
# Multi-view Clustering via Joint Feature Selection and Partially Constrained Cluster Label Learning

Qiyue Yin, Junge Zhang, Shu Wu, Hexi Li

PII: S0031-3203(19)30170-0
DOI: https://doi.org/10.1016/j.patcog.2019.04.024
Reference: PR 6889

To appear in: *Pattern Recognition*

Received date: 19 July 2018
Revised date: 4 April 2019
Accepted date: 24 April 2019

Please cite this article as: Qiyue Yin, Junge Zhang, Shu Wu, Hexi Li, Multi-view Clustering via Joint Feature Selection and Partially Constrained Cluster Label Learning, *Pattern Recognition* (2019), doi: https://doi.org/10.1016/j.patcog.2019.04.024

**Highlights**

- Cluster indicator is improvedassisted by prior knowledge for multi-view data.
- Feature selectionhelps to select discriminative views and features in a view.
- Cluster indicator is a better embedding than an orthogonal embedding.
- Running time of our method is similar with that of the mainstream approaches.

# Multi-view Clustering via Joint Feature Selection and Partially Constrained Cluster Label Learning

Qiyue Yin[a], Junge Zhang[a,*], Shu Wu[a], Hexi Li[b,*]

[a] *Institute of Automation, Chinese Academy of Sciences, HaiDian, Beijing, China*
[b] *Faculty of Intelligent Manufacturing, Wuyi university, Jiangmen, Guangdong, China*

**Abstract**

Real world data are often represented by multiple distinct feature sets, and some prior knowledge is provided, such as labels of some examples or pairwise constraints between several sample pairs. Accordingly, task of multi-view clustering arises from a complex information aggregation of multiple sources of feature sets and knowledge prior. In this paper, we propose to optimize the cluster indicator, which representing the class labels is an intuitive reflection of the clustering structure. Besides, the prior indicating the same level of semantics can be directly utilized guiding the learned clustering structure. Furthermore, feature selection is embedded into the above process to select views and features in each view, which leads to the most discriminative views and features chosen for every single cluster. To these ends, an objective is accordingly proposed with an efficient optimization strategy and convergence analysis. Extensive experiments demonstrate that our model performs better than the state-of-the-art methods.

*Keywords:* Multi-view clustering, feature selection, prior information, cluster indicator

## 1. Introduction

In real world, data are often represented by multiple distinct feature sets, which are called multi-view data. For example, an image can be represented

*Corresponding authors
*Email addresses:* qyyin@nlpr.ia.ac.cn (Qiyue Yin), jgzhang@nlpr.ia.ac.cn (Junge Zhang), shu.wu@nlpr.ia.ac.cn (Shu Wu), Jmlihexi@163.com (Hexi Li)

by SIFT and GIST descriptors, and a webpage can be described using images, texts, videos and hyperlinks. Usually, by exploring complementary information of these multiple sources of information, multi-view clustering provides a better way to discover intrinsical grouping patterns among data points than every single view [1]. On the other hand, semi-supervised clustering techniques have recently shown to substantially improve single view clustering by incorporating prior knowledge [2]. Such knowledge can be label information, e.g., partial examples have class labels, or pairwise constraint information, e.g., partial observations of whether two examples belong to the same cluster or not. More recently, a few studies have been made by integrating the above semi-supervised information for multi-view clustering and have obtained relatively good results [3, 4, 5, 6].

Generally, to solve the above problem, we confront two basic challenges. The first one is how to learn from multiple feature sets to boost clustering performance. Since different views may consist of heterogeneous feature sets and they are of different importance for clustering, a suitable strategy should well joint all the views. Typical methods usually learn a latent space, where different views can be compared to explore the complementary information [7, 8, 9][1]. Besides, feature learning is always performed to select discriminative views and features in each view [10]. Recently, some methods proposed to learn the cluster indicator matrix (the cluster index of examples) as a latent space [11, 12], which can well capture the data clustering structure. However, none of the above methods learn the clustering structure along with the views and features selection, which is a non-trivial problem considering the model design and optimization.

Another challenge is how to incorporate prior information, i.e., partial labels or partial pairwise constraints, for better discovering grouping structure. To make use of the prior, some methods utilize such information to adjust the learned new embeddings of multi-view features through soft constraints

[1] The code for [7] is available in https://github.com/singularity4/NonlinearOrthogonalNMF.

[13, 14, 3]. For example, if two examples are from the same cluster, they usually make the learned new features of the two examples be similar. Furthermore, some algorithms just keep the prior knowledge fixed with hard constraints when performing clustering [15]. Overall, those methods discover the grouping structure by indirectly utilizing the prior knowledge, and we argue that we can directly use such high level semantics to enhance learning of the grouping structure, which will be more effective.

To alleviate the above problems, a novel clustering method is proposed based on joint feature selection and partially constrained cluster label learning. The cluster indicator, which serves as an intuitive reflection of clustering labels, is optimized. The learned representation can be directly applied for clustering without relying on computationally expensive spectral clustering step. Besides, prior knowledge, as the same high-level semantics with the clustering labels, can be utilized to adjust the learned clustering structure. Furthermore, feature selection and view selection can be embedded into the above learning process in a relatively simple manner, from which discriminative views and features are selected for every single cluster. Finally, an overall objective consisting of all above parts is developed, and an efficient optimization strategy is designed with rigorous convergence analysis.

The main contributions are listed as follows: 1) We joint data clustering structure learning, feature selection (discriminative features in a view and importance of different views) and different kinds of prior knowledge learning into a unified objective for multi-view clustering. The learned low dimensional embedding can be directly utilized for clustering without relying on computationally expensive spectral clustering. 2) We propose an efficient algorithm to optimize the above objective, which is ingenious considering the mixture of complete and partial constraints due to the regularizer encoding the prior. Besides, convergence is guaranteed via theoretical analysis. 3) We conduct experiments on four public databases with five widely used metrics, achieving the state-of-the-art clustering results.

## 2. Related work

To explore complementary information between different views, plenty of promising multi-view clustering approaches have been developed [16, 17, 18, 19]. Generally, those methods can be classified into four categories based on when to use multiple sources of information. [20, 21, 22]. Subspace based ones learn a unified embedding of multiple views relying on techniques such as probabilistic approach, matrix factorization, spectral analysis and canonical correlation analysis [23, 24, 25, 10, 12, 26, 24, 1]. Co-training and Co-EM based methods just use multiple sources of information in the clustering process with typical examples such as [27, 28, 15, 29]. Late fusion based ones integrate the clustering results from each view through voting or other strategies [30, 31, 32]. And the last type of methods just learn a unified similarity matrix, which serves as an affinity matrix for spectral clustering [33, 34, 35, 36, 37, 38]. Some recently proposed methods [39, 33, 35, 40, 41][2] are extensions of typical single view subspace segmentation approaches.

For most multi-view clustering approaches, subspace-based ones, which rely on various kinds of techniques for a low dimensional embedding learning, are widely studied. Those methods are easy to be explained and have the advantage to reduce dimensionality of original data. Co-training and Co-EM frameworks are conventionally designed for semi-supervised classification, which need strong assumptions [21] , such as sufficiency, compatibility and conditional independence, for their success. If the above conditions are not satisfied, clustering performance will be harmed. Late fusion based approaches obtain clustering through a decision-level fusion, which are direct and heavily rely on single view clustering results. Methods for unified similarity matrix learning are similar with multiple kernel learning. Similarities between data need to be calculated, whose computation cost is high, and the final representation becomes more voluminous with the increasing data size. With the comparison between different

[2] Authors in [40] provide codes for readers' reference.

kinds of approaches, we base our method on subspace learning.

One the other hand, due to convenience of obtaining partial label or pairwise constraint information, semi-supervised multi-view clustering has significantly boosted the performance by embedding such prior knowledge [13, 42, 15, 3, 43, 44]. The most popular kinds of methods use prior information to guide learning of new embedding for multiple views [4, 13, 45]. For example, Liu et al. [4] proposed a sparse regression model to project the learned embedding to the known class labels. Tang [46] developed a simple constraint to enforce the embedding of must-linked pairs to be similar and cannot-linked pairs to be far away. Some methods fix the semi-supervised information when performing multi-view clustering or just propagate the prior knowledge. For example, several kernel or canonical correlation analysis based methods use partial pairwise constraints or label information to modify the kernel matrix or covariance matrix [47, 48], and some methods [44] aim to propagate the prior information in a co-EM based style.

To the best of our knowledge, none of above clustering methods can merge clustering labels learning, views and features selection, and prior knowledge learning into one objective, which is our aim in this paper. Note Wang et al. [10] propose to learn an orthogonal subspace with views and features selection. However, our learned latent subspace is different with theirs due to their lacking of nonnegative constraints, and we show the effectiveness of this part in the experimental results (Figures 2 and 3). Besides, method proposed in [10] cannot deal with semi-supervised setting, and accordingly they cannot utilize prior knowledge in the same way as ours. Finally, with the non-negative, orthogonal and partially constrained conditions, optimizing our objective and guaranteeing the convergence is much harder.

## 3. Model

Let $\mathbf{X}^l = [\mathbf{x}_1^l, \mathbf{x}_2^l, ..., \mathbf{x}_n^l] \in \Re^{d_l \times n}$ denote feature matrix of the $l$-th view for $n$ data samples, where $d_l$ is the dimensionality. Then we have $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in$

$\Re^{d\times n}$ representing the whole dataset of total $m$ views, where $d = \sum_{l=1}^{m} d_l$ is the total dimensionality. Furthermore, we are given prior information, i.e., partial labels or partial pairwise constraints. Our goal is to conduct clustering based on all above information. More specifically, we design an objective to jointly consider clustering structure learning, feature selection and prior knowledge learning, which will be elaborated in the following subsections.

