

Enhancing multi-view clustering through common subspace integration by considering both global similarities and local structures

Wanlin Weng^a, Weiwei Zhou^a, Jiazhou Chen^a, Hong Peng^a, Hongmin Cai^{a,b,c,*}

^aSchool of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

^bGuangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information, Guangzhou 510006, China

^cBioinformatics Center, Institute for Chemical Research, Kyoto University China

ARTICLE INFO

Article history:

Received 7 March 2019

Revised 4 September 2019

Accepted 12 October 2019

Available online 23 October 2019

Communicated by Dr. Chenping Hou

Keywords:

Multi-view learning

Subspace learning

Graph learning

Clustering

Integration

ABSTRACT

Multi-view clustering seeks to partition objects based on various observations by utilizing cross-views to provide a complementary description of the same objects. It remains challenging to effectively fuse the multi-view data with various dimensions as well as different structures into a new yet highly informative form, thus facilitating adequate assignment of the objects. To tackle the issue, we propose a common subspace integration (CSI) model. The CSI actively learns a common subspace by jointly preserving the local geometry of each view, while incorporating a global partition information to enhance its separability during the learning process. It can be easily generalized to its kernel version, thereby popularizing its general usages. An effective alternative optimization scheme is designed to solve the proposed model. Extensive experiments on six real-world datasets were conducted to demonstrate its superiority by comparing with the twelve state-of-art methods.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Currently, multi-view data is widely used in the engineering field. For example, a web page can be divided into two views, presenting words that occur on a given page and words that occur in the hyperlink to it [1,2]. Objects can be visualized by various imaging modalities, where image descriptors are developed to quantify objects. Different views usually provide supplemental information, and information obtained from a single view does not fully describe all related objects. Various collections from different domains usually exhibit a large number of heterogeneous properties [3,32,33]. It remains a challenge in accurately clustering multi-view objects by unifying the commonalities and complementarities from different views.

Multi-view clustering seeks to partition data points based on multiple representations by assuming that all views share the same cluster structure. By combining information from different views, the multi-view clustering algorithm attempts to achieve a more accurate clustering assignment than which is obtained by simply concatenating features from different views. In recent years, multi-view learning has gained great research interest as a new paradigm which uses redundant views of the same object to improve

learning performance. The general multi-view learning algorithms can be divided into three categories: 1) co-training; 2) multiple kernel learning; 3) subspace learning algorithms [4].

A groundbreaking co-training algorithm was proposed in [5] for semi-supervised classification of two views [6]. Recently, [7] uses a co-training scheme to train deep neural networks with extremely noisy labels. [8] combines active learning with co-training and proposes a robust semi-supervised learning algorithm. A Bayesian undirected graph model for co-training was proposed in [9]. The authors in [10] unified graph-based and disagree-based semi-supervised learning into a framework in which combinatorial label propagation was trained together on two views. It has been shown that if the two views satisfy the sufficiency and independence assumptions, then it can be guaranteed that the co-training works well [11]. If the above assumptions are violated, co-training may produce undesirable results. A practical guide [11] was designed to determine whether co-training should be applied.

Another popular method of analyzing multi-view data is through Multiple Kernel Learning (MKL) [12]. It was originally developed to control the capacity of the search space of possible kernel matrices to achieve good nonlinear generalization. MKL is considered to build multiple kernels in the sample space [4]. Recent years, MKL has been applied to many fields. Ramazzotti et al. [13] used MKL to integrate multi-omic tumor data so as to reveal the diversity of molecular mechanisms that correlate with survival, Li et al. [14] applied deep feature based MKL method to video

* Corresponding author at: School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China.

E-mail address: hmcai@scut.edu.cn (H. Cai).

emotion recognition. However, one of the main drawbacks of MKL is that the algorithms currently used to solve it are usually time-consuming thereby cannot be extended to the big dataset [15].

Multi-view subspace learning is one of the most popular techniques in multi-view learning. It assumes that representations of objects on different views share a common subspace. Most other methods are based on subspace learning [16–19]. Assume that each sample is sampled from a low-dimensional subspace. Therefore, clustering can be performed accurately on the learned subspace instead of each individual original view. In the processing of supervised data, a multi-view discriminant analysis algorithm is proposed in [20] by minimizing intra-class scattering while maximizing inter-class dispersion. For unsupervised learning, Zhang et al. [21] proposed a multi-view latent space clustering method which explores underlying complementary information from multiple views and performs clustering on the learned latent representation. Zhang et al. [22] viewed the multiple subspaces as a tensor and imposed a low-rank constraint on it to excavate the cross-view information. [23] exploits the complementary as well as the common information among views via an position-aware exclusivity term and a label consistency term, respectively. Zhuge et al. [24] clusters data points by considering the importance of multiple views and the importance of features in each view simultaneously.

In this paper, we propose a multi-view common subspace integration (CSI) method. CSI actively learns a common subspace by preserving the local geometry of each view. Moreover, global partition information is incorporated into CSI to enhance its separability during the subspace learning process. Therefore, by integrating both the local and global similarities, CSI is shown to achieve superior performance in clustering multi-view data. We also generalize the CSI to its kernel version (KCSI) to popularize its usage. An efficient iterative numerical algorithm is designed to solve the CSI. Extensive experiments on six real-world data sets demonstrate the effectiveness of our approaches.

The proposed method CSI and KCSI provide three advantages over the competition algorithm:

- The CSI exploits the key characteristics of each single view and incorporates them during subspace learning, thus achieving a well utilization of multiple views.
- The CSI learns a common subspace to extract the global commonalities of each view by minimizing the difference between the global and local structures, thus possessing superior separability.
- The CSI could be naturally generalized to its kernel counterpart, thus enjoying a large popularization in various scenarios.

The remainder of the paper is organized as follows: In Section 2, recent advances related to multi-view subspace clustering is introduced. In Section 3, we detail the formulation of the proposed CSI and KCSI. The corresponding numerical algorithm is reported in Section 3.4. In Section 4, extensive experiments on six empirical datasets were conducted and compared with the benchmark methods to evaluate their performances. We draw concluding remarks in Section 5.

2. Notations and related work

2.1. Notations

Throughout this paper, we let vectors and matrices be denoted by bold lowercase and uppercase characters, respectively. Let $\text{Tr}(\mathbf{X})$ denotes the trace of a matrix $\mathbf{X}$, and $\text{rank}(\mathbf{X})$ denotes its rank. The matrix norms, such as l_1 -norm, Frobenius norm and nuclear norm,

are defined as

$$\|\mathbf{X}\|_1 = \sum_{j=1}^n \sum_{i=1}^d |x_{ij}|$$

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^d \sum_{j=1}^n x_{ij}^2}$$

$$\|\mathbf{X}\|_* = \sum_{i=1}^r \sigma_i, \text{ where } \sigma_i\text{'s are the singular values of } \mathbf{X}.$$

2.2. Related work

In this section, we introduce subspace clustering and constrained Laplacian rank method that is related to our method.

Subspace Clustering: Subspace clustering is an extension of traditional clustering that seeks to find clusters in different subspaces within a dataset. Consider a set of n data points $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^n$ that lie in several hidden linear subspaces, the aim of subspace clustering is to assign each sample into an adequate subspace. To ease clustering in subspace, one may incorporate additional constraints to characterize the subspace. For example, one can enforce a low-rank representation (LRR) aims to obtain an affinity matrix with block diagonal matrices. Each block is used to denote a subspace. Another popular configuration is to enforce a sparse space representation (SSC).

Low-Rank Representation (LRR) aims to find a low-rank representation matrix $\mathbf{Z} \in \mathbb{R}^{n \times n}$ for input data $\mathbf{X}$. The model of LRR is:

$$\begin{aligned} \min_{\mathbf{Z}} \|\mathbf{Z}\|_* \\ \text{s.t. } \mathbf{X} = \mathbf{XZ} \end{aligned} \quad (1)$$

SSC assumes that each sample could be represented by a small number of samples from its own subspace by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{Z}} \|\mathbf{Z}\|_1 \\ \text{s.t. } \mathbf{X} = \mathbf{XZ}, \text{diag}(\mathbf{Z}) = 0 \end{aligned} \quad (2)$$

Once the subspace $\mathbf{Z}$ is obtained by LRR or SSC, an affinity matrix could be estimated by its average. Finally, one can use popular clustering methods, such as spectral clustering, to have the clustering assignment.

Constrained Laplacian Rank Method (CLR): CLR is a heuristic yet effective graph learning method. It aims to learn a new graph $\mathbf{S}$ from a given graph $\mathbf{Z}$. The learned new graph $\mathbf{S}$ has clearer block-diagonal structure, and it can be used directly to partition the data points into k clusters actually. Hence, clustering is more easily to be applied in such a new graph. The CLR is formulated into the following equation:

$$\begin{aligned} \min_{\mathbf{S}} \|\mathbf{S} - \mathbf{Z}\|_F^2 \\ \text{subject to:} \end{aligned} \quad (3)$$

$$\sum_j \mathbf{S}_{ij} = 1, \mathbf{S}_{ij} \geq 0, \text{rank}(\mathbf{L}_\mathbf{S}) = n - k$$

where $\mathbf{L}_\mathbf{S} = \mathbf{D}_\mathbf{S} - (\mathbf{S}^T + \mathbf{S})/2$, and $\mathbf{D}$ is the degree matrix of $\mathbf{S}$. The parameter k is the number of connected components in graph $\mathbf{S}$ and it turns to be same as the cluster number, as is shown bellow.

Theorem 1 [25]. The multiplicity c of the eigenvalue zero of the Laplacian matrix $\mathbf{L}_\mathbf{S}$ is equal to the number of connected components in the graph associated with $\mathbf{X}$.

Similarly to the SSC method, once the graph $\mathbf{S}$ is obtained, one can use popular clustering methods, such as spectral clustering, to have the clustering assignment.

