and Normalized Mutual Information (NMI) to give a comprehensive analysis. Each experiment is repeated 20 times for average performance and standard deviation. All results are produced by released codes, some of which may be inconsistent with published information due to different parameter ranges and preprocessing.

5.4 Results and Analysis

Figure 3 and Table 1 report the results of ACC and NMI values on one synthetic and four real-world datasets with different PDR settings, respectively. From these data and curves, the following observations and discussion are made.

- The trends of ACC and NMI with varied PDR are similar for the following groups of methods, {MultiNMF, MIC, DAIMC}, {PVC, IMG}, and {SC[C], SC[A]}, respectively. This can be explained by the mechanisms of their models. MIC extends MultiNMF by introducing weighted NMF and $L_{2,1}$-norm. DAIMC carries forward MIC through semi-NMF and $L_{2,1}$-norm regularized regression. IMG integrates PVC by adding manifold learning. SC[C] and SC[A] conduct spectral clustering on two different poor estimations of similarities.
- As PDR increases, the performance of all algorithms drops. The improvement of APMC over SC[C], SC[A], PVC, IMG becomes larger. When the dataset is complete, all methods exhibit a relatively high performance. This is an evidence that partial multi-view clustering is more challenging than complete multi-view clustering.
- APMC dramatically outperforms two simple baselines SC[C] and SC[A], which further indicates the effectiveness of our method. It is the anchor-based similarity matrix reconstruction, rather than simple spectral clustering, that contributes to the performance improvement.
- Over 3Sources dataset, our proposed APMC method performs consistently higher than other competitors in three two-view cases and one three-view case. When the number of views increases from two to three, APMC yields even better results, which demonstrates its capability to extend to more than two views.

The superiority of APMC is analyzed as follow. The nearest anchors serve as a set of bases to represent all instances, and the instance-to-anchor similarities computed by kernel function are just like latent representations. This is intrinsically similar to conventional matrix factorization based competitors that learn latent representations for all instances. The difference is that APMC captures more non-linear relations while previous works involve in linear correlations.

As for parameter study, our proposed APMC method has only one parameter m to be fine-tuned. We set PDR from 0% to 90% as aforementioned, and explore the clustering performance of APMC by ranging m within $\{2, 4, \cdots, 14\}$. Due to the limit of space, we only report the results on Digit dataset and similar trends can be observed over other datasets. As shown in Figure 4, APMC is not obviously sensitive to m in a relatively wide range.

6 Conclusion and Future Work

In this paper, we propose a simple yet effective approach dubbed APMC for partial multi-view clustering. Specifically, APMC utilizes anchors to firstly construct instance-to-anchor similarity matrices. Then, it integrates intra- and inter- view relations into fused similarities and finally performs spectral clustering to obtain a unified clustering result. Our proposed APMC method can solve the main drawbacks of existing matrix factorization based methods for partial multi-view clustering. Experimental results well validate its superiority over state-of-the-arts.

As for future work, it will be interesting to utilize adaptive graphs (Zhao, Liu, and Fu 2016; Zhang et al. 2018c) to fuse view-specific truncated similarities. We also plan to adjust our model to better handle outliers and noise in partial multi-view data. For example, we can adopt correntropy induced metric (He, Zheng, and Hu 2011; Guo et al. 2016) and self-paced learning (Liang et al. 2016; Fan et al. 2017).

Acknowledgments

J. Guo and J. Ye are listed as co-first authors alphabetized by last name. They are funded by Tsinghua-Berkeley Shenzhen Institute (TBSI) Grant. Guo is partly supported by the 2018 Tencent Rhino-Bird Elite Training Program.

Here is author contribution statement. Guo conceived the whole idea, built a basic implementation, and ran an early simulation. Ye conducted all experiments and analyzed the results. Two authors wrote the manuscript together.

References

Bai, S.; Sun, S.; Bai, X.; Zhang, Z.; and Tian, Q. 2016. Smooth neighborhood structure mining on multiple affinity graphs with applications to context-sensitive similarity. In *ECCV*, 592–608.

Bai, S.; Zhou, Z.; Wang, J.; Bai, X.; Latecki, L. J.; and Tian, Q. 2017. Ensemble diffusion for retrieval. In *ICCV*, 774–783.

Cai, L.; Wang, Z.; Gao, H.; Shen, D.; and Ji, S. 2018. Deep adversarial learning for multi-modality missing data completion. In *SIGKDD*, 1158–1166.

Dick, U.; Haider, P.; and Scheffer, T. 2008. Learning from incomplete data with infinite imputations. In *ICML*, 232–239.

Fan, Y.; He, R.; Liang, J.; and Hu, B. 2017. Self-paced learning: an implicit regularization perspective. In *AAAI*, 1877–1883.

Gao, H.; Peng, Y.; and Jian, S. 2016. Incomplete multi-view clustering. In *ICIIP*, 245–255.

Greene, D., and Cunningham, P. 2009. A matrix factorization approach for integrating multiple data views. In *ECML/PKDD*, 423–438.

Guo, J., and Zhu, W. 2018. Partial multi-view outlier detection based on collective learning. In *AAAI*, 774–783.

Guo, J.; Guo, Y.; Kong, X.; Zhang, M.; and He, R. 2016. Discriminative analysis dictionary learning. In *AAAI*, 1617–1623.

He, R.; Zheng, W.; and Hu, B. 2011. Maximum correntropy criterion for robust face recognition. *TPAMI* 33(8):1561–1576.

Holmes, M.; Gray, A.; and Isbell, C. 2007. Fast svd for large-scale matrices. In *NIPS workshop*, volume 28, 249–252.

Hu, M., and Chen, S. 2018. Doubly aligned incomplete multi-view clustering. In *IJCAI*, 2262–2268.