### *3.1. Cluster Label Learning*

Given the dataset $\mathbf{X}$, we construct $\mathbf{F} \in \Re^{n\times c}$ to be the cluster indicator matrix with the assumption that each data point belongs to only one of the $c$ classes. Then we have $\mathbf{F}(i,j) = 1$ if $\mathbf{x}_i$ belongs to class $j$, otherwise $\mathbf{F}(i,j) = 0$. According to the definition of $\mathbf{F}$, its structure satisfies the following constraints:

$$\mathbf{F} \in \{0,1\}^{n\times c},\ \mathbf{F}\mathbf{1}_c = \mathbf{1}_n \tag{1}$$

where $\mathbf{1}_c$ and $\mathbf{1}_n$ are $c$ and $n$ dimensional vectors with their values all being one. To obtain such an $\mathbf{F}$, we use a regression like method to project the original feature matrix to the cluster indicator matrix [10]. By doing so, we can conduct feature learning in a simple manner that will be elaborated later. Then, the optimization of $\mathbf{F}$ and other variables are written as:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{F},\mathbf{b}} ||\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2 \\ &s.t.\ \mathbf{F} \in \{0,1\}^{n\times c},\ \mathbf{F}\mathbf{1}_k = \mathbf{1}_n \end{aligned} \tag{2}$$

where $\mathbf{W}$ is the projection matrix, $\mathbf{b}$ is the intercept vector, which is necessary for minimization of the regression loss as in [10, 49].

### *3.2. Encoding prior knowledge*

We show how to incorporate prior knowledge, i.e., partial labels or partial pairwise constraints, into the above optimization problem. Without loss of generality, we use $[\mathbf{y}_l, \mathbf{y}_u] = \{1, ..., c, ?\}^{n\times 1}$ to indicate known labels and unknown labels of data, where $\mathbf{y}_l = \{1, ..., c\}^{n_l\times 1}$ denotes $n_l$ labeled points. As for partial pairwise constraints, we use $\mathbf{Y} = \{1, -1, ?\}^{n\times n}$ to indicate two examples

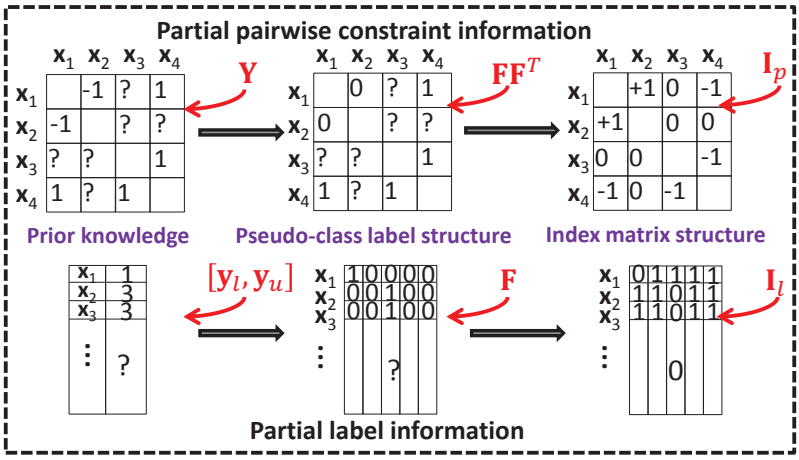


Figure 1: Process of using prior knowledge to design regularizers on the leaned pseudo-class label matrix ($\mathbf{F}$). The middle column indicates the ideal structure of $\mathbf{F}$ based on the prior, and the right column represents structure of the designed index matrix for constructing the regularizers.

belonging to the same cluster, to different clusters and unknown relation, respectively. Overall, above prior knowledge reflects group structure of partial data. For example, partial labels provide cluster index of data and partial pairwise constraints give true relation between two examples. Since our model aims to obtain the similar group structure of data, i.e., the cluster indicator matrix, we utilize prior knowledge to guide learning of $\mathbf{F}$ through a regularizer $\varphi(\mathbf{F})$. As for different prior knowledge, the regularizers are encoded in a sightly different manner, which are elaborated in the following subsections.

### *3.2.1. Partial labels*

We divide $\mathbf{F}$ into $\mathbf{F}_l$ and $\mathbf{F}_u$ corresponding to the labeled and unlabeled parts of data. Then the construction of the pseudo-class label matrix based on partial labels is illustrated in the bottom of Figure 1. Suppose $\mathbf{x}_i$ is labeled and belongs to the $k$-th cluster, then $\mathbf{F}_l(i,:)$, namely the $i$-th row of $\mathbf{F}_l$, should satisfy that the $k$-th element is large and other elements are zeros. To achieve this, we introduce an index matrix $\mathbf{I}_l$ with size of $m \times c$ corresponding to numbers of

labeled data and total clusters, and it is defined as: if $\mathbf{x}_i$ belongs to the $k$-th cluster, $\mathbf{I}_l(i,k) = 0$, otherwise $\mathbf{I}_l(i,k) = 1$. Then the regularizer is formulated as:

$$\varphi(\mathbf{F}) = ||\mathbf{I}_l \odot \mathbf{F}_l||_F^2 \tag{3}$$

where $\odot$ is element-wise product.

Using Equation 3, if $\mathbf{x}_i$ belongs to the $k$-th cluster, we make no punishment to $\mathbf{F}_l(i,k)$, and for the other positions of $\mathbf{F}_l(i,:)$, we give a penalty if the value is not zero. Noting non-negative constraint is imposed on $\mathbf{F}$, so no negative values are allowed for $\mathbf{F}$ to reduce the objective of Equation 3. Since we are given no prior for $\mathbf{F}_u$, no regularizer is utilized for this part.

#### *3.2.2. Partial pairwise constraints*

Suppose the cluster indicator matrix is known, we can easily obtain the relation between any two examples through $\mathbf{F}\mathbf{F}^T$. For example, given the true relation between $\mathbf{x}_i$ and $\mathbf{x}_j$, and if the two data samples belong to the same cluster, the $(i,j)$-th element of $\mathbf{F}\mathbf{F}^T$ should be large, otherwise zero. Then the regularizer imposed on $\mathbf{F}$ based on partial pairwise constraints is illustrated in the upper of Figure 1. We bring in an index matrix $\mathbf{I}_p \in \Re^{n\times n}$ and it is defined as: if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same cluster, $\mathbf{I}_p(i,j) = -1$, and if they belong to different clusters, $\mathbf{I}_p(i,j) = 1$, otherwise zeros. Then the regularizer is formulated as:

$$\varphi(\mathbf{F}) = \sum\nolimits_{ij}(\mathbf{I}_p \odot (\mathbf{F}\mathbf{F}^T)) \tag{4}$$

where $\odot$ is element-wise product.

Once $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same cluster, we give a negative weight to make $\mathbf{F}\mathbf{F}^T(i,j)$ be large so as to reduce the loss, and if they belong to different clusters, a penalty is given when $\mathbf{F}\mathbf{F}^T(i,j)$ does not equal to zero. As for unknown relation, no regularizer is given.

### *3.3. Feature selection*

$\mathbf{W}$ is a learned projection matrix for all the views, which is specified as $\mathbf{W} = [\mathbf{w}_1^1, ..., \mathbf{w}_c^1; ...; \mathbf{w}_1^m, ..., \mathbf{w}_c^m] \in \Re^{d\times c}$, and $\mathbf{w}_p^q$ is the weights of features

in the $q$-th view for the $p$-th cluster. Generally, different views play distinct roles for different clusters, for example, for image clustering based on color and textural features, the former one is important for clusters such as "tree" and "sky" and the latter one is more suitable for clusters such as "buildings". So we use a group $\ell_1$-norm ($G_1$-norm) imposed on $\mathbf{w}_p^q$ to select relevant views for each cluster as did in [10], and it is defined as:

$$||\mathbf{W}||_{G_1} = \sum_{p=1}^{c} \sum_{q=1}^{m} ||\mathbf{w}_p^q||_2 \tag{5}$$

Furthermore, even though some views are important for a cluster, there will still be features of these views that are noisy or redundant, and even if most features of a view are not discriminative for most clusters, a small number of features may still be highly discriminative. So a widely used $\ell_{21}$-norm constraint is also imposed on $\mathbf{W}$ as did in [50, 10] to select relevant features in each view and it is defined as:

$$||\mathbf{W}||_{21} = \sum_{i=1}^{n} ||\mathbf{w}^i||_2 \tag{6}$$

where $\mathbf{w}^i$ is the $i$th row of $\mathbf{W}$. Adding the above two constraints, we have:

$$\phi(\mathbf{W}) = ||\mathbf{W}||_{G_1} + ||\mathbf{W}||_{21} \tag{7}$$

Finally, adding all parts, we obtain the objectives

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{W},\mathbf{b}} ||\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_1||\mathbf{W}||_{G_1} + \lambda_2||\mathbf{W}||_{21} + \lambda_3||\mathbf{I}_l \odot \mathbf{F}_l||_F^2 \\ &s.t.\ \mathbf{F} \in \{0,1\}^{n\times c},\ \mathbf{F}\mathbf{1}_k = \mathbf{1}_n \end{aligned} \tag{8}$$

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{W},\mathbf{b}} ||\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_1||\mathbf{W}||_{G_1} + \lambda_2||\mathbf{W}||_{21} + \lambda_3 \textstyle\sum_{ij}(\mathbf{I}_p \odot (\mathbf{F}\mathbf{F}^T)) \\ &s.t.\ \mathbf{F} \in \{0,1\}^{n\times c},\ \mathbf{F}\mathbf{1}_k = \mathbf{1}_n \end{aligned} \tag{9}$$

for partial labels and pairwise constraints, respectively. $\lambda_1$, $\lambda_2$ and $\lambda_3$ are parameters balancing different terms.

## 4. Optimization and analysis

### *4.1. Optimization*

Since all the variables are coupled together, it may be difficult to optimize them at the same time. Hence, we propose an alternating optimization strategy

to find a local solution of Equations 8 and 9.