3. Proposed method

3.1. Local structure preserved multi-view subspace learning

Given a dataset containing n objects from m views, the data in the v th view is represented as $\mathbf{X}^v = [\mathbf{x}_1^v, \mathbf{x}_2^v, \dots, \mathbf{x}_n^v] \in \mathbb{R}^{d_v \times n}$, where d_v is the dimensionality of the v th view. We firstly map each view into different subspace, with $\mathbf{Z}^v = [\mathbf{z}_1^v, \mathbf{z}_2^v, \dots, \mathbf{z}_n^v]$ to denoting mapped value in the subspace for the v th view. To preserve the local affinity of the same sample among different views, one may impose a linear transform by which each pair of samples should be closed to each other within subspace if they are close in the observed space. The goal could be achieved by using a graph constraint:

$$\sum_v \sum_{i,j=1}^n w_{ij}^v \|\mathbf{z}_i^v - \mathbf{z}_j^v\|_2^2 = \sum_v \text{Tr}(\mathbf{Z}^v \mathbf{L}^v (\mathbf{Z}^v)^T) \quad (4)$$

where $\mathbf{W}^v = (w_{ij}^v)$ is the weighting matrix of $\mathbf{X}^v$ used to penalize the similarity between i th and j th sample in raw space. The corresponding Laplacian matrix is given by $\mathbf{L}^v = \mathbf{D}^v - \mathbf{W}^v$, in which $\mathbf{D}_{ii}^v = \sum_j w_{ij}^v$. Throughout this paper, each weighting matrix $\mathbf{W}^v = [w_{ij}^v]$ is estimated by Gaussian scaled distance of the k -nearest neighbors, and k is set to be 5 for simplicity. The analysis of the effect of k is demonstrated in Section 4.3. The weighting matrix is defined by,

$$w_{ij}^v = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i^v - \mathbf{x}_j^v\|^2}{2}\right), & \mathbf{x}_j^v \in k\text{-nearest neighbors of } \mathbf{x}_i^v \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Naturally, by combining subspace clustering and above Laplacian regularization, one can enforce the estimated subspaces possessing graph-like characteristics,

$$\begin{aligned} \min_{\mathbf{Z}^v} \quad & \sum_{v=1}^m \|\mathbf{X}^v - \mathbf{X}^v \mathbf{Z}^v\|_F^2 + \alpha \sum_{v=1}^m \text{Tr}(\mathbf{Z}^v \mathbf{L}^v (\mathbf{Z}^v)^T) \\ \text{s.t.} \quad & \text{diag}(\mathbf{Z}^v) = 0 \end{aligned} \quad (6)$$

This method learns each subspace independently. Finally, one could directly take an average of all $\mathbf{Z}^v$ to an integrated result. Thus, clustering of the multiple view data could be carried out on the integrated matrix.

3.2. Common subspace integration

The aforementioned model seamlessly incorporates the local manifold into the learning process. However, it is a decoupled system in that it failed to take advantage of the complementarity among different views. Moreover, in order to cluster the multi-view data, one has to fuse the learned representation $\mathbf{Z}^v$, $v = 1, 2, \dots, m$ into a unified representation. Obviously, simply averaging the new representations could not fulfill the task.

To overcome the drawbacks, we proposed a common subspace integration model (CSI). The CSI aims to learn a centroid-based common subspace representation which enforces view-specific representations diffused towards a common centroid $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n]$. The CSI also considers the local structures within each view to guide the fusion of common centroid. Both the global and local constraints work competitively to promote the learning of a common centroid which serves as an unique but informative representation of the multiple views. The CSI is

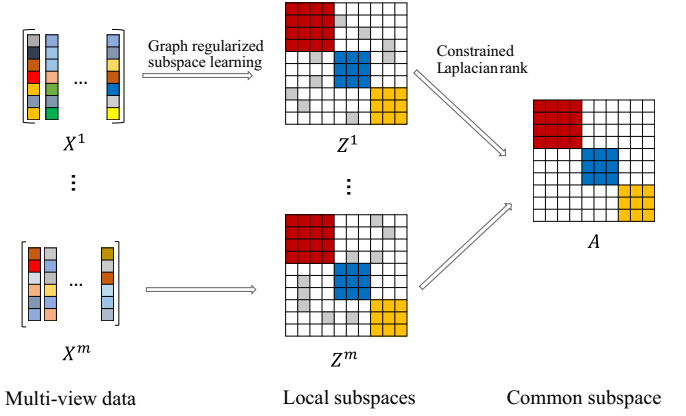


Fig. 1. The framework of proposed multi-view common subspace integration (CSI). With the multi-view input, our method independently learns the local subspaces by graph regularized subspace learning, then a common subspace is obtained by constrained Laplacian rank method. We use the common subspace to explore the global consistent information from all single views.

formulated as an energy minimization problem as follows,

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{Z}^v} \quad & \sum_{v=1}^m \|\mathbf{X}^v - \mathbf{X}^v \mathbf{Z}^v\|_F^2 + \lambda \sum_{v=1}^m \|\mathbf{A} - \mathbf{Z}^v\|_F^2 \\ & + \alpha \sum_{v=1}^m \text{Tr}(\mathbf{Z}^v \mathbf{L}^v (\mathbf{Z}^v)^T) \end{aligned} \quad (7)$$

subject to:

$$\begin{aligned} \text{diag}(\mathbf{Z}^v) &= 0, \\ \mathbf{1}^T \mathbf{a}_i &= 1, 0 \leq \mathbf{a}_i \leq 1, i = 1, 2, \dots, n, \\ \text{rank}(\mathbf{L}_A) &= n - c \end{aligned}$$

where $\mathbf{L}^v$ is Laplacian matrix of $\mathbf{X}^v$, $\mathbf{L}_A = \mathbf{D} - \frac{\mathbf{A}^T \mathbf{A}}{2}$ is the Laplacian matrix corresponding to $\mathbf{A}$, $\mathbf{a}_i$ is the i -th column of common subspace representation $\mathbf{A}$.

The Theorem 1 indicates that if $\text{rank}(\mathbf{L}_A) = n - c$, then the graph derived from the common subspace $\mathbf{A}$ is an ideal graph which we can partition into c clusters. There is a trial case where some columns of the common subspace representation $\mathbf{A}$ are all being zeros. To avoid the pitfall, we impose two constraints on the common subspace representation $\mathbf{A}$. Its column-wise sum is unit, and each element is lying between zero and one.

In CSI formation Eq. (7), each affinity matrix from different views is learned from Laplacian regularized subspaces. They work jointly to create a common subspace representation. The constrained Laplacian rank method in multi-view clustering offers three main advantages: (1) preserves consistency among multiple views; (2) discards misleading noises from each view; (3) makes the block-diagonal structure of affinity matrix clearer. On this account, we use the common subspace obtained by CLR as the input of spectral clustering to promote clustering performance. The framework of the CSI is demonstrated in Fig. 1.

3.3. Kernel extension of CSI

When the v -th view space $\mathbf{X}^v$ lies in infinite-dimensional Hilbert space, there exists a nonlinear mapping $\phi^v: \mathbb{R}^{d^v} \rightarrow \mathcal{H}$ that maps the samples from the input space to a reproducing kernel Hilbert space $\mathcal{H}$. For $\mathbf{X}^v = [\mathbf{x}_1^v, \dots, \mathbf{x}_n^v]$ containing n samples, the map is $\phi^v(\mathbf{X}^v) = [\phi^v(\mathbf{x}_1^v), \dots, \phi^v(\mathbf{x}_n^v)]$. We can define the kernel similarity between data samples $\mathbf{x}_i^v$ and $\mathbf{x}_j^v$ as $\mathbf{K}_{\mathbf{x}_i^v, \mathbf{x}_j^v}^v = \langle \phi^v(\mathbf{x}_i^v), \phi^v(\mathbf{x}_j^v) \rangle$. Using kernel trick, we can easily compute all similarities exclusively using kernel function and do not need to know

the exact ϕ^v . The kernel extension of CSI (KCSI) can be obtained through

$$\begin{aligned} \|\mathbf{x}_i^v - \mathbf{x}_i^v \mathbf{z}_i^v\|_F^2 &= \langle \phi^v(\mathbf{x}_i^v) - \phi^v(\mathbf{x}_i^v) \mathbf{z}_i^v, \phi^v(\mathbf{x}_i^v) - \phi^v(\mathbf{x}_i^v) \mathbf{z}_i^v \rangle \\ &= \langle \phi^v(\mathbf{x}_i^v), \phi^v(\mathbf{x}_i^v) \rangle - 2 \langle \phi^v(\mathbf{x}_i^v), \phi^v(\mathbf{x}_i^v) \mathbf{z}_i^v \rangle \\ &\quad + \langle \phi^v(\mathbf{x}_i^v) \mathbf{z}_i^v, \phi^v(\mathbf{x}_i^v) \mathbf{z}_i^v \rangle \\ &= \mathbf{K}_{\mathbf{x}_i^v, \mathbf{x}_i^v}^v - 2 \mathbf{K}_{\mathbf{x}_i^v, \mathbf{x}_i^v}^v \mathbf{z}_i^v + (\mathbf{z}_i^v)^T \mathbf{K}_{\mathbf{x}_i^v, \mathbf{x}_i^v}^v \mathbf{z}_i^v \end{aligned} \quad (8)$$

Therefore, we can get the kernel extension of (7) as below:

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{Z}^v} \quad & \sum_{v=1}^m \text{Tr}(\mathbf{K}^v - 2 \mathbf{K}^v \mathbf{Z}^v + (\mathbf{Z}^v)^T \mathbf{K}^v \mathbf{Z}^v) \\ & + \lambda \sum_{v=1}^m \|\mathbf{A} - \mathbf{Z}^v\|_F^2 + \alpha \sum_{v=1}^m \text{Tr}(\mathbf{Z}^v \mathbf{L}^v (\mathbf{Z}^v)^T) \end{aligned} \quad (9)$$

subject to:

$$\begin{aligned} \text{diag}(\mathbf{Z}^v) &= 0, \\ \mathbf{1}^T \mathbf{a}_i &= 1, 0 \leq \mathbf{a}_i \leq 1, i = 1, 2, \dots, n, \\ \text{rank}(\mathbf{L}_A) &= n - c \end{aligned}$$

The kernelled problem (9) can be optimized using the same technique as the linear CSI defined by Eq. (7).