Hull, J. J. 1994. A database for handwritten text recognition research. *TPAMI* 16(5):550–554.

Jain, A. K.; Duin, R. P. W.; and Mao, J. 2000. Statistical pattern recognition: a review. *TPAMI* 22(1):4–37.

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

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

Lahat, D.; Adali, T.; and Jutten, C. 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proc. IEEE* 103(9):1449–1477.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haaffner, P. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86(11):2278–2324.

Li, S.; Jiang, Y.; and Zhou, Z. 2014. Partial multi-view clustering. In *AAAI*, 1968–1974.

Liang, J.; Li, Z.; Cao, D.; He, R.; and Wang, J. 2016. Self-paced cross-modal subspace matching. In *SIGIR*, 569–578.

Little, R. J. A., and Rubin, D. B. 2014. *Statistical analysis with missing data*, volume 333. John Wiley & Sons.

Liu, W.; Wang, J.; Kumar, S.; and Chang, S. F. 2011. Hashing with graphs. In *ICML*, 1–8.

Liu, J.; Wang, C.; Gao, J.; and Han, J. 2013. Multi-view clustering via joint nonnegative matrix factorization. In *SDM*, 252–260.

Liu, W.; He, J.; and Chang, S. F. 2010. Large graph construction for scalable semi-supervised learning. In *ICML*, 679–686.

Nilsback, M. E., and Zisserman, A. 2006. A visual vocabulary for flower classification. In *CVPR*, 1447–1454.

Qian, B.; Shen, X.; Gu, Y.; Tang, Z.; and Ding, Y. 2016. Double constrained NMF for partial multi-view clustering. In *DICTA*, 1–7.

Rai, N.; Negi, S.; Chaudhury, S.; and Deshmukh, O. 2016. Partial multi-view clustering using graph regularized NMF. In *ICPR*, 2192–2197.

Sa, V. R. D.; Gallagher, P. W.; Lewis, J. M.; and Malave, V. L. 2010. Multi-view kernel construction. *Mach. Learning* 79(1):47–71.

Shao, W.; He, L.; Lu, C. T.; and Yu, P. S. 2016. Online multi-view clustering with incomplete views. In *ICBD*, 1012–1017.

Shao, W.; He, L.; and Yu, P. S. 2015. Multiple incomplete views clustering via weighted NMF with $l_{2,1}$ regularization. In *ECML/PKDD*, 318–334.

Shao, W.; Shi, X.; and Yu, P. S. 2013. Clustering on multiple incomplete datasets via collective kernel learning. In *ICDM*, 1181–1186.

Wang, R.; Nie, F.; and Yu, W. 2017. Fast spectral clustering with anchor graph for large hyperspectral images. *IEEE Geosci. Remote Sensing Lett.* 14(11):2003–2007.

Williams, D., and Carin, L. 2005. Analytical kernel matrix completion with incomplete multi-view data. In *ICML Workshop*.

Williams, D.; Liao, X.; Xue, Y.; Carin, L.; and Krishnapuram, B. 2007. On classification with incomplete data. *TPAMI* 29(3):427–436.

Xiang, S.; Yuan, L.; Fan, W.; Wang, Y.; Thompson, P. M.; and Ye, J. 2013. Multi-source learning with block-wise missing data for alzheimer's disease prediction. In *SIGKDD*, 185–193.

Xu, N.; Guo, Y.; Zheng, X.; Wang, Q.; and Luo, X. 2018. Partial multi-view subspace clustering. In *ACMMM*, 1794–1801.

Xu, C.; Tao, D.; and Xu, C. 2015. Multi-view learning with incomplete views. *TIP* 24(12):5812–5825.

Yang, W.; Shi, Y.; Gao, Y.; Wang, L.; and Yang, M. 2018a. Incomplete-data oriented multiview dimension reduction via sparse low-rank representation. *TNNLS*.

Yang, Y.; Zhan, D.; Sheng, X.; and Jiang, Y. 2018b. Semi-supervised multi-modal learning with incomplete modalities. In *IJCAI*, 2998–3004.

Yin, Q.; Wu, S.; and Wang, L. 2015. Incomplete multi-view clustering via subspace learning. In *CIKM*, 383–392.

Yin, Q.; Wu, S.; and Wang, L. 2017. Unified subspace learning for incomplete and unlabeled multi-view data. *PR* 67:313–327.

Yuan, L.; Wang, Y.; Thompson, P. M.; Narayan, V. A.; and Ye, J. 2012. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *SIGKDD*, 1149–1157.

Zhang, Z.; Liu, L.; Qin, J.; Zhu, F.; Shen, F.; Y. Xu, L. S.; and Shen, H. T. 2018a. Highly-economized multi-view binary compression for scalable image clustering. In *ECCV*, 717–732.

Zhang, Z.; Liu, L.; Shen, F.; Shen, H. T.; and Shao, L. 2018b. Binary multi-view clustering. *TPAMI*.

Zhang, Z.; Shao, L.; Xu, Y.; Liu, L.; and Yang, J. 2018c. Marginal representation learning with graph structure self-adaptation. *TNNLS* 29(10):4645–4659.

Zhao, Z.; Lu, H.; Cai, D.; He, X.; and Zhuang, Y. 2016. Partial multi-modal sparse coding via adaptive similarity structure regularization. In *ACMMM*, 152–156.

Zhao, H.; Liu, H.; and Fu, Y. 2016. Incomplete multi-modal visual data grouping. In *IJCAI*, 2392–2398.

Zheng, W.; Zhu, X.; Zhu, Y.; and Zhang, S. 2018. Robust feature selection on incomplete data. In *IJCAI*, 3191–3197.

Zhu, X. 2006. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison* 2(3):4.