*4.1.1. Updating the cluster indicator matrix*

$\mathbf{F}$ is mixed with integer programming and partial constraint, which makes the problem difficult to be solved. Similar to previous studies [12], we firstly relax them to:

$$\mathbf{F}^T\mathbf{F} = \mathbf{I}, \quad \mathbf{F} \geq 0 \tag{10}$$

where the orthogonal and nonnegative constraints guarantee that there is one positive value in each row of $\mathbf{F}$ and others are zeros. Then, we optimize $\mathbf{F}$ under Equation 10 along with partial regularization encoding the prior.

- For partial pairwise constraints:

The objective becomes

$$\begin{aligned} &\min_{\mathbf{F}} ||\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_3 \textstyle\sum_{ij}(\mathbf{I}_p \odot (\mathbf{F}\mathbf{F}^T)) \\ &s.t. \quad \mathbf{F}^T\mathbf{F} = \mathbf{I},\ \mathbf{F} \geq 0 \end{aligned} \tag{11}$$

The Lagrangian function of the above function is:

$$\begin{aligned} &L(\mathbf{F}) = Tr(\mathbf{F}^T\mathbf{I}_p\mathbf{F}) + Tr(\mathbf{\Gamma}(\mathbf{F}^T\mathbf{F} - \mathbf{I})) \\ &+1/\lambda_3 Tr(-2(\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T)^T\mathbf{F} + \mathbf{F}^T\mathbf{F}) - Tr(\mathbf{\Lambda}\mathbf{F}) \end{aligned} \tag{12}$$

where $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$ are Lagrangian multipliers. Using the KKT condition, namely $\mathbf{\Lambda}(i,j)\mathbf{F}(i,j) = 0$, we have:

$$(\mathbf{I}_p\mathbf{F} + 1/\lambda_3(-\mathbf{X}^T\mathbf{W} - \mathbf{1}_n\mathbf{b}^T + \mathbf{F}) + \mathbf{F}\mathbf{\Gamma})(i,j)\mathbf{F}(i,j) = 0 \tag{13}$$

and the updating rule for $\mathbf{F}$ is:

$$\mathbf{F}(i,j) = \mathbf{F}(i,j)\sqrt{\frac{(\mathbf{I}_p^-\mathbf{F} + \gamma\mathbf{Z}^+ + \mathbf{F}\mathbf{\Gamma}^-)(i,j)}{(\mathbf{I}_p^+\mathbf{F} + \gamma(\mathbf{Z}^- + \mathbf{F}) + \mathbf{F}\mathbf{\Gamma}^+)(i,j)}} \tag{14}$$

where $\gamma = 1/\lambda_3$, $\mathbf{Z} = \mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T$ and for a matrix $\mathbf{B}$, $\mathbf{B}^+(i,j) = (|\mathbf{B}(i,j)| + \mathbf{B}(i,j))/2$ and $\mathbf{B}^-(i,j) = (|\mathbf{B}(i,j)| - \mathbf{B}(i,j))/2$.

In Equation 13, summing over $i$, we have $\mathbf{\Gamma}(i,i) = (\gamma(\mathbf{F}^T\mathbf{Z} - \mathbf{I}) - \mathbf{F}^T\mathbf{I}_p\mathbf{F})(i,i)$, and the off-diagonal elements are approximated by deleting the non-negative of $\mathbf{F}$, which results in $\mathbf{\Gamma}(i,j) = (\gamma(\mathbf{F}^T\mathbf{Z} - \mathbf{I}) - \mathbf{F}^T\mathbf{I}_p\mathbf{F})(i,j)$. In summary, $\mathbf{\Gamma} = \gamma(\mathbf{F}^T\mathbf{Z} - \mathbf{I}) - \mathbf{F}^T\mathbf{I}_p\mathbf{F}$.

- For partial labels:

The objective becomes:

$$\begin{aligned} &\min_{\mathbf{F}} ||\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_3||\mathbf{I}_l \odot \mathbf{F}_l||_F^2 \\ &s.t. \quad \mathbf{F}^T\mathbf{F} = \mathbf{I},\ \mathbf{F} \geq 0 \end{aligned} \tag{15}$$

We divide $\mathbf{F}$ into two parts, i.e., $\mathbf{F}_l$ and $\mathbf{F}_u$, corresponding to parts of the labeled and unlabeled data respectively. Then the problem is relaxed as:

$$\min_{\mathbf{F}_l} ||\mathbf{X}^T\mathbf{W} + \mathbf{1}_l\mathbf{b}_l^T - \mathbf{F}_l||_F^2 + \lambda_3||\mathbf{I}_l \odot \mathbf{F}_l||_F^2,\ s.t. \mathbf{F}_l \geq \mathbf{0} \tag{16}$$

$$\min_{\mathbf{F}_u} ||\mathbf{X}^T\mathbf{W} + \mathbf{1}_u\mathbf{b}_u^T - \mathbf{F}_u||_F^2,\ s.t.\ \mathbf{F}_u^T\mathbf{F}_u = \mathbf{I}, \mathbf{F}_u \geq \mathbf{0} \tag{17}$$

As for Equation 16, by using the penalty $\lambda_3||\mathbf{I}_l \odot \mathbf{F}_l||_2^2$ and the constraint $\mathbf{F}_l \geq \mathbf{0}$, we can approximate the condition in Equation 10 that there is only one positive value of each row in $\mathbf{F}_l$. Then Equation 16 is solved by taking derivative of $\mathbf{F}_l$:

$$\mathbf{F}_l = \max((\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T)./(\lambda_3\mathbf{I}_l + 1), \mathbf{0}) \tag{18}$$

As for Equation 17, we use $\mathbf{F}_u^T\mathbf{F}_u = \mathbf{I}, \mathbf{F}_u \geq \mathbf{0}$ to constrain the structure of $\mathbf{F}_u$. Even though it may be wrong when the unlabeled data do not contain all classes, we ignore this slight influence because the percentage of labeled data is usually small. In turn, it makes our optimization very compact. Then the solution can be obtained by ignoring the part of regularizer in Equation 14 when compared with Equation 11.

*4.1.2. Updating the projection matrix and intercept vector*

The objective for updating $\mathbf{W}$ is written as:

$$\min_{\mathbf{W}} ||\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_1||\mathbf{W}||_{G_1} + \lambda_2||\mathbf{W}||_{21} \tag{19}$$

Taking the derivative of the above objective with respect to the $k$th column of $\mathbf{W}$, i.e., $\mathbf{w}_k$, and set it to zero, we have[3]:

$$\mathbf{X}\mathbf{X}^T\mathbf{w}_k - \mathbf{X}(\mathbf{f}_k - \mathbf{1}_n b_k) + \lambda_1\mathbf{P}_k\mathbf{w}_k + \lambda_2\mathbf{Q}\mathbf{w}_k = 0 \tag{20}$$

[3]When $||\mathbf{w}_k^i||_2 = 0$ or $||\mathbf{w}^i||_2 = 0$, a small perturbation is added as in [51].

where $\mathbf{f}_k$ is the $k$-th column of $\mathbf{F}$. $\mathbf{P}_k$ is a block diagonal matrix with its $i$-th diagonal block being all $\frac{1}{2||\mathbf{w}_k^i||_2}$, which corresponds to weights of the $i$-th view to $k$-th cluster. $\mathbf{Q}$ is a diagonal matrix with its $i$-th diagonal element being $\frac{1}{2||\mathbf{w}^i||_2}$, which corresponds to the weights of the $i$-th feature in $\mathbf{X}$ to all clusters. Based on above equation, we have:

$$\mathbf{w}_k = (\mathbf{X}\mathbf{X}^T + \lambda_1\mathbf{P}_k + \lambda_2\mathbf{Q})^{-1}(\mathbf{X}(\mathbf{f}_k - \mathbf{1}_n b_k)) \tag{21}$$

In the solution, $\mathbf{P}_k$ and $\mathbf{Q}$ are dependent on $\mathbf{W}$, and an iterative solution is proposed to solve $\mathbf{P}_k$, $\mathbf{Q}$ and $\mathbf{w}_k$.

Finally $\mathbf{b}$ is obtained as:

$$\mathbf{b} = (\mathbf{F} - \mathbf{X}^T\mathbf{W})^T\mathbf{1}_n/n \tag{22}$$

The overall solution of our model is summarized in Algorithm 1. After obtaining the cluster indicator matrix, we can directly obtain clustering results by regarding the largest value of each row as cluster labels, or use the $k$means algorithm imposed on $\mathbf{F}$ for final results.

### *4.2. Convergence analysis*

Since the intercept vector $\mathbf{b}$ is calculated through analytic solution, we just elaborate the convergence for $\mathbf{W}$ and $\mathbf{F}$, respectively.

#### *4.2.1. Convergence for the cluster indicator matrix*

Since Equations 17 and 11 have similar optimization strategies, we just analyze the updating rule for Equation 11. Let

$$H(\mathbf{F}) = Tr(\mathbf{F}^T\mathbf{I}_p\mathbf{F} + \mathbf{\Gamma}(\mathbf{F}^T\mathbf{F} - \mathbf{I}) + \gamma(-2\mathbf{Z}^T\mathbf{F} + \mathbf{F}^T\mathbf{F})) \tag{23}$$

where $\gamma = 1/\lambda_3$, $\mathbf{Z} = \mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T$, and the function is further rewritten as:

$$\begin{aligned} H(\mathbf{F}) = &Tr(\mathbf{F}^T\mathbf{I}_p^+\mathbf{F} + 2\gamma\mathbf{Z}^-\mathbf{F}^T + \gamma\mathbf{F}^T\mathbf{F} + \mathbf{\Gamma}^+\mathbf{F}^T\mathbf{F}) \\ &-Tr(\mathbf{F}^T\mathbf{I}_p^-\mathbf{F} + 2\gamma\mathbf{Z}^+\mathbf{F}^T + \mathbf{\Gamma}^-\mathbf{F}^T\mathbf{F}) \end{aligned} \tag{24}$$

**Algorithm 1** Optimization strategy for Equations (8) and (9)

**Input:**

Multi-view dataset $\mathbf{X} \in \Re^{d \times n}$; Semi-supervised information; Parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$.
Initialize $\mathbf{W}$, $\mathbf{b}$ and $\mathbf{F}$ uniformly from $(-1, +1)$.
**while** not converge **do**
    Calculate diagonal matrices $\mathbf{P}_k (k = 1, ..., c)$ and $\mathbf{Q}$;
    Calculate each column of $\mathbf{W}$ using Equation (21);
    Calculate $\mathbf{b}$ using Equation (22);
    **if** Given partial labels **then**
        Calculate $\mathbf{F}$ using Equations (16) and (17);
    **else if** Given partial pairwise constraints **then**
        Calculate $\mathbf{F}$ using Equations (11);
    **end if**
**end while**

**Output:**

Approximated pseudo-class label matrix of dataset $\mathbf{X}$.