3.4. Numerical solution of CSI

We now sketch the numerical scheme to solve the proposed model (7). Since the Lapacian matrix $\mathbf{L}_A$ is positive semi-definite, its i -th smallest eigenvalue $\sigma_i(\mathbf{L}_A)$ is non-negative. One can verify that the rank constraint $\text{rank}(\mathbf{L}_A) = n - c$ is equivalent to $\sum_{i=1}^c \sigma_i(\mathbf{L}_A) = 0$. Therefore, the minimization problem (7) could be reformatted as:

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{Z}^v} \quad & \sum_{v=1}^m \|\mathbf{X}^v - \mathbf{X}^v \mathbf{Z}^v\|_F^2 + \lambda \sum_{v=1}^m \|\mathbf{A} - \mathbf{Z}^v\|_F^2 \\ & + \alpha \sum_{v=1}^m \text{Tr}(\mathbf{Z}^v \mathbf{L}^v (\mathbf{Z}^v)^T) + \beta \sum_{i=1}^c \sigma_i(\mathbf{L}_A) \end{aligned} \quad (10)$$

subject to:

$$\begin{aligned} \text{diag}(\mathbf{Z}^v) &= 0, \\ \mathbf{1}^T \mathbf{a}_i &= 1, 0 \leq \mathbf{a}_i \leq 1, i = 1, 2, \dots, n \end{aligned}$$

where β is a Lagrange dual parameter.

According to Ky Fans Theorem [26],

$$\sum_{i=1}^c \sigma_i(\mathbf{L}_A) = \min_{\mathbf{P}^T \mathbf{P} = \mathbf{I}} \text{Tr}(\mathbf{P}^T \mathbf{L}_A \mathbf{P}) \quad (11)$$

for some orthogonal matrix $\mathbf{P}$.

Therefore, Eq. (10) can be further formatted as

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{Z}^v, \mathbf{P}} \quad & \sum_{v=1}^m \|\mathbf{X}^v - \mathbf{X}^v \mathbf{Z}^v\|_F^2 + \lambda \sum_{v=1}^m \|\mathbf{A} - \mathbf{Z}^v\|_F^2 \\ & + \alpha \sum_{v=1}^m \text{Tr}(\mathbf{Z}^v \mathbf{L}^v (\mathbf{Z}^v)^T) + \beta \text{Tr}(\mathbf{P}^T \mathbf{L}_A \mathbf{P}) \end{aligned} \quad (12)$$

subject to:

$$\begin{aligned} \text{diag}(\mathbf{Z}^v) &= 0, \\ \mathbf{1}^T \mathbf{a}_i &= 1, 0 \leq \mathbf{a}_i \leq 1, i = 1, 2, \dots, n \\ \mathbf{P}^T \mathbf{P} &= \mathbf{I} \end{aligned}$$

The final objective function (12) is convex, we use an alternat optimization strategy to solve it.

In the **first step**, by fixing both $\mathbf{A}$ and $\mathbf{P}$, (12) is reduced to

$$\begin{aligned} \min_{\mathbf{Z}^v} \quad & \|\mathbf{X}^v - \mathbf{X}^v \mathbf{Z}^v\|_F^2 + \lambda \|\mathbf{A} - \mathbf{Z}^v\|_F^2 + \alpha \text{Tr}(\mathbf{Z}^v \mathbf{L}^v (\mathbf{Z}^v)^T) \end{aligned} \quad (13)$$

subject to:

$$\text{diag}(\mathbf{Z}^v) = 0$$

Setting the gradient the of equation with respect to $\mathbf{Z}^v$ to zero,

$$(\mathbf{X}^v)^T \mathbf{X}^v \mathbf{Z}^v - (\mathbf{X}^v)^T \mathbf{X}^v + \lambda (\mathbf{Z}^v - \mathbf{A}) + \alpha \mathbf{L}^v \mathbf{Z}^v = 0 \quad (14)$$

s.t. $\text{diag}(\mathbf{Z}^v) = 0$

Otherwise, equivalently, we can obtain an explicit solution:

$$\mathbf{Z}^v = [(\mathbf{X}^v)^T \mathbf{X}^v + \lambda \mathbf{I} + \alpha \mathbf{L}^v]^{-1} ((\mathbf{X}^v)^T \mathbf{X}^v + \lambda \mathbf{A}) \quad (15)$$

with $\text{diag}(\mathbf{Z}^v) = 0$

In the **second step**, by fixing $\mathbf{Z}^v$ and $\mathbf{A}$, (12) is reduced to

$$\min_{\mathbf{P}^T \mathbf{P} = \mathbf{I}} \text{Tr}(\mathbf{P}^T \mathbf{L}_A \mathbf{P}) \quad (16)$$

It turns out to be an generalized eigenvalue problem. Its optimal solution is obtained by the first c smallest eigenvectors of $\mathbf{L}_A$.

In the **third step**, by fixing both $\mathbf{Z}^v$ and $\mathbf{P}$, (12) is reduced to

$$\begin{aligned} \min_{\mathbf{A}} \quad & \sum_{v=1}^m \|\mathbf{A} - \mathbf{Z}^v\|_F^2 + \beta \text{Tr}(\mathbf{P}^T \mathbf{L}_A \mathbf{P}) \end{aligned} \quad (17)$$

subject to:

$$\begin{aligned} \mathbf{1}^T \mathbf{a}_i &= 1, \\ 0 \leq \mathbf{a}_i &\leq 1, i = 1, 2, \dots, n \end{aligned}$$

By expanding the second term $\text{Tr}(\mathbf{P}^T \mathbf{L}_A \mathbf{P})$ as

$$\text{Tr}(\mathbf{P}^T \mathbf{L}_A \mathbf{P}) = \sum_{i,j} \frac{1}{2} \|\mathbf{P}_{i,:} - \mathbf{P}_{j,:}\|^2 \mathbf{a}_{ij} \quad (18)$$

The equation (17) can be reformulated column-wisely as

$$\begin{aligned} \min_{\mathbf{a}_i} \quad & \sum_{v=1}^m \|\mathbf{a}_i - \mathbf{Z}_i^v\|^2 + \beta \mathbf{d}_i^T \mathbf{a}_i \end{aligned} \quad (19)$$

s.t. $\forall i, \mathbf{1}^T \mathbf{a}_i = 1, 0 \leq \mathbf{a}_i \leq 1$

where $\mathbf{d}_i \in \mathbb{R}^{n \times 1}$ is a vector with its j -th element being $\mathbf{d}_{ij} = \|\mathbf{P}_{i,:} - \mathbf{P}_{j,:}\|^2$. Therefore, with respect to each column, it is a constrained least square problem. Thus, it can be efficiently solved by standard quadratic programming methods.

The above three steps are iteratively solved and updated until convergence. Once the intergraded central subspace representation $\mathbf{A}$ is obtained, standard clustering method can be employed on it to have the cluster assignments.

The complete algorithm is summarized in Algorithm 1.

Algorithm 1 The algorithm for solving CSI.

Input: $\mathbf{X} = \{\mathbf{x}_i^v | 1 \leq i \leq n, 1 \leq v \leq m\}$, λ , α

1: Initialize $\mathbf{A}$, $\mathbf{P}$, $\mathbf{Z}^v$ as identity matrix. Construct Laplacian matrix $\mathbf{L}^v$.

2: **repeat**

3: Update $\mathbf{Z}^v (v = 1, 2, \dots, m)$ by solving Eq.(15).

4: Update $\mathbf{P}$ by solving Eq.(16).

5: Update $\mathbf{A}$ by solving Eq.(19).

6: **until** convergence

Output: $\mathbf{A}$.

3.5. Complexity analysis

For the readability, we only provide complexity analysis for CSI. The numerical scheme in solving CSI consists of iterating of three major steps. In each round of the iterations, the three steps are all having explicit solutions and thus the whole scheme is rather efficient. In particular, it involves a matrix inversion in **step 1**. Fortunately, the inversion of $(\mathbf{X}^v)^T \mathbf{X}^v$ in Eq. (15) needs to be calculated just once. It costs $O(mn^3 + dn^2)$, with $d = \sum_{v=1}^m d_v$ being the total number of features in all views and n being sample size.

Table 1

Summary of the multi-view datasets used in our experiments.

Datasets	Size	Number of views	Number of clusters
MIT-8-scene	2688	4	8
Caltech-101	1474	6	7
BBCSports	282	3	5
3Sources	169	3	6
ORL	400	3	40
Yale	165	3	15

The **step 2** is a typical eigenvalue problem with cost known to be $O(cn^2)$, where c is the cluster number. In **step 3**, computing $\mathbf{A}$ is a quadratic programming problem and its complexity is $O(mn^3)$. Finally, the total cost of **Algorithm 1** is $O(mn^3 + dn^2 + T(cn^2 + mn^3))$, where T is the total iteration number. In our experiments, we found that the objective value decreased very fast and the convergence could be reached within a small number of iterations. Therefore, the total cost is about $O(mn^3)$.

4. Experiments

We are now testing the proposed CSI and KCSI method extensively on six real-world datasets. The polynomial kernel $\mathbf{K}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + 1)^2$ is used through the experiments when testing KCSI. Six public available datasets which are widely evaluated as benchmark examples are used for experiments. The characteristics of the datasets are summarized in **Table 1**.

To validate the effectiveness of our method, twelve recently reported methods are borrowed and compared with the proposed CSI and KCSI. We shall give a short introduction to the twelve methods below.

1. **Single View Spectral Clustering (SC)** [27], spectral clustering is applied on every single view to demonstrate the improvement of multi-view methods.
2. **ConcatFea** concatenates all the features and performs standard spectral clustering.
3. **ConcatPCA** concatenates all the features and then projects original features into low dimensional subspace that preserve 99% message via PCA. Spectral clustering is applied in the low dimensional space.
4. **MVGL** [28] learns a global graph from different single view graphs. We use the default setting as the authors' suggestion.
5. **SNF** [29] uses the k -nearest neighbor graphs, and iteratively updates similarity. The setting of the parameter l is searched in the range of [0.3, 0.8] with interval 0.1 as authors' suggestion and other parameters by default.
6. **Co-Reg** [30] co-training strategy to alternatively modify the graph structure of each view using other views information.co-regularizes the clustering hypotheses to enforce the memberships from different views admit with each other.
7. **Co-Pair** [30] is the pairwise co-regularization version of Co-Reg.
8. **DiMSC** [16] investigates the complementary information of representations of multi-view data by introducing a diversity term.
9. **LMSC** [21] learns a latent representation with complementarity from multi-view data and then performs subspace clustering on the learned latent representation.
10. **LT-MSC** [22] regards the multi-view representations as a tensor and imposes a low-rank constraint on it to explore the high-order correlation among views.
11. **RAMSC** [24] learns a common representation from multiple views with auto weights.

12. **ECMSC** [23] is a subspace clustering method which considers complementarity and consistency simultaneously by introducing a position-aware exclusivity term and an indicator consistency term, respectively.

To make a comprehensive evaluation, we use different evaluation metrics including normalized mutual information (**NMI**), accuracy (**ACC**), adjusted rand index (**ARI**), **F-score**, **Precision**, **Recall** and **Purity**. For all the metrics, a higher value implies a better performance. Different measurements favor different properties, thus a comprehensive view can be acquired from the diverse results. For fair comparison, each experiment was repeated ten times independently and the mean values and standard deviations were reported. Moreover, the parameters in each compared method is tuned to meet the best performance in the suggested range. All our experiments were conducted on a desktop computer with a 4.0 GHz Intel Core i7 CPU and 12 GB of RAM and used MATLAB R2016a (64 bit).

4.1. Dataset description

MIT-8-scene comprises of 2688 pictures belonging to 8 different outdoor scene categories. This dataset was originally released by Oliva and Torralba [31] to do scene classification tasks. The categories are Tall Building (with 356 images), City Center (with 308 images), Street (with 292 images), Highway (with 260 images), Coast and Beach (with 360 images), Open Country (with 410 images), Mountain (with 374 images), Forest (with 328 images). We extracted four types of features: a 512-dimension GIST feature, a 432-dimension HOG feature, a 256-dimension LBP feature and a 48-dimension Gabor feature.

Caltech-101 is a dataset of pictures of objects belonging to 101 categories. Following past work, we extracted five features from all images: a 48-dimension Gabor feature, 40 dimension wavelet moments (Wm), a 254-dimension centrist feature, a 1984-dimension HOG feature, a 512-dimension Gist feature, and a 928-dimension LBP feature. Each feature could be considered a view of an object. Seven widely used classes (Face, Motorbikes, Dolla-Bill, Garfield, Snoopy, Stop-Sign, and Windsor-Chair) consisting of 1474 images were selected to form the Caltech7 subset. The Caltech7 dataset consisted of six feature sets.

BBCSports Consists of 737 documents from the BBC Sports website corresponding to sports news articles in five topical areas from 2004 to 2005. For our multi-view experiments, we choose 282 reports that appear in all three views.

3Sources is a three-view text dataset collected from three online news sources. In total there are 948 news articles covering 416 distinct news stories. All articles are in the bag-of-words representation. Of these stories, 169 were reported in all three sources. For our multi-view experiments, the dataset containing 169 articles was used, it contains 6 clusters.

ORL dataset contains 10 different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open or closed eyes, smiling or not smiling) and facial details (glasses or no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). We extracted 4096 dimensional raw pixel values, 3304 dimensional LBP feature, and 6750 dimensional Gabor feature as 3 views for our experiments.

Yale dataset contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, with glasses, happy, left-light, without glasses, normal, right-light, sad, sleepy, surprised, and wink. We extracted 4096 dimensional raw pixel

Table 2
Clustering results on ORL, Yale and 3Sources.

Dataset	Method	NMI	ACC	AR	F-score	Precision	Recall	Purity
ORL	SC1	0.807±0.004	0.663±0.012	0.535±0.013	0.546±0.012	0.519±0.017	0.576±0.010	0.692±0.009
	SC2	0.912±0.003	0.820±0.008	0.753±0.006	0.759±0.006	0.731±0.008	0.789±0.008	0.843±0.006
	SC3	0.869±0.008	0.774±0.014	0.666±0.018	0.673±0.017	0.651±0.019	0.698±0.016	0.789±0.016
	ConcatFea	0.808±0.007	0.663±0.018	0.536±0.017	0.547±0.016	0.523±0.017	0.572±0.015	0.691±0.014
	ConcatPCA	0.821±0.005	0.689±0.013	0.564±0.012	0.574±0.011	0.552±0.014	0.599±0.011	0.711±0.009
	MVGL [28]	0.894±0.000	0.735±0.000	0.545±0.000	0.558±0.000	0.428±0.000	0.801±0.000	0.795±0.000
	SNF [29]	0.905±0.006	0.784±0.026	0.712±0.027	0.719±0.027	0.662±0.038	0.787±0.012	0.827±0.016
	Co-Pair [30]	0.841±0.004	0.680±0.006	0.577±0.009	0.587±0.008	0.543±0.008	0.640±0.008	0.719±0.005
	Co-Reg [30]	0.872±0.003	0.724±0.007	0.642±0.008	0.651±0.008	0.603±0.008	0.707±0.008	0.761±0.005
	DiMSC [16]	0.901±0.014	0.817±0.026	0.764±0.035	0.770±0.034	0.733±0.035	0.811±0.036	0.846±0.022
	LMSC [21]	0.916±0.024	0.805±0.033	0.743±0.048	0.749±0.047	0.699±0.042	0.807±0.057	0.836±0.033
	LT-MSC [22]	0.917±0.011	0.808±0.027	0.753±0.029	0.759±0.028	0.716±0.034	0.807±0.023	0.840±0.019
	RAMSC [24]	0.895±0.009	0.786±0.016	0.706±0.019	0.713±0.018	0.675±0.020	0.757±0.018	0.816±0.014
	ECMSC [23]	0.877±0.013	0.721±0.028	0.631±0.033	0.640±0.032	0.576±0.033	0.720±0.032	0.766±0.026
	CSI($\lambda = 0$)	0.927±0.007	0.861±0.015	0.802±0.018	0.806±0.018	0.776±0.021	0.837±0.015	0.880±0.016
	CSI($\alpha = 0$)	0.895±0.008	0.746±0.014	0.667±0.020	0.675±0.020	0.606±0.028	0.763±0.015	0.796±0.013
	CSI	0.930±0.009	0.863±0.023	0.805±0.025	0.810±0.025	0.782±0.029	0.840±0.020	0.883±0.016
	KCSI	0.920±0.003	0.850±0.009	0.793±0.009	0.798±0.009	0.773±0.013	0.823±0.006	0.873±0.008
Yale	SC1	0.622±0.003	0.615±0.006	0.389±0.005	0.428±0.004	0.409±0.005	0.449±0.004	0.627±0.007
	SC2	0.673±0.008	0.664±0.005	0.462±0.009	0.497±0.008	0.464±0.007	0.534±0.013	0.664±0.005
	SC3	0.736±0.007	0.696±0.012	0.549±0.014	0.577±0.013	0.564±0.012	0.590±0.013	0.708±0.012
	ConcatFea	0.626±0.005	0.611±0.006	0.392±0.007	0.431±0.007	0.412±0.008	0.452±0.006	0.623±0.006
	ConcatPCA	0.603±0.005	0.588±0.004	0.361±0.007	0.402±0.006	0.386±0.007	0.419±0.005	0.599±0.003
	MVGL [28]	0.678±0.000	0.642±0.000	0.383±0.000	0.428±0.000	0.362±0.000	0.524±0.000	0.648±0.000
	SNF [29]	0.713±0.003	0.694±0.003	0.493±0.008	0.526±0.007	0.490±0.012	0.567±0.005	0.694±0.003
	Co-Pair [30]	0.643±0.008	0.595±0.013	0.423±0.010	0.460±0.009	0.436±0.010	0.488±0.009	0.605±0.013
	Co-Reg [30]	0.651±0.007	0.602±0.015	0.435±0.012	0.472±0.011	0.447±0.014	0.500±0.009	0.611±0.012
	DiMSC [16]	0.676±0.034	0.648±0.031	0.478±0.049	0.511±0.046	0.494±0.045	0.531±0.047	0.648±0.031
	LMSC [21]	0.706±0.008	0.673±0.018	0.483±0.015	0.517±0.014	0.481±0.019	0.559±0.009	0.678±0.017
	LT-MSC [22]	0.762±0.003	0.734±0.002	0.591±0.006	0.616±0.005	0.596±0.007	0.639±0.004	0.734±0.002
	RAMSC [24]	0.756±0.010	0.726±0.021	0.583±0.012	0.609±0.012	0.589±0.012	0.631±0.011	0.726±0.021
	ECMSC [23]	0.690±0.010	0.662±0.006	0.442±0.012	0.480±0.011	0.437±0.010	0.532±0.012	0.668±0.004
	CSI($\lambda = 0$)	0.764±0.003	0.733±0.000	0.582±0.005	0.608±0.005	0.582±0.006	0.638±0.003	0.739±0.000
	CSI($\alpha = 0$)	0.691±0.016	0.655±0.032	0.415±0.018	0.456±0.016	0.399±0.022	0.532±0.027	0.661±0.026
	CSI	0.777±0.000	0.734±0.000	0.606±0.000	0.631±0.000	0.606±0.000	0.659±0.000	0.739±0.000
	KCSI	0.749±0.000	0.727±0.000	0.573±0.000	0.600±0.000	0.575±0.000	0.627±0.000	0.727±0.000
3Sources	SC1	0.613±0.000	0.615±0.000	0.473±0.000	0.584±0.000	0.646±0.000	0.533±0.000	0.781±0.000
	SC2	0.562±0.018	0.549±0.031	0.392±0.037	0.518±0.027	0.582±0.046	0.468±0.019	0.730±0.024
	SC3	0.568±0.003	0.555±0.004	0.423±0.003	0.546±0.001	0.597±0.005	0.502±0.002	0.715±0.004
	ConcatFea	0.670±0.002	0.639±0.000	0.499±0.001	0.608±0.001	0.649±0.001	0.572±0.002	0.787±0.000
	ConcatPCA	0.637±0.001	0.682±0.003	0.513±0.003	0.611±0.003	0.702±0.002	0.541±0.002	0.775±0.000
	MVGL [28]	0.677±0.000	0.757±0.000	0.538±0.000	0.666±0.000	0.558±0.000	0.827±0.000	0.793±0.000
	SNF [29]	0.707±0.007	0.766±0.003	0.658±0.003	0.731±0.003	0.796±0.004	0.675±0.002	0.837±0.003
	Co-Pair [30]	0.614±0.007	0.573±0.005	0.444±0.009	0.558±0.007	0.635±0.009	0.499±0.006	0.762±0.006
	Co-Reg [30]	0.609±0.008	0.567±0.011	0.434±0.011	0.551±0.009	0.622±0.012	0.496±0.013	0.756±0.009
	DiMSC [16]	0.700±0.000	0.740±0.000	0.595±0.000	0.679±0.000	0.754±0.000	0.618±0.000	0.822±0.000
	LMSC [21]	0.648±0.038	0.633±0.061	0.487±0.062	0.587±0.053	0.692±0.042	0.511±0.058	0.733±0.033
	LT-MSC [22]	0.700±0.005	0.782±0.002	0.655±0.006	0.737±0.004	0.722±0.007	0.753±0.006	0.835±0.002
	RAMSC [24]	0.692±0.000	0.775±0.000	0.657±0.000	0.741±0.000	0.704±0.000	0.781±0.000	0.817±0.000
	ECMSC [23]	0.112±0.066	0.356±0.022	0.013±0.053	0.368±0.005	0.244±0.041	0.867±0.183	0.382±0.047
	CSI($\lambda = 0$)	0.747±0.000	0.763±0.000	0.622±0.000	0.709±0.000	0.717±0.000	0.702±0.000	0.834±0.000
	CSI($\alpha = 0$)	0.366±0.000	0.550±0.000	0.301±0.000	0.508±0.000	0.399±0.000	0.702±0.000	0.609±0.000
	CSI	0.729±0.000	0.757±0.000	0.608±0.000	0.699±0.000	0.702±0.000	0.696±0.000	0.828±0.000
	KCSI	0.747±0.000	0.763±0.000	0.622±0.000	0.709±0.000	0.717±0.000	0.702±0.000	0.834±0.000