Based on the auxiliary function approach [52, 12], we can derive that the following function $h(\mathbf{F}, \tilde{\mathbf{F}})$

$$\begin{aligned} h(\mathbf{F}, \tilde{\mathbf{F}}) = & \sum_{ij} \left( \frac{(\tilde{\mathbf{F}}\mathbf{\Gamma}^{+})(i,j)\mathbf{F}^2(i,j)}{\tilde{\mathbf{F}}(i,j)} + \frac{(\mathbf{I}_p^{+}\tilde{\mathbf{F}})(i,j)\mathbf{F}^2(i,j)}{\tilde{\mathbf{F}}(i,j)} \right) \\ & + \sum_{ij} \gamma(\mathbf{Z}^{-}(i,j) \frac{\mathbf{F}^2(i,j) + \tilde{\mathbf{F}}^2(i,j)}{\tilde{\mathbf{F}}(i,j)} + \frac{\tilde{\mathbf{F}}(i,j) + \mathbf{F}^2(i,j)}{\tilde{\mathbf{F}}(i,j)}) \\ & - \sum_{ij} 2\gamma \mathbf{Z}(i,j) \tilde{\mathbf{F}}(i,j) (1 + \log \frac{\mathbf{F}(i,j)}{\tilde{\mathbf{F}}(i,j)}) \\ & - \sum_{ijl} \mathbf{\Gamma}^{-}(j,l) \tilde{\mathbf{F}}(i,j) \tilde{\mathbf{F}}(i,l) (1 + \log \frac{\mathbf{F}(i,j)\mathbf{F}(i,l)}{\tilde{\mathbf{F}}(i,j)\tilde{\mathbf{F}}(i,l)}) \\ & - \sum_{ijl} \mathbf{I}_p^{-}(j,l) \tilde{\mathbf{F}}(j,i) \tilde{\mathbf{F}}(l,i) (1 + \log \frac{\mathbf{F}(j,i)\mathbf{F}(l,i)}{\tilde{\mathbf{F}}(j,i)\tilde{\mathbf{F}}(l,i)}) \end{aligned} \tag{25}$$

is an auxiliary function of $H(\mathbf{F})$, which is convex with its minimum being Equation 14. Besides, $H(\mathbf{F})$ is the Lagrangian function of Equation 11 with the KKT condition. Then we have the following inequality chain:

$$H(\mathbf{F}^0) = h(\mathbf{F}^0, \mathbf{F}^0) \geq h(\mathbf{F}^0, \mathbf{F}^1) \geq H(\mathbf{F}^1) ... \tag{26}$$

which means the updating rule for $\mathbf{F}$ can monotonically decrease the objective function.

#### 4.2.2. Convergence for the projection matrix

Based on Equation 21, we can derive that:

$$\mathbf{W}^{(t+1)} = \min_{\mathbf{W}} ||\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2 + \lambda_1 \sum_{k=1}^{c} \mathbf{w}_k^T \mathbf{P}_k \mathbf{w}_k + \lambda_2 \mathbf{W}^T\mathbf{Q}\mathbf{W} \tag{27}$$

and then we have:

$$\begin{aligned} &L^{t+1} + \lambda_1 \textstyle\sum_{k=1}^{c} (\mathbf{w}_k^{t+1})^T \mathbf{P}_k^{t+1} \mathbf{w}_k^{t+1} + \lambda_2 (\mathbf{W}^{t+1})^T \mathbf{Q} \mathbf{W}^{t+1} \\ &\leq L^t + \lambda_1 \textstyle\sum_{k=1}^{c} (\mathbf{w}_k^{t})^T \mathbf{P}_k^{t+1} \mathbf{w}_k^{t} + \lambda_2 (\mathbf{W}^{t})^T \mathbf{Q} \mathbf{W}^{t} \end{aligned} \tag{28}$$

where $L^t = ||\mathbf{X}^T\mathbf{W}^t + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2$. By substituting matrices $\mathbf{P}$ and $\mathbf{Q}$, we have:

$$\begin{aligned} &L^{t+1} + \lambda_1 \sum_{k=1}^{c} \sum_{i=1}^{m} \frac{||(\mathbf{w}_k^i)^{t+1}||^2}{2||(\mathbf{w}_k^i)^t||} + \lambda_2 \sum_{i=1}^{d} \frac{||(\mathbf{w}^i)^{t+1}||^2}{2||(\mathbf{w}^i)^t||} \\ &\leq L^t + \lambda_1 \sum_{k=1}^{c} \sum_{i=1}^{m} \frac{||(\mathbf{w}_k^i)^{t}||^2}{2||(\mathbf{w}_k^i)^t||} + \lambda_2 \sum_{i=1}^{d} \frac{||(\mathbf{w}^i)^{t}||^2}{2||(\mathbf{w}^i)^t||} \end{aligned} \tag{29}$$

Note that it can be verified that for a function $f(x) = x - 0.5x^2/a$ $(a > 0)$, given any $x \neq a$, $f(x) \leq f(a)$ holds. Then we take $(x, a)$ pairs being $(||(\mathbf{w}_k^i)^{t+1}||_2, ||(\mathbf{w}_k^i)^t||_2)$ and $(||(\mathbf{w}^i)^{t+1}||_2, ||(\mathbf{w}^i)^t||_2)$ and add the obtained two inequations to the above inequation. Finally, we can obtain:

$$\begin{aligned} &L^{t+1} + \lambda_1 \sum_{k=1}^{c} \sum_{i=1}^{m} ||(\mathbf{w}_k^i)^{t+1}||_2 + \lambda_2 \sum_{i=1}^{d} ||(\mathbf{w}^i)^{t+1}||_2 \\ &\leq L^t + \lambda_1 \sum_{k=1}^{c} \sum_{i=1}^{m} ||(\mathbf{w}_k^i)^{t}||_2 + \lambda_2 \sum_{i=1}^{d} ||(\mathbf{w}^i)^{t}||_2 \end{aligned} \tag{30}$$

which means the updating rule for $\mathbf{W}$ can monotonically decrease the objective function.

### 4.3. Complexity analysis

In Algorithm 1, as for $\mathbf{F}$, the main computation consists of some matrix multiplication like in Equations (14) and (18). As for $\mathbf{W}$, we need to solve $\mathbf{w}_k (k = 1, ..., c)$ as described in Equation (21), which solves an inverse problem of cubic complexity [10]. Instead, we can update $\mathbf{w}_k$ by solving a linear system for about $O(cd^2)$ complexity with $c$ and $d$ being the number of clusters and the dimensionality of $\mathbf{X}$ respectively.

## 5. Experiments

### 5.1. Datasets and settings

We report results on four public datasets, i.e., USPS[4], Pascal VOC 2007[5] (VOC), MIR-Flickr[6] (MIR) and 3Source[7] datasets. Those databases have diverse views, and are widely used for multi-view clustering. Their characteristics are illustrated below.

**USPS** dataset consists of features of handwritten numerals and there are 2,000 examples uniformly distributed in 10 categories. Three types of features, i.e., the fourier coefficients, profile correlations and zernike moments, are used.

**VOC** dataset consists of a total of 9,963 images divided into twenty categories and two types of features, i.e., the tag feature (one-hot representation) and color feature are utilized.

**MIR** dataset is crawled from Flickr and contains images of 38 classes. Here a subset with five categories about 23,691 images are used with two types of features, i.e., edge histogram and homogenous texture descriptors. Finally, for datasets with multiple labels, we just remove those examples.

**3Source** dataset is constructed using three online news sources, i.e., BBC, Reuters and the Guardian. In this dataset, 169 news are reported by all the three sources, which are used with each source serving as a view. As for the representation of news, word frequency feature is utilized.

To mimic the two kinds of prior information, we randomly select partial examples with labels and also use this to construct the pairwise constraints. As for the competing approaches, we compare our method with the following typical multi-view and semi-supervised multi-view clustering algorithms.

**SULF** [42] and **PSLF** [4] are competing methods using partial labels, which add soft constrains of regression error between the samples and their class labels.

[4] http://archive.ics.uci.edu/ml/datasets/Multiple+Features
[5] http://pascallin.ecs.soton.ac.uk/challenges/VOC/
[6] http://www.cs.toronto.edu/ nitish/multimodal/index.html
[7] http://mlg.ucd.ie/datasets/3sources.html

**SCCA** [48], **PSC** and **CSC** [24] are the ones utilizing pairwise constraints. Those methods use hard constraints to guide the construction of affinity matrix, and then rely on spectral clustering for multi-view clustering.

**OurP, OurL** are our proposed methods handling partial labels and pairwise constraints, respectively. Besides, **SinP, SinL** are simple baselines that use one view for clustering without feature learning. **ConP, ConL** are another baselines that concatenate all the views for clustering without feature selection.

For methods SULF and PSLF, we use the codes released by their authors to achieve the best performance. We revise SCCA, PSC and CSC based on codes of their original vision to obtain the semi-supervised clustering results. For our method, we empirically adjust the parameters to obtain the best results, and their influence for the performance are discussed in the following section. To alleviate variance caused by the sampling of labels, all the experiments are run ten times with different sampling, and the average are reported. Finally, following [28, 4], Accuracy (ACC), F-score (F1), normalized mutual information (NMI), average entropy, and adjusted rand index are utilized for performance evaluation. For these measures, the higher values represent better performance, except for the **average entropy**.

### *5.2. Experimental results*

We test all the methods with the labeled example ratios (LER) being 10%, 20% and 30%, respectively, and the results on the USPS, VOC, MIR and 3Source datasets are shown in Tables 1, 2, 3 and 4, respectively. As a baseline, all the methods without using semi-supervised information are also reported. Overall, it can be seen that our proposed methods, whether based on partial label information or partial pairwise constraint information, achieve relatively better clustering results.

From the tables, with increasing number of labeled examples, performances of most methods are improved, which show the effectiveness of prior information for multi-view clustering. However, some methods may confront performance degradation with more labels, e.g. PSC and CSC. The main reason lies behind is

Table 1: Clstering results on the USPS datasets.

| ACC | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 73.07 | 78.88 | 83.2 | 52.46 | 57.27 | 42.62 | 42.62 | 34.50 | 34.50 | 85.18 | 85.18 |
| 0.1 | 76.18 | 76.36 | 84.45 | 62.89 | 79.51 | 57.78 | 50.11 | 83.71 | 51.57 | 93.00 | 87.03 |
| 0.2 | 74.89 | 74.15 | 80.37 | 69.76 | 81.23 | 61.21 | 46.65 | 92.85 | 61.91 | 94.50 | 87.98 |
| 0.3 | 70.47 | 69.39 | 83.59 | 66.67 | 84.03 | 61.79 | 60.44 | 94.07 | 71.87 | 94.55 | 92.18 |
| NMI | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 71.54 | 73.73 | 74.38 | 44.09 | 44.89 | 34.18 | 34.18 | 30.86 | 30.86 | 75.61 | 75.61 |
| 0.1 | 72.97 | 76.35 | 77.46 | 61.89 | 73.56 | 45.95 | 40.08 | 78.23 | 48.70 | 86.89 | 78.06 |
| 0.2 | 72.74 | 77.04 | 79.07 | 66.33 | 75.39 | 50.10 | 42.91 | 86.98 | 60.19 | 88.69 | 78.50 |
| 0.3 | 70.81 | 74.89 | 80.58 | 66.11 | 77.63 | 53.98 | 54.52 | 89.19 | 73.48 | 88.07 | 91.71 |
| F-score | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 65.17 | 69.92 | 72.15 | 40.17 | 42.58 | 29.15 | 29.15 | 22.06 | 22.06 | 74.14 | 74.14 |
| 0.1 | 67.11 | 70.94 | 75.41 | 54.68 | 70.43 | 40.65 | 33.07 | 66.46 | 28.41 | 86.79 | 76.96 |
| 0.2 | 67.23 | 70.64 | 74.14 | 59.61 | 72.17 | 45.71 | 31.79 | 85.98 | 37.41 | 89.41 | 78.17 |
| 0.3 | 63.50 | 66.44 | 76.77 | 59.49 | 75.69 | 46.50 | 43.60 | 88.06 | 55.45 | 89.45 | 89.75 |
| Adjusted Rand Index | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 61.15 | 66.49 | 69.05 | 33.38 | 36.16 | 20.29 | 20.29 | 7.52 | 7.52 | 71.26 | 71.26 |
| 0.1 | 63.40 | 67.53 | 72.64 | 49.30 | 66.99 | 33.59 | 33.59 | 62.29 | 16.03 | 85.32 | 74.41 |
| 0.2 | 63.47 | 67.07 | 71.04 | 54.85 | 69.06 | 39.30 | 39.30 | 84.41 | 27.30 | 88.24 | 75.74 |
| 0.3 | 59.11 | 62.31 | 74.07 | 54.59 | 72.90 | 39.87 | 39.87 | 86.72 | 48.98 | 88.28 | 88.54 |
| Average Entropy | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 0.97 | 0.89 | 0.85 | 1.86 | 1.83 | 2.22 | 2.22 | 2.49 | 2.49 | 0.82 | 0.82 |
| 0.1 | 0.91 | 0.82 | 0.76 | 1.30 | 0.90 | 1.82 | 2.03 | 0.78 | 1.90 | 0.44 | 0.73 |
| 0.2 | 0.92 | 0.82 | 0.74 | 1.15 | 0.82 | 1.68 | 1.96 | 0.44 | 1.50 | 0.38 | 0.72 |
| 0.3 | 1.00 | 0.90 | 0.67 | 1.17 | 0.76 | 1.57 | 1.56 | 0.37 | 1.04 | 0.40 | 0.30 |

Table 2: Clstering results on the VOC datasets.

| ACC | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 27.92 | 55.90 | 51.37 | 53.22 | 53.88 | 50.44 | 50.44 | 53.92 | 53.92 | 65.77 | 65.77 |
| 0.1 | 27.88 | 53.61 | 54.74 | 59.71 | 61.61 | 53.43 | 53.98 | 58.23 | 53.48 | 68.90 | 63.86 |
| 0.2 | 26.58 | 55.12 | 56.21 | 62.32 | 62.23 | 57.01 | 55.10 | 64.00 | 56.68 | 70.85 | 67.46 |
| 0.3 | 28.99 | 54.14 | 51.91 | 64.48 | 64.48 | 57.27 | 46.31 | 69.59 | 61.67 | 72.39 | 69.18 |
| NMI | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 22.05 | 53.88 | 50.61 | 51.80 | 52.46 | 48.78 | 48.78 | 53.60 | 53.60 | 66.43 | 66.43 |
| 0.1 | 20.87 | 56.25 | 55.41 | 58.09 | 55.95 | 61.68 | 63.86 | 60.10 | 52.71 | 70.98 | 66.99 |
| 0.2 | 19.49 | 62.00 | 62.68 | 32.79 | 60.57 | 57.69 | 56.10 | 62.58 | 56.58 | 72.4 | 64.73 |
| 0.3 | 20.87 | 64.99 | 63.11 | 65.82 | 63.85 | 61.27 | 46.04 | 70.48 | 61.78 | 73.90 | 68.64 |
| F-score | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 20.24 | 44.81 | 41.97 | 42.05 | 42.04 | 32.34 | 32.34 | 33.93 | 33.93 | 57.40 | 57.40 |
| 0.1 | 19.91 | 43.96 | 44.28 | 49.47 | 49.47 | 38.17 | 46.65 | 37.75 | 36.23 | 62.66 | 57.91 |
| 0.2 | 19.21 | 48.05 | 48.76 | 50.70 | 50.70 | 38.34 | 41.59 | 42.79 | 39.06 | 65.05 | 56.77 |
| 0.3 | 21.04 | 47.96 | 45.64 | 56.54 | 56.54 | 40.77 | 21.04 | 48.60 | 43.08 | 67.45 | 63.2 |
| Adjusted Rand Index | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 12.95 | 40.78 | 37.70 | 37.38 | 36.77 | 24.04 | 24.04 | 24.72 | 24.72 | 53.42 | 53.42 |
| 0.1 | 12.10 | 39.47 | 40.04 | 45.45 | 46.88 | 29.88 | 41.36 | 29.39 | 28.26 | 59.43 | 54.31 |
| 0.2 | 10.74 | 43.93 | 44.73 | 46.47 | 48.44 | 31.39 | 34.88 | 35.53 | 31.32 | 61.87 | 53.14 |
| 0.3 | 13.03 | 43.71 | 41.04 | 53.07 | 53.12 | 33.42 | 21.27 | 42.10 | 35.74 | 64.52 | 59.95 |
| Average Entropy | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 3.05 | 1.71 | 1.84 | 1.82 | 1.83 | 2.11 | 2.11 | 2.03 | 2.03 | 1.32 | 1.32 |
| 0.1 | 3.11 | 1.65 | 1.66 | 1.56 | 1.67 | 1.71 | 1.44 | 1.82 | 2.00 | 1.10 | 1.25 |
| 0.2 | 3.18 | 1.42 | 1.38 | 1.38 | 1.51 | 1.74 | 1.81 | 1.67 | 1.87 | 1.08 | 1.33 |
| 0.3 | 3.11 | 1.30 | 1.39 | 1.26 | 1.37 | 1.67 | 2.24 | 1.37 | 1.69 | 1.02 | 1.20 |

Table 3: Clstering results on the MIR datasets.