values, 3304 dimensional LBP feature, and 6750 dimensional Gabor feature as 3 views for our experiments.

4.2. Evaluation on clustering performance

Tables 2 and 3 tabulates the results on six real-world datasets. We show the clustering results in terms of NMI, ACC, AR, F-score, Precision, Recall, and Purity. In Tables 2 and 3, “SC1” represents performing Spectral Clustering in the first view of the dataset, “SC2” represents performing Spectral Clustering in the second view of the dataset, and so on. For a fair comparison, each method was run ten times on the tested dataset and the average performance is recorded. The optimal and suboptimal results are highlighted in red and blue, respectively.

For the experiments on each dataset, we divided the tested method into two categories, concatenating the multi-view data directly and integrating them indirectly. The SNF, Co-Pair, Co-Reg, DiMSC, LMSC, LT-MSC, RAMSC, ECMSC and the proposed CSI

method fall into the second category, which is highlighted in gray background.

We firstly observed that ConcatFea and ConcatPCA got a worse result than single view spectral clustering on the MIT-8-scene, Caltech 101, ORL and Yale datasets, and performed better on the other datasets. This indicated that the performance of directly concatenating features is data-dependent. All the single view based methods performed uniformly less satisfactory than the multi-view based ones across all experiments. It suggests that adequate fusion of multi-view data will enhance its separability performance over simple augmentation of single-view data.

In testing the multi-view based methods, Co-Pair and Co-Reg achieved compromising results over different views, this is due to the aim of Co-Pair and Co-Reg at reducing the dissimilarity of every two views. MVGL, SNF, DiMSC, LMSC, LT-MSC, RAMSC and ECMSC achieved improvements over the single view based methods on most datasets. However, DiMSC did not perform well on the Caltech 101 and MIT-8-scene. The possible reason for such

Table 3
Clustering results on BBCSports, Caltech101-7 and MIT-8-scene.

Dataset	Method	NMI	ACC	AR	F-score	Precision	Recall	Purity
BBCSports	SC1	0.555±0.000	0.699±0.000	0.471±0.000	0.603±0.000	0.619±0.000	0.589±0.000	0.766±0.000
	SC2	0.625±0.000	0.784±0.000	0.633±0.000	0.729±0.000	0.703±0.000	0.756±0.000	0.848±0.000
	SC3	0.643±0.004	0.703±0.003	0.621±0.008	0.711±0.006	0.767±0.004	0.663±0.007	0.806±0.004
	ConcatFea	0.758±0.000	0.809±0.000	0.774±0.000	0.832±0.000	0.832±0.000	0.832±0.000	0.901±0.000
	ConcatPCA	0.796±0.000	0.830±0.000	0.803±0.000	0.853±0.000	0.859±0.000	0.848±0.000	0.901±0.000
	MVGL [28]	0.752±0.000	0.812±0.000	0.704±0.000	0.787±0.000	0.724±0.000	0.861±0.000	0.865±0.000
	SNF [29]	0.789±0.000	0.812±0.000	0.797±0.000	0.849±0.000	0.854±0.000	0.844±0.000	0.904±0.000
	Co-Pair [30]	0.601±0.014	0.680±0.013	0.538±0.017	0.647±0.013	0.706±0.015	0.597±0.012	0.783±0.011
	Co-Reg [30]	0.584±0.035	0.668±0.030	0.522±0.042	0.634±0.032	0.698±0.036	0.581±0.029	0.773±0.027
	DiMSC [16]	0.669±0.001	0.784±0.003	0.648±0.003	0.734±0.002	0.774±0.000	0.698±0.004	0.869±0.000
	LMSC [21]	0.632±0.022	0.733±0.033	0.606±0.042	0.699±0.033	0.763±0.031	0.646±0.044	0.827±0.021
	LT-MSC [22]	0.590±0.009	0.718±0.012	0.507±0.015	0.633±0.011	0.635±0.013	0.632±0.013	0.812±0.007
	RAMSC [24]	0.822±0.000	0.812±0.000	0.825±0.000	0.869±0.000	0.887±0.000	0.853±0.000	0.904±0.000
	ECMSC [23]	0.471±0.077	0.596±0.084	0.377±0.101	0.549±0.058	0.527±0.107	0.589±0.050	0.682±0.081
	CSI($\alpha = 0$)	0.801±0.002	0.812±0.000	0.815±0.002	0.862±0.001	0.868±0.003	0.856±0.000	0.911±0.000
	CSI($\alpha = 0$)	0.364±0.032	0.605±0.056	0.259±0.070	0.486±0.038	0.412±0.050	0.598±0.021	0.636±0.051
	CSI	0.823±0.005	0.920±0.006	0.856±0.005	0.892±0.004	0.928±0.002	0.858±0.007	0.920±0.006
	KCSI	0.842±0.000	0.933±0.000	0.868±0.000	0.900±0.000	0.934±0.000	0.869±0.000	0.933±0.000
Caltech101-7	SC1	0.369±0.008	0.507±0.011	0.298±0.007	0.481±0.002	0.714±0.024	0.362±0.009	0.791±0.000
	SC2	0.234±0.003	0.357±0.001	0.151±0.001	0.341±0.000	0.586±0.003	0.240±0.000	0.674±0.001
	SC3	0.573±0.000	0.690±0.000	0.473±0.000	0.649±0.000	0.748±0.000	0.573±0.000	0.807±0.000
	SC4	0.525±0.012	0.553±0.000	0.411±0.004	0.550±0.003	0.779±0.003	0.416±0.003	0.805±0.009
	SC5	0.540±0.000	0.562±0.000	0.400±0.000	0.561±0.000	0.805±0.000	0.430±0.000	0.844±0.000
	SC6	0.345±0.002	0.384±0.001	0.207±0.002	0.389±0.001	0.654±0.004	0.277±0.000	0.785±0.001
	ConcatFea	0.397±0.002	0.474±0.001	0.304±0.002	0.480±0.001	0.730±0.001	0.358±0.000	0.801±0.002
	ConcatPCA	0.398±0.002	0.468±0.002	0.302±0.001	0.478±0.000	0.730±0.001	0.355±0.000	0.801±0.003
	MVGL [28]	0.542±0.000	0.665±0.000	0.407±0.000	0.628±0.000	0.651±0.000	0.606±0.000	0.849±0.000
	SNF [29]	0.601±0.004	0.559±0.008	0.428±0.004	0.572±0.003	0.874±0.005	0.425±0.002	0.857±0.003
	Co-Pair [30]	0.419±0.005	0.399±0.007	0.256±0.005	0.416±0.005	0.744±0.006	0.289±0.004	0.782±0.004
	Co-Reg [30]	0.460±0.004	0.393±0.010	0.273±0.005	0.429±0.004	0.770±0.009	0.297±0.003	0.799±0.005
	DiMSC [16]	0.427±0.006	0.462±0.013	0.403±0.007	0.483±0.001	0.511±0.001	0.396±0.002	0.720±0.001
	LMSC [21]	0.570±0.027	0.525±0.027	0.400±0.036	0.547±0.030	0.859±0.031	0.402±0.026	0.865±0.016
	LT-MSC [22]	0.644±0.001	0.601±0.001	0.474±0.001	0.612±0.000	0.897±0.001	0.465±0.000	0.887±0.001
	RAMSC [24]	0.600±0.000	0.606±0.000	0.466±0.007	0.594±0.000	0.860±0.000	0.453±0.000	0.856±0.000
	ECMSC [23]	0.519±0.040	0.521±0.002	0.337±0.008	0.508±0.002	0.757±0.019	0.383±0.003	0.830±0.001
	CSI($\alpha = 0$)	0.588±0.011	0.643±0.004	0.481±0.008	0.626±0.006	0.866±0.006	0.490±0.006	0.840±0.009
	CSI($\alpha = 0$)	0.586±0.001	0.645±0.002	0.483±0.002	0.627±0.002	0.866±0.001	0.492±0.002	0.837±0.000
	CSI	0.609±0.000	0.693±0.000	0.521±0.000	0.676±0.000	0.808±0.000	0.580±0.000	0.858±0.000
	KCSI	0.617±0.000	0.703±0.000	0.538±0.000	0.687±0.000	0.821±0.000	0.590±0.000	0.859±0.000
MIT-8-scene	SC1	0.468±0.000	0.596±0.000	0.389±0.000	0.469±0.000	0.454±0.000	0.484±0.000	0.615±0.000
	SC2	0.488±0.000	0.576±0.000	0.336±0.000	0.427±0.000	0.396±0.000	0.465±0.000	0.579±0.000
	SC3	0.274±0.000	0.363±0.000	0.189±0.000	0.295±0.000	0.285±0.000	0.307±0.000	0.420±0.000
	SC4	0.525±0.012	0.553±0.000	0.411±0.004	0.550±0.003	0.779±0.003	0.416±0.003	0.805±0.009
	ConcatFea	0.274±0.000	0.363±0.000	0.189±0.000	0.295±0.000	0.285±0.000	0.307±0.000	0.420±0.000
	ConcatPCA	0.273±0.000	0.361±0.000	0.187±0.000	0.294±0.000	0.283±0.000	0.306±0.000	0.419±0.000
	MVGL [28]	0.479±0.000	0.463±0.000	0.201±0.000	0.352±0.000	0.236±0.000	0.697±0.000	0.464±0.000
	SNF [29]	0.584±0.000	0.655±0.000	0.504±0.000	0.571±0.000	0.539±0.000	0.607±0.000	0.677±0.000
	Co-Pair [30]	0.478±0.005	0.594±0.007	0.396±0.005	0.473±0.005	0.473±0.005	0.474±0.005	0.597±0.006
	Co-Reg [30]	0.456±0.007	0.584±0.017	0.378±0.008	0.457±0.007	0.457±0.007	0.458±0.006	0.601±0.011
	DiMSC [16]	0.146±0.001	0.307±0.001	0.110±0.001	0.227±0.001	0.219±0.001	0.237±0.001	0.323±0.001
	LMSC [21]	0.508±0.026	0.670±0.030	0.434±0.027	0.508±0.024	0.494±0.024	0.525±0.028	0.675±0.022
	LT-MSC [22]	0.596±0.001	0.713±0.000	0.494±0.000	0.559±0.000	0.549±0.000	0.569±0.001	0.713±0.000
	RAMSC [24]	0.490±0.001	0.634±0.000	0.404±0.001	0.479±0.001	0.482±0.001	0.477±0.001	0.634±0.000
	ECMSC [23]	0.560±0.000	0.661±0.000	0.463±0.000	0.533±0.000	0.520±0.000	0.546±0.000	0.671±0.000
	CSI($\alpha = 0$)	0.611±0.001	0.673±0.001	0.512±0.000	0.582±0.000	0.519±0.000	0.661±0.001	0.694±0.001
	CSI($\alpha = 0$)	0.552±0.000	0.540±0.000	0.409±0.000	0.492±0.000	0.444±0.000	0.552±0.000	0.589±0.000
	CSI	0.618±0.000	0.662±0.000	0.508±0.000	0.579±0.000	0.505±0.000	0.679±0.000	0.690±0.000
	KCSI	0.628±0.000	0.716±0.000	0.531±0.000	0.595±0.000	0.556±0.000	0.639±0.000	0.716±0.000