| ACC | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 30.92 | 26.84 | 26.95 | 27.90 | 30.79 | 32.25 | 32.25 | 30.93 | 30.93 | 32.66 | 33.66 |
| 0.1 | 28.37 | 28.36 | 28.55 | 29.53 | 33.44 | 31.66 | 28.15 | 40.74 | 30.53 | 45.45 | 35.08 |
| 0.2 | 25.93 | 33.01 | 28.67 | 43.43 | 42.39 | 40.51 | 31.53 | 45.37 | 38.32 | 51.41 | 41.52 |
| 0.3 | 24.55 | 43.35 | 42.88 | 48.04 | 48.49 | 41.34 | 33.47 | 48.37 | 39.94 | 52.95 | 45.96 |
| NMI | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 5.59 | 4.33 | 4.26 | 3.98 | 4.86 | 3.19 | 3.19 | 3.82 | 3.82 | 7.10 | 7.10 |
| 0.1 | 2.89 | 6.01 | 4.52 | 6.82 | 8.07 | 4.27 | 5.22 | 8.88 | 6.45 | 11.93 | 9.53 |
| 0.2 | 1.78 | 10.79 | 5.35 | 12.03 | 17.00 | 5.62 | 8.92 | 11.28 | 7.64 | 17.35 | 18.01 |
| 0.3 | 1.15 | 19.73 | 19.84 | 18.64 | 18.88 | 9.35 | 11.76 | 15.23 | 8.17 | 17.80 | 26.08 |
| F-score | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 25.96 | 24.03 | 24.03 | 24.40 | 24.80 | 27.24 | 27.24 | 25.13 | 25.13 | 28.39 | 28.39 |
| 0.1 | 23.89 | 26.17 | 24.57 | 29.32 | 29.80 | 25.60 | 27.26 | 28.47 | 27.85 | 31.11 | 35.08 |
| 0.2 | 23.12 | 27.56 | 25.36 | 34.58 | 35.40 | 33.51 | 29.76 | 30.64 | 29.40 | 34.60 | 37.05 |
| 0.3 | 22.88 | 30.58 | 30.77 | 35.91 | 35.42 | 31.77 | 30.44 | 32.90 | 31.33 | 36.40 | 38.20 |
| Adjusted Rand Index | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 3.36 | 1.90 | 1.86 | 1.86 | 2.66 | 3.45 | 3.45 | 2.41 | 2.41 | 4.35 | 4.35 |
| 0.1 | 1.85 | 1.46 | 2.74 | 2.06 | 3.04 | 3.06 | 1.00 | 6.52 | 2.17 | 9.57 | 3.22 |
| 0.2 | 1.08 | 3.20 | 2.14 | 6.59 | 6.46 | 5.94 | 2.16 | 9.65 | 7.21 | 14.82 | 6.63 |
| 0.3 | 0.57 | 8.27 | 8.52 | 12.16 | 12.21 | 6.17 | 2.37 | 12.12 | 8.32 | 16.12 | 8.77 |
| Average Entropy | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 2.03 | 2.06 | 2.06 | 2.07 | 2.05 | 2.09 | 2.09 | 2.00 | 2.00 | 2.00 | 2.00 |
| 0.1 | 2.09 | 2.03 | 2.05 | 2.02 | 1.99 | 2.06 | 2.05 | 1.96 | 2.02 | 1.89 | 1.99 |
| 0.2 | 2.11 | 1.93 | 2.04 | 1.92 | 1.86 | 2.04 | 1.98 | 1.91 | 1.98 | 1.77 | 1.83 |
| 0.3 | 2.13 | 1.73 | 1.73 | 1.77 | 1.76 | 1.96 | 1.92 | 1.82 | 1.88 | 1.75 | 1.67 |

Table 4: Clstering results on the 3Source datasets.

| ACC | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 34.44 | 56.45 | 58.11 | 53.25 | 56.81 | 56.45 | 56.45 | 55.50 | 55.50 | 68.88 | 68.88 |
| 0.1 | 34.32 | 67.57 | 67.57 | 68.16 | 68.76 | 65.68 | 62.60 | 68.75 | 59.53 | 70.89 | 75.15 |
| 0.2 | 34.44 | 56.33 | 57.99 | 62.96 | 66.75 | 69.11 | 70.41 | 66.63 | 66.51 | 72.43 | 72.07 |
| 0.3 | 35.15 | 43.43 | 44.62 | 67.21 | 76.87 | 73.49 | 73.73 | 73.25 | 74.56 | 79.17 | 77.04 |
| NMI | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 13.78 | 59.41 | 61.12 | 51.29 | 56.00 | 54.33 | 54.33 | 52.52 | 52.52 | 58.09 | 58.09 |
| 0.1 | 10.26 | 59.64 | 59.54 | 62.17 | 61.55 | 58.45 | 60.75 | 57.35 | 60.75 | 62.97 | 64.26 |
| 0.2 | 8.21 | 50.94 | 53.08 | 59.56 | 61.64 | 60.08 | 61.62 | 60.80 | 62.47 | 62.74 | 68.94 |
| 0.3 | 7.34 | 34.25 | 35.94 | 65.16 | 66.37 | 65.24 | 64.16 | 64.87 | 66.31 | 65.48 | 66.97 |
| F-score | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 30.46 | 54.54 | 56.29 | 48.62 | 54.75 | 53.09 | 53.09 | 50.48 | 50.48 | 61.63 | 61.63 |
| 0.1 | 36.97 | 65.97 | 67.20 | 65.01 | 65.37 | 66.42 | 62.57 | 55.70 | 58.82 | 68.96 | 69.72 |
| 0.2 | 36.33 | 53.80 | 56.44 | 60.49 | 63.54 | 61.74 | 64.52 | 64.60 | 66.01 | 66.17 | 68.56 |
| 0.3 | 36.59 | 35.95 | 37.18 | 61.28 | 74.42 | 72.46 | 71.41 | 68.06 | 67.23 | 75.23 | 71.81 |
| Adjusted Rand Index | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 6.01 | 42.79 | 44.58 | 35.50 | 42.04 | 40.74 | 40.74 | 35.70 | 35.70 | 51.00 | 51.00 |
| 0.1 | 3.81 | 55.86 | 56.82 | 55.80 | 55.89 | 57.00 | 51.89 | 42.17 | 48.57 | 59.13 | 61.29 |
| 0.2 | 1.79 | 38.90 | 43.01 | 49.84 | 53.55 | 50.01 | 55.17 | 54.28 | 55.13 | 55.45 | 60.15 |
| 0.3 | 1.83 | 5.81 | 6.78 | 53.90 | 66.86 | 63.62 | 62.88 | 58.65 | 57.26 | 67.15 | 64.01 |
| Average Entropy | | | | | | | | | | | |
| ratio | SCCA | PSC | CSC | SULF | PSLF | SinP | SinL | ConP | ConL | OurP | OurL |
| 0 | 1.96 | 0.88 | 0.84 | 1.07 | 0.98 | 1.00 | 1.00 | 1.09 | 1.09 | 0.94 | 0.94 |
| 0.1 | 2.05 | 0.93 | 0.97 | 0.82 | 0.85 | 0.94 | 0.90 | 0.98 | 0.83 | 0.90 | 0.80 |
| 0.2 | 2.04 | 1.16 | 1.07 | 0.89 | 0.85 | 0.93 | 0.87 | 0.89 | 0.85 | 0.91 | 0.70 |
| 0.3 | 2.12 | 1.64 | 1.61 | 0.81 | 0.80 | 0.86 | 0.83 | 0.79 | 0.75 | 0.85 | 0.74 |

the unreasonable utilization of the prior information. For example, the guiding of affinity matrix construction maybe harmed due to the mixture of labels and the calculated similarities.

SULF and PSLF utilize non-negative matrix factorization for the latent representation learning, and the partial label information is used to guide learning of the learned latent representation through a regression loss. Compared with them, we can directly punish the difference between the true label and the cluster indicator, and it is more direct and effective. Besides, feature selection is implemented to improve clustering performance. So, our proposed method outperforms SULF and PSLF.

Compared with SemiCCA, PairwiseSC and CentroidSC that use pairwise constraint information to guide the construction of the kernel matrix or covariance matrix, we increase the loss if the observed pairwise constraints are different between the one directly constructed from the cluster indicator, and it is more useful.

SinP and SinL achieve clustering through a view that obtains the best performance among all views. Compared with them, we utilize all the views, and accordingly complementarity among views are explored. This further proves the usefulness of utilizing multiple views for clustering.

ConP and ConL concatenate all views and cluster data without feature learning. Compared with them, view selection and feature selection in each view are performed, and accordingly more discriminative features are utilized. This validates that feature selection is necessary for clustering multiple low level features.

OurP and OurL use partial labels and pairwise constraints as prior information, respectively. From the tables, we observe that neither of them achieves the best clustering performance among all the datasets. So, given partial labels, we can change the prior to pairwise constraint, and compare their performance on the validation set before the future testing.

Next, we compare our model with [10], which learns an orthogonal subspace to approximate the cluster labels. Since [10] cannot deal with semi-supervised information, we ignore such prior knowledge with our method denoted as Mul-

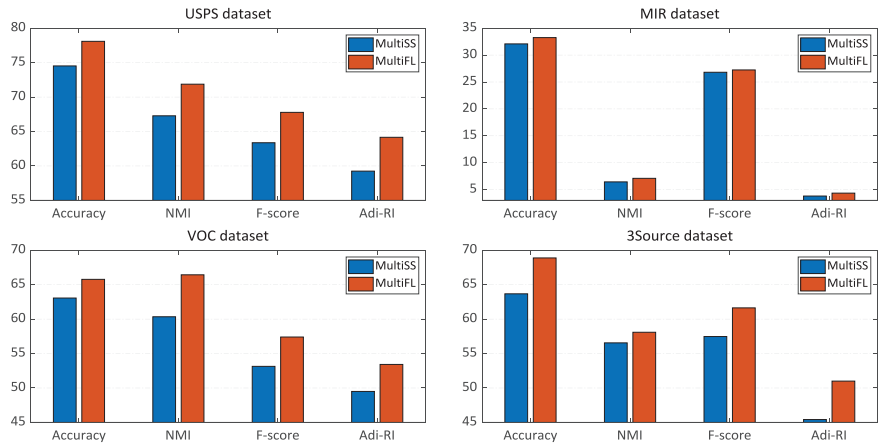


Figure 2: performance of MultiFL vs. MultiSS.

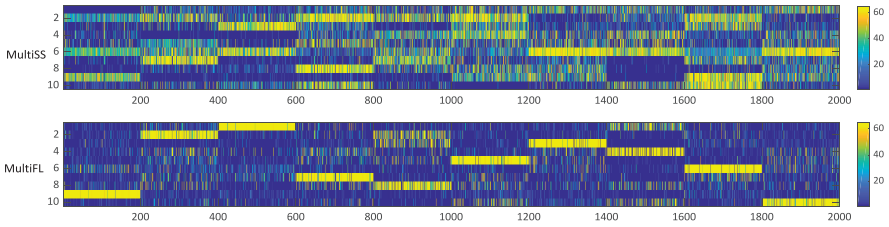


Figure 3: Visualization of MultiSS and MultiFL.

tiFL and [10] as MultiSS. The comparison results on the four datasets are shown in Figure 2. It can be seen that our method learns better cluster labels, and thus our method outperforms MultiSS. Furthermore, to clearly illustrate the learned cluster labels, we visualize them on the USPS dataset, and the results are shown in Figure 3. From the figure, the cluster structure learned by our method is obviously better than that of MultiSS, which further shows the effectiveness of our proposed method.