an unsatisfactory result may lie in that DiMSC is configured to learn complementary information from different views. However, if there exists large distortions among different views, DiMSC may be misled to have incorrect results. LT-MSC achieved all-right performances in most datasets except BBCSports. ECMSC also performed poorly on BBCSports and 3Sources. This may be because the multi-view features from the two datasets come from actual multiple sources, not like datasets such as ORL and Yale whose multi-view features are extracted artificially from the same object, namely, the multi-view features are lie in different sample spaces. The feature positions of each view are not aligned, therefore, the position-aware exclusivity term in ECMSC may lead to uncorrect clustering results. The RAMSC had good performance yet is inferior to CSI occasionally. The merits of RAMSC lie in its auto-weighting property so that it can make good use of information of all views. However, compared with CSI, the common subspace of RAMSC is learned directly from multi-view input data other than by CLR from

multi-view subspaces. Therefore, the affinity matrix by RAMSC is not as robust as by CSI. From the experimental results, one can find the performance of RAMSC was not stable. For example, RAMSC performed unsatisfactorily on the large dataset MIT-8-scene.

It can be seen that our proposed method CSI and KCSI outperformed uniformly over the single view based methods. It also achieved comparable or superior performances when compared to other multi-view methods. The special case was that the performance on 3Sources dataset was not so outstanding as on other datasets. The is mainly because of the incomplete views and partially missing patterns of 3Sources dataset, which is contrary to the view-consistency of CLR in CSI. Hence CSI can only achieve comparable performance on 3Sources dataset. On the whole, we can get the conclusion that the proposed CSI and KCSI are robust with respect to datasets. This owes to the graph constraint that preserves the local structure of the single view, and the constraint Laplacian rank in common space that guarantees the final clusters. The

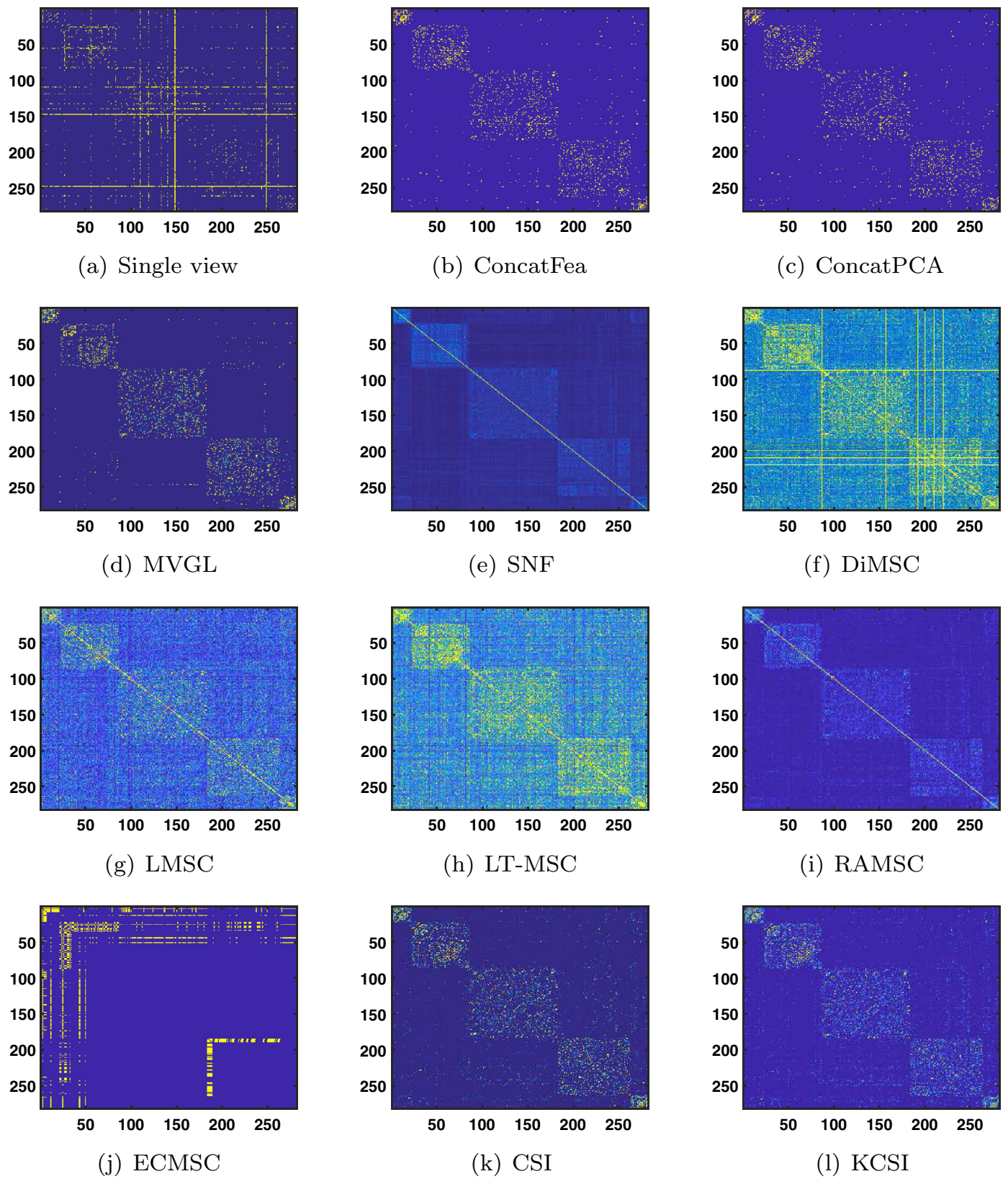


Fig. 2. Visualization of similarity matrices on the BBCSports dataset.

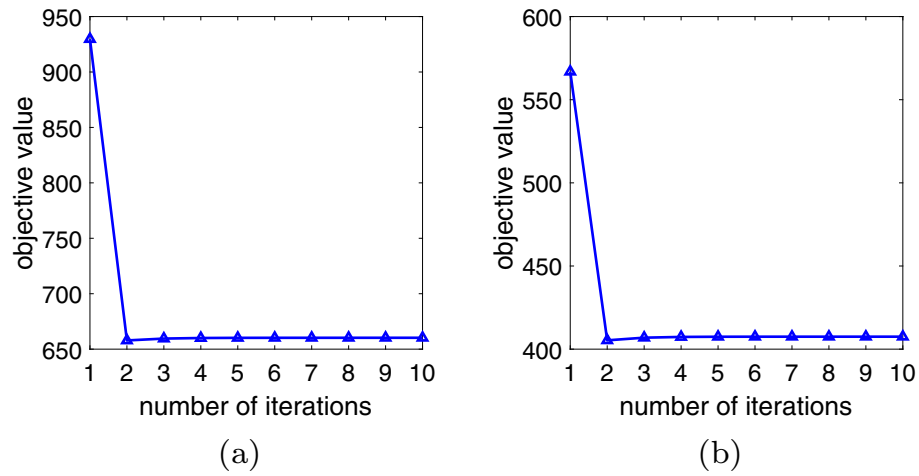


Fig. 3. The proposed CSI quickly converged to stable state with the fast decreasing curve of the objective value on (a) BBCSports; (b) 3Sources.