### *5.3. Convergence analysis*

In previous sections, we prove the convergence of our proposed optimization method. In this section, we give the convergence and NMI curves on the USPS

dataset with 10% labeled samples. Due to space limitation, we have conducted our experiments on this database, and it is possible that similar results would be obtained with other datasets. From Figure 4, the objectives in Equations (8) and (9) decrease with the increasing number of iterations, and the NMI results become better until no big changes. The results further illustrate the convergence of the proposed optimization approach.

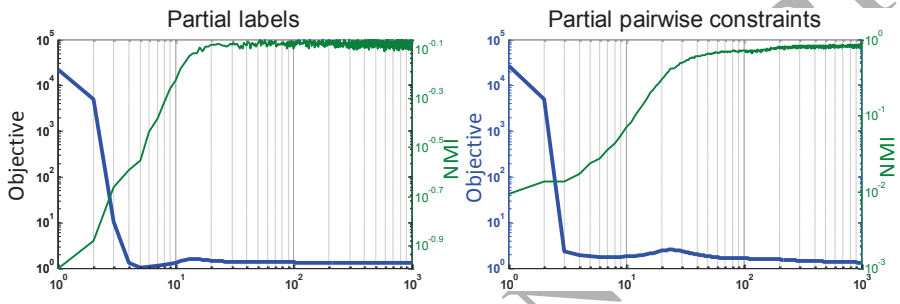


Figure 4: Convergence and NMI with varying iterations.

### *5.4. Running time*

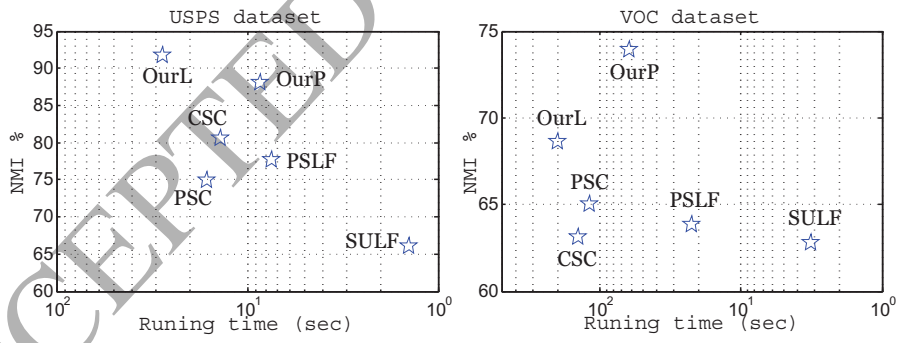


Figure 5: Running time on the USPS and VOC datasets.

In this section, we show the running time for obtaining the embeddings on the USPS and VOC datasets, where all the methods are run on the same machine (Intel CPU 3.1GHz and 12 GB memory). Besides, The publicly available codes

of all the methods, compared here, are written in MATLAB. From Figure 5, we can see that OurP and OurL obtain the best results, and the time used for obtaining an embedding is in the same magnitudes with the mainstream partial label based and pairwise constraints based methods, respectively.

Implementation of OurP is faster than that of OurL due to less matrix multiplication operation in Equation (14). For CSC and PSC methods, the kernel matrix of each view has to be calculated, and eigenvalue decomposition should be conducted for all views in each learning iteration. Compared with them, the learning of projection matrix of OurL is time consuming, which results in a little longer time cost for learning embeddings. Compared with PSLF and SULF methods, whose time cost are mainly for matrix multiplication operation, OurP needs more time with the same reason of OurL.

### *5.5. Parameter selection*

In our model, $\lambda_1$ and $\lambda_2$ are parameters balancing the view selection and feature selection in each view, respectively. From Figure 6, $\lambda_1$ and $\lambda_2$ can be empirically searched in $\{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$ once the data are normalized and will lead to acceptable results. Apart from the results conducted on the USPS dataset to validate the selection of $\lambda_1$ and $\lambda_2$, we emphasize how the performance varies with $\lambda_3$. From Figure 7, for partial labels, there is a large interval to achieve acceptable results, and for partial pairwise constraints, the interval is relatively small and this is because the absolute value of such regularizer is larger than that of partial labels and should be carefully controlled. When applied in real-world applications, it is possible to select these parameters based on the above suggestions. More advanced improvements will be left in the future work discussed in the Conclusion section.

### *5.6. Discussion*

#### *5.6.1. Initialization When Optimizing the Objective*

In the proposed optimization strategy as shown in Algorithm 1, we use random value to initialize the cluster indicator matrix $\mathbf{F}$. However, a better

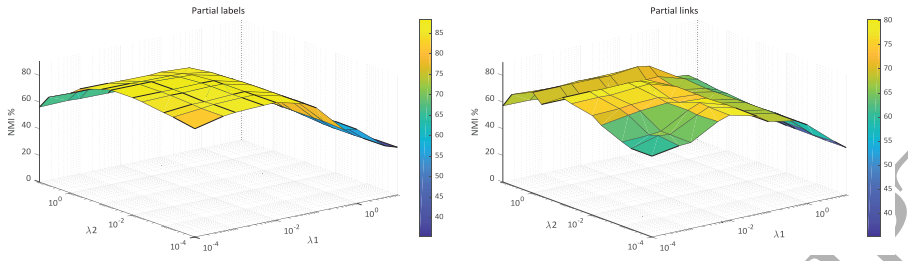


Figure 6: Performance vs. parameters $\lambda_1$ and $\lambda_2$ on the USPS dataset with LER being 0.1.

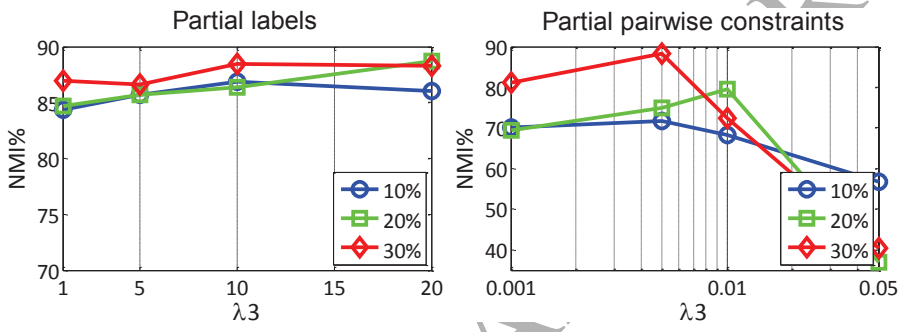


Figure 7: Performance vs. parameter $\lambda_3$ on the USPS dataset.

initialization may offer a more suitable starting point, so as to learn a better first latent space and avoid falling too fast in a bad local optimum. In this section, we show even a simple $k$means on the concatenated views without semi-supervised guidance can speed up the convergence. We conduct experiments on the USPS dataset with the partial labels and links being 0.1, respectively, and the results are shown in Figure 8. From the figure, the clustering results with $k$means initialization become saturated faster than the method with just random initialization.

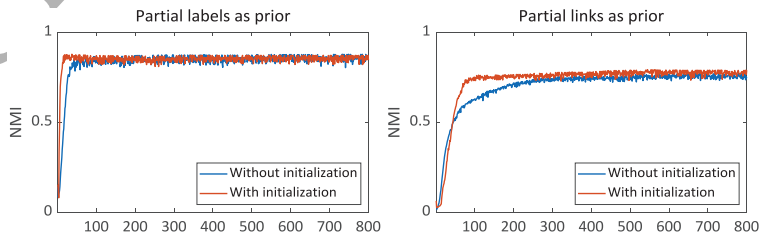


Figure 8: $k$means initialization on the cluster indicator matrix.

*5.6.2. Intercept Term in the Objective*

In this section, we investigate how the intercept term is related to the clustering performance, hoping to give a clue for selecting views. More specifically, we select any two views in a three view dataset, and plot the regression loss along with its clustering performance. The above loss and performance are compared with the results utilizing all the three views in the dataset. The intercept loss (curves) along with its NMI performance (values in the legend) on the USPS and the 3Source datasets are shown in Figure 9. It can be seen that the regression loss with the smallest value is always corresponding to the best performance, which may give a clue on selecting views based the intercept term. More specifically, a new view that can reduce loss of the intercept term maybe a good candidate for clustering. More systematic experiments are left in the future work due to the limitation of space.

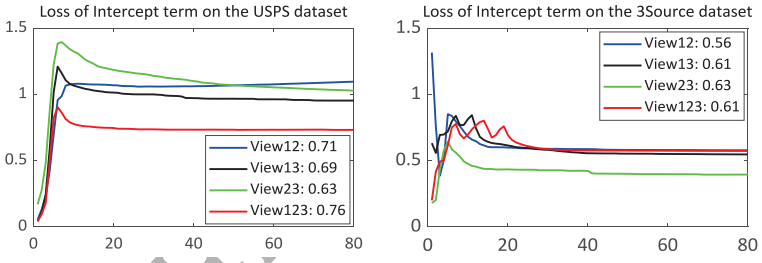


Figure 9: Loss of the intercept term.

## 6. Conclusion and future work

In this paper, we have proposed a novel multi-view clustering method by jointly considering feature learning and partially constrained cluster labels learning. Clustering label is the goal to be optimized, which is directly guided by the prior encoding the same high-level semantics. Besides, feature selection is embedded into the above framework for discriminative views and features selection. To solve the proposed objective, an effective optimization strategy is designed, especially for optimizing the clustering labels mixed with complete and partial

constrains. Finally, extensive experiments have validated the effectiveness of our model.