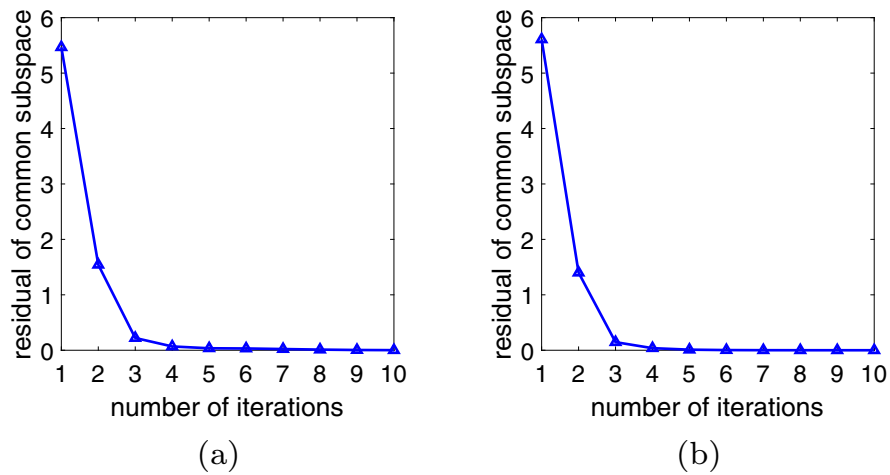


Fig. 4. The proposed CSI quickly converged to stable state with the fast decreasing curve of the residual of common subspace A on (a) BBCSports; (b) 3Sources.

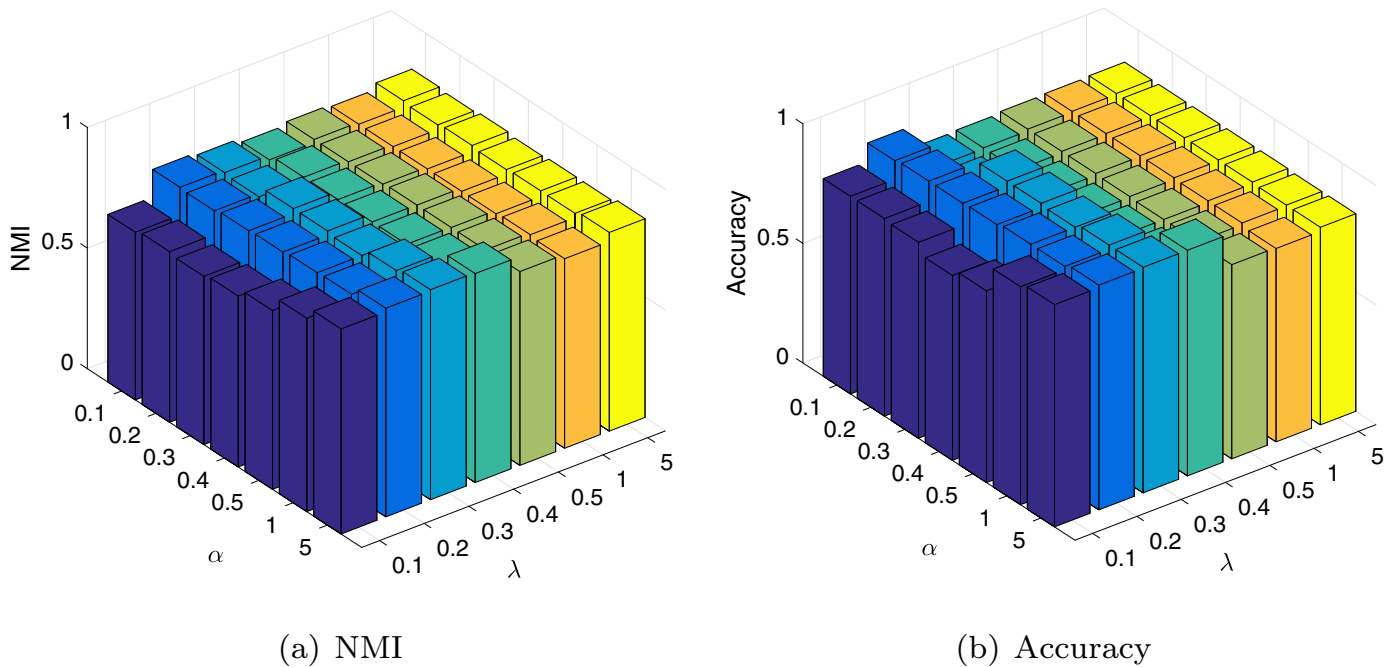


Fig. 5. Clustering performance on the BBCSports dataset w.r.t λ and α .

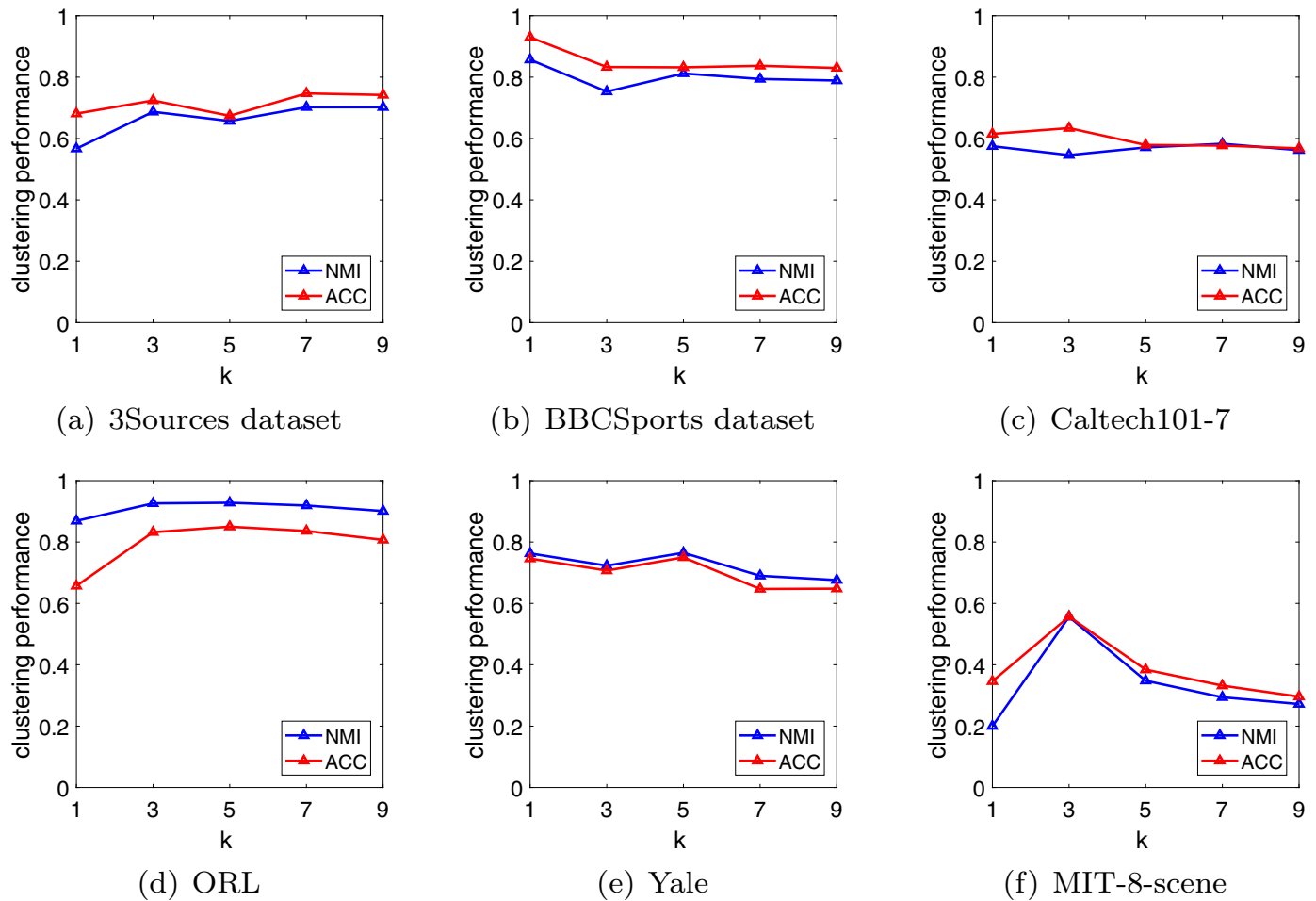


Fig. 6. The clustering performances with respect to different k on six datasets.

graph constraint and constraint Laplacian rank worked jointly to extract a separable common structure of data, thus achieving nice performances.

In particular to the dataset of BBCSports, we draw the obtained similarity matrices by testing the twelve methods in Fig. 2, the colors are scaled to $[0,0.1]$. Because Co-Pair and Co-Reg do not use similarity matrix to clustering, we did not draw the similarity matrices. One may find that our method can obtain an integrated common subspace with a better structure than the single view. There is one point needed to be emphasized that the SNF seems to led to a clearer diagonal-block structure than by CSI and KCSI. However, in fact, the quantitative measurements including NMI and ACC adequately demonstrate the clustering performance. From Table 3, one can find that the proposed CSI and KCSI achieves dramatically improvements around 5%, 13% over SNF in terms of NMI and ACC, respectively. The higher the NMI value, the closer the clustering result is to the real class distribution. ACC is a measure of statistical bias, and it refers to closeness of the clustering label to the ground truth label. In conclusion, the visualizations of CSI and KCSI are superior over other baselines.

It can also be seen that the CSI performed superior than KCSI on datasets of ORL and Yale, while performed inferior on the rest datasets. The reason is that CSI and KCSI are applicable to different scenarios. The kernel version of CSI is prone to process more complex datasets. KCSI provides a mapping for the input multi-view data. Hence the typical issues in multi-view learning such as different feature dimensions and different sample spaces

can be alleviated. Nevertheless, both CSI and KCSI ranked in the top two among all the methods. It demonstrated the separability of common subspace integration by considering both global similarities and local structures.

4.3. Convergence and parameter analysis

We first analyze the convergence property of Algorithm 1. During each iteration, a convex minimization problem is solved for each Z -step, P -step and A -step, and the global optimum solution is obtained for each sub-step. Thus the overall objective is non-increasing with the iterations and Algorithm 1 is guaranteed to convergence.