In our proposed method, we consider partial labels and partial pairwise constraints, respectively. If given mixed prior, one possible solution is to change partial labels to partial pairwise constraints. However, such operation may leads to performance degeneration due to different regularizers for encoding prior. In the future, we may consider jointing different kinds of prior knowledge into a unified objective to deal with such a scenario. Furthermore, utilization of feature selection and prior knowledge brings in three parameters to be tuned, which maybe difficult in real-world applications. We will consider auto-weighted multi-view learning framework to alleviate the above problem in the future.

## Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2016YFB1001004), by the National Natural Science Foundation of China (Grant No. 61876181), by the Guangdong Natural Science Foundation (Grant No. 2016A030313003).

## References

[1] L. Zhao, Z. Chen, L. Yang, Z. J. Wang, V. C. Leung, Incomplete multi-view clustering via deep semantic mapping, Neurocomputing 275 (2018) 1053–1062.

[2] S. Ding, H. Jia, M. Du, Y. Xue, A semi-supervised approximate spectral clustering algorithm based on hmrf model, Information Sciences 429 (2018) 215–228.

[3] S. Günnemann, I. Färber, M. Rüdiger, T. Seidl, Smvc: semi-supervised multi-view clustering in subspace projections, International Conference on Knowledge Discovery and Data Mining, (2014) 253–262.

[4] J. Liu, Y. Jiang, Z. Li, Z.-H. Zhou, H.-H. Lu, Partially shared latent factor learning with multiview data, IEEE Transactions on Neural Networks and Learning Systems 26 (6) (2015) 1233–1246.

[5] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: recent progress and new challenges, Information Fusion 38 (2017) 43–54.

[6] F. Nie, G. Cai, J. Li, X. Li, Auto-weighted multi-view learning for image clustering and semi-supervised classification, IEEE Transactions on Image Processing 27 (3) (2018) 1501–1511.

[7] D. Tolic, N. Antulov-Fantulin, I. Kopriva, Non-negative subspace clustering in nonlinear orthogonal non-negative matrix factorization framework, Pattern Recognition 82 (2018) 40–55.

[8] Y. Wang, L. Wu, X. Lin, J. Gao, Multiview spectral clustering via structured low-rank matrix factorization, IEEE Transactions on Neural Networks and Learning Systems 29 (10) (2018) 4833–4843.

[9] Z. Ding, Y. Fu, Dual low-rank decompositions for robust cross-view learning, IEEE Transactions on Image Processing 28 (1) (2019) 194–204.

[10] H. Wang, F. Nie, H. Huang, Multi-view clustering and feature learning via structured sparsity, International Conference on Machine Learning (2013) 352–360.

[11] Q. Yin, S. Wu, L. Wang, Unified subspace learning for incomplete and unlabeled multi-view data, Pattern Recognition 67 (2017) 313–327.

[12] J. Tang, X. Hu, H. Gao, H. Liu, Unsupervised feature selection for multi-view data in social media, SIAM International Conference on Data Mining (2013) 270–278.

[13] F. Nie, G. Cai, X. Li, Multi-view clustering and semi-supervised classification with adaptive neighbours, AAAI Conference on Artificial Intelligence (2017) 2408–2414.

[14] C. Hou, C. Zhang, Y. Wu, F. Nie, Multiple view semi-supervised dimensionality reduction, Pattern Recognition 43 (3) (2010) 720–730.

[15] X. Zhao, N. Evans, J.-L. Dugelay, A subspace co-training framework for multi-view clustering, Pattern Recognition Letters 41 (2014) 73–82.

[16] Z. Zhang, L. Liu, F. Shen, H. T. Shen, L. Shao, Binary multi-view clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[17] M. Hu, S. Chen, Doubly aligned incomplete multi-view clustering., in: International Joint Conference on Artificial Intelligence, 2018, pp. 2262–2268.

[18] Y. Yang, H. Wang, Multi-view clustering: A survey, Big Data Mining and Analytics 1 (2) (2018) 83–107.

[19] Q. Yin, S. Wu, L. Wang, Multiview clustering via unified and view-specific embeddings learning, IEEE Transactions on Neural Networks and Learning Systems 29 (11) (2018) 5541–5553.

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

[21] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, arXiv:1304.5634.

[22] X. Dong, P. Frossard, P. Vandergheynst, N. Nefedov, Clustering on multi-layer graphs via subspace analysis on grassmann manifolds, IEEE Transactions on Signal Processing 62 (4) (2014) 905–918.

[23] K. Chaudhuri, S. M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, International Conference on Machine Learning (2009) 129–136.

[24] A. Kumar, P. Rai, H. Daume, Co-regularized multi-view spectral clustering, Advances in Neural Information Processing Systems (2011) 1413–1421.

[25] J. Liu, C. Wang, J. Gao, J. Han, Multi-view clustering via joint nonnegative matrix factorization, SIAM International Conference on Data Mining (2013) 252–260.

[26] A. Klami, S. Kaski, Probabilistic approach to detecting dependencies between data sets, Neurocomputing 72 (1) (2008) 39–46.

[27] S. Bickel, T. Scheffer, Multi-view clustering, IEEE International Conference on Data Mining 4 (2004) 19–26.

[28] A. Kumar, H. Daumé, A co-training approach for multi-view spectral clustering, International Conference on Machine Learning (2011) 393–400.

[29] T.-L. Liu, Guided co-training for large-scale multi-view spectral clustering, arXiv:1707.09866.

[30] E. Bruno, S. Marchand-Maillet, Multiview clustering: a late fusion approach using latent models, International Conference on Research and Development in Information Retrieval (2009) 736–737.

[31] D. Greene, P. Cunningham, A matrix factorization approach for integrating multiple data views, International Conference on Knowledge Discovery and Data Mining (2009) 423–438.

[32] S. F. Hussain, M. Mushtaq, Z. Halim, Multi-view document clustering via ensemble method, Journal of Intelligent Information Systems 43 (1) (2014) 81–99.

[33] Q. Yin, S. Wu, R. He, L. Wang, Multi-view clustering via pairwise sparse subspace representation, Neurocomputing 156 (2015) 12–21.

[34] P. Muthukrishnan, D. Radev, Q. Mei, Edge weight regularization over multiple graphs for similarity learning, IEEE International Conference on Data Mining (2010) 374–383.

[35] R. Xia, Y. Pan, L. Du, J. Yin, Robust multi-view spectral clustering via low-rank and sparse decomposition, AAAI Conference on Artificial Intelligence (2014) 2149–2155.

[36] X. Cao, C. Zhang, H. Fu, S. Liu, H. Zhang, Diversity-induced multi-view subspace clustering, IEEE Conference on Computer Vision and Pattern Recognition (2015) 586–594.

[37] F. Nie, J. Li, X. Li, Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification, International Joint Conference on Artificial Intelligence (2016) 1881–728.

[38] F. Nie, W. Zhu, X. Li, Unsupervised feature selection with structured graph optimization, in: AAAI Conference on Artificial Intelligence, 2016, pp. 1302–1308.

[39] C. Lu, S. Yan, Z. Lin, Convex sparse spectral clustering: Single-view to multi-view, IEEE Transactions on Image Processing 25 (6) (2016) 2833–2843.

[40] M. Brbic, I. Kopriva, Multi-view low-rank sparse subspace clustering, Pattern Recognition 73 (2018) 247–268.

[41] M. Abavisani, V. M. Patel, Multimodal sparse and low-rank subspace clustering, Information Fusion 39 (2018) 168–177.

[42] Y. Jiang, J. Liu, Z. Li, H. Lu, Semi-supervised unified latent factor learning with multi-view data, Machine Vision and Applications 25 (7) (2014) 1635–1645.

[43] X. Shen, Q. Sun, A novel semi-supervised canonical correlation analysis and extensions for multi-view dimensionality reduction, Journal of Visual Communication and Image Representation 25 (8) (2014) 1894–1904.

[44] E. Eaton, M. Desjardins, S. Jacob, Multi-view clustering with constraint propagation for learning with an incomplete mapping between views, International Conference on Information and Knowledge Management (2010) 389–398.

[45] H. Yu, X. Wang, G. Wang, A semi-supervised three-way clustering framework for multi-view data, International Joint Conference on Rough Sets (2017) 313–325.

[46] W. Tang, Z. Lu, I. S. Dhillon, Clustering with multiple graphs, IEEE International Conference on Data Mining (2009) 1016–1021.

[47] X. Chen, S. Chen, H. Xue, X. Zhou, A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data, Pattern Recognition 45 (5) (2012) 2005–2018.

[48] D.-Q. Z. Peng Yan, Semi-supervised canonical correlation analysis algorithm, Journal of Software 19 (11) (2008) 2822–2832.

[49] F. Nie, D. Xu, I. W. Tsang, C. Zhang, Spectral embedded clustering, in: International Joint Conference on Artificial Intelligence, 2009, pp. 1181–1186.

[50] Q. Yin, S. Wu, L. Wang, Incomplete multi-view clustering via subspace learning, International Conference on Information and Knowledge Management (2015) 383–392.

[51] I. F. Gorodnitsky, B. D. Rao, Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm, IEEE Transactions on Signal Processing 45 (3) (1997) 600–616.

[52] C. Ding, T. Li, M. Jordan, et al., Convex and semi-nonnegative matrix factorizations, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (1) (2010) 45–55.

Qiyue Yin received PhD degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2017. Now he serves as an Assistant Professor in CASIA. His major research interests include pattern recognition, deep learning and artificial intelligence on games.

Junge Zhang received PhD degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2013. Now he serves as a Special-appointed Professor in CASIA. His major research interests include computer vision, pattern recognition, deep learning and general artificial intelligence.

Shu Wu received the Ph.D. degree in computer science from University of Sherbrooke, Canada, in 2012. He is an associate professor in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include data mining, recommendation systems, and pervasive computing.

Hexi Li is an associate professor and M.S. supervisor at Faculty of Intelligent Manufacturing, Wuyi University. His research interests include artificial intelligence and robot vision.