To verify the convergence property of the proposed method, Figs. 3 and 4 show the speed of convergence on two datasets. In each subfigure of Fig. 3, the x-axis and the y-axis denote the number of iterations and the corresponding objective value, respectively. We see that the objective value decreases sharply within ten iterations, and then becomes steady. Another figure can further verify the convergence property. Fig. 4 shows the variation trend of the residual of common subspace A with the number of iterations. This is in accord with the setting of convergence in our experiments. We used the residual of common subspace to terminate the iteration when it is lower than a tiny threshold of 10^{-4} . It can be easily observed from the Fig. 4 that, as the iteration increasing, the residual of common subspace is almost zero and the decreas-

Table 4

Performances of ablation study on six real-world bench mark datasets. The methods with ablated term are highlighted in gray background. Each method was run ten times on the tested datasets. Their mean and standard deviation value were recorded for performance comparing. The optimal results were highlighted in bold.

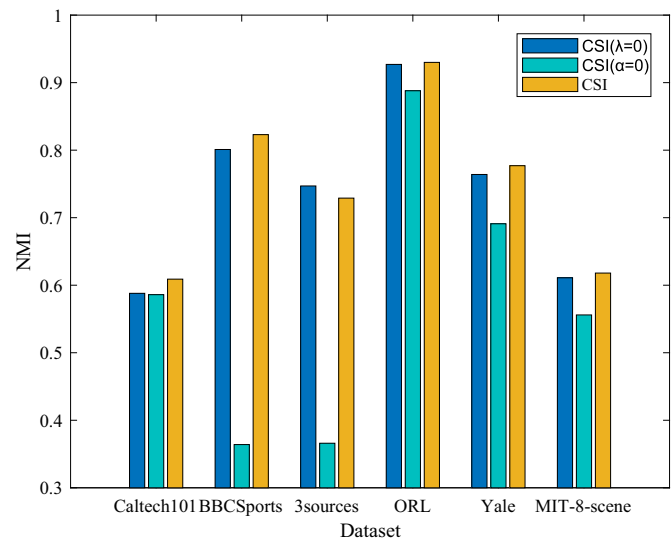
Dataset	Method	NMI	ACC	AR	F-score	Precision	Recall	Purity
Caltech101-7	CSI($\lambda = 0$)	0.588 ± 0.011	0.643 ± 0.004	0.481 ± 0.008	0.626 ± 0.006	0.866 ± 0.006	0.490 ± 0.006	0.840 ± 0.009
	CSI($\alpha = 0$)	0.586 ± 0.001	0.645 ± 0.002	0.483 ± 0.002	0.627 ± 0.002	0.866 ± 0.001	0.492 ± 0.002	0.837 ± 0.000
	CSI	0.609 ± 0.000	0.693 ± 0.000	0.521 ± 0.000	0.676 ± 0.000	0.808 ± 0.000	0.580 ± 0.000	0.858 ± 0.000
BBCSports	CSI($\lambda = 0$)	0.801 ± 0.002	0.812 ± 0.000	0.815 ± 0.002	0.862 ± 0.001	0.868 ± 0.003	0.856 ± 0.000	0.911 ± 0.000
	CSI($\alpha = 0$)	0.364 ± 0.032	0.605 ± 0.056	0.259 ± 0.070	0.486 ± 0.038	0.412 ± 0.050	0.598 ± 0.021	0.636 ± 0.051
	CSI	0.823 ± 0.005	0.920 ± 0.006	0.856 ± 0.005	0.892 ± 0.004	0.928 ± 0.002	0.858 ± 0.007	0.920 ± 0.006
3Sources	CSI($\lambda = 0$)	0.747 ± 0.000	0.763 ± 0.000	0.622 ± 0.000	0.709 ± 0.000	0.717 ± 0.000	0.702 ± 0.000	0.834 ± 0.000
	CSI($\alpha = 0$)	0.366 ± 0.000	0.550 ± 0.000	0.301 ± 0.000	0.508 ± 0.000	0.399 ± 0.000	0.702 ± 0.000	0.609 ± 0.000
	CSI	0.729 ± 0.000	0.757 ± 0.000	0.608 ± 0.000	0.699 ± 0.000	0.702 ± 0.000	0.696 ± 0.000	0.828 ± 0.000
ORL	CSI($\lambda = 0$)	0.927 ± 0.007	0.861 ± 0.015	0.802 ± 0.018	0.806 ± 0.018	0.776 ± 0.021	0.837 ± 0.015	0.880 ± 0.016
	CSI($\alpha = 0$)	0.895 ± 0.008	0.746 ± 0.014	0.667 ± 0.020	0.675 ± 0.020	0.606 ± 0.028	0.763 ± 0.015	0.796 ± 0.013
	CSI	0.930 ± 0.009	0.863 ± 0.023	0.805 ± 0.025	0.810 ± 0.025	0.782 ± 0.029	0.840 ± 0.020	0.883 ± 0.016
Yale	CSI($\lambda = 0$)	0.764 ± 0.003	0.733 ± 0.000	0.582 ± 0.005	0.608 ± 0.005	0.582 ± 0.006	0.638 ± 0.003	0.739 ± 0.000
	CSI($\alpha = 0$)	0.691 ± 0.016	0.655 ± 0.032	0.415 ± 0.018	0.456 ± 0.016	0.399 ± 0.022	0.532 ± 0.027	0.661 ± 0.026
	CSI	0.777 ± 0.000	0.734 ± 0.000	0.606 ± 0.000	0.631 ± 0.000	0.606 ± 0.000	0.659 ± 0.000	0.739 ± 0.000
MIT-8-scene	CSI($\lambda = 0$)	0.611 ± 0.001	0.673 ± 0.001	0.512 ± 0.000	0.582 ± 0.000	0.519 ± 0.000	0.661 ± 0.001	0.694 ± 0.001
	CSI($\alpha = 0$)	0.552 ± 0.000	0.540 ± 0.000	0.409 ± 0.000	0.492 ± 0.000	0.444 ± 0.000	0.552 ± 0.000	0.589 ± 0.000
	CSI	0.618 ± 0.000	0.662 ± 0.000	0.508 ± 0.000	0.579 ± 0.000	0.505 ± 0.000	0.679 ± 0.000	0.690 ± 0.000

ing trend of it is becoming steady. These two figures indicate that the proposed method converges sufficiently well.

To investigate the influence of parameter (λ , α) in CSI, we plotted the clustering performances with respect to the two hyper-parameters (λ , α) on BBCSports. The results are shown in Fig. 5. We choose 0.3, 0.3 as the value of parameter λ and α on BBCSports, respectively. One may observe that CSI is insensitive to parameters pruning.

We also provided an experiment to analyze and demonstrate the effect of k . The quantitative measurements on clustering performances with respect to the number of nearest neighbors k on six real-world bench mark datasets are plotted in Fig. 6. The blue and the red curves are used to denote the values of NMI and Accuracy, respectively. The value of k take ranges of [1, 3, 5, 7, 9]. Due to the nonlinear mapping by the Gaussian scaled distance, the two values of NMI and Accuracy are rather stable with respect to different k , as is shown in Fig. 6. The clustering performance of CSI varies smoothly on the six real datasets. As rule of thumb, we set the value of k to be 5.

To verify the effectiveness of our proposed method, we also conducted experiments to compare the different clustering performances when $\lambda = 0$ ($\alpha \neq 0$), $\alpha = 0$ ($\lambda \neq 0$), and $\lambda, \alpha \neq 0$ in CSI, respectively. The results are supplemented in Table 4 and visualized in Fig. 7. We can see that CSI has better performance than CSI($\lambda = 0$) and CSI($\alpha = 0$) on most datasets. In the special case, such as 3Sources dataset, the evaluation metrics value of CSI($\lambda = 0$) is slightly higher than CSI and is equal to KCSI. The main reason is that the 3Sources dataset has incomplete views and partially missing patterns. Nevertheless, the CLR term controlled by parameter λ is under the assumption of view-consistency. On the account, when $\lambda = 0$, the clustering performance was slightly improved. On the MIT-8-scene dataset, some of the evaluation metrics values of CSI($\lambda = 0$) is slightly higher, the possible reason may be that the dimensions of multi-view features vary greatly. When we use CLR to learn the common subspace from such multiple subspaces, the new common subspace may not always capture the features fully, thus resulting in the unstable performance compared with not using CLR. Another parameter α controls local affinity. From the comparison, we can obviously recognize the significance of the regularization term controlled by α . Owing to the two regularization terms, our method possesses both global and local similarity properties. The experimental results of ablation study can verify the effectiveness of our method.

**Fig. 7.** Ablation studies on six real-world bench mark datasets.

5. Conclusion

To enhance multi-view clustering performance, in this paper, we propose to combine graph regularized subspace learning with constrained Laplacian rank method to explore a common subspace, which has the exact number of connected components. The proposed approach has three advantages: (1) CSI leverages the key features of every single view and combines them during subspace learning to achieve good use of multiple views; (2) CSI learns common subspaces by minimizing the difference between global and local structures to extract the global commonality from each view, thus achieving excellent separability; (3) CSI can be easily generalized to kernel space, thus it can be popularized for many occasions. The experimental results on the six real-world datasets demonstrate the superiority of the proposed methods compared to the state-of-art methods.

Declaration of Competing Interest

None.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (61771007), Science and Technology Planning Project of Guangdong Province (2016A010101013, 2017B020226004), the Health & Medical Collaborative Innovation Project of Guangzhou City (201803010021), and the Fundamental Research Fund for the Central Universities (2017ZD051).

References

- [1] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: recent progress and new challenges, *Inf. Fusion* 38 (2017) 43–54.
- [2] R. Qi, A. Ma, Q. Ma, Q. Zou, Clustering and classification methods for single-cell RNA-sequencing data, *Brief. Bioinform.* 00(00) (2019) 1–13.
- [3] A. Kar, C. Häne, J. Malik, Learning a multi-view stereo machine, in: *Advances in Neural Information Processing Systems*, 2017, pp. 365–376.
- [4] C. Zhang, H. Fu, Q. Hu, X. Cao, Y. Xie, D. Tao, D. Xu, Generalized latent multi-view subspace clustering, *IEEE Trans. Pattern Anal. Mach. Intell.* (2018), doi:10.1109/TPAMI.2018.2877660.
- [5] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: *Proceedings of the 17th Conference on Computer Learning Theory*, 1998, pp. 92–100.
- [6] S. Sun, A survey of multi-view machine learning, *Neural Comput. Appl.* 23 (7–8) (2013) 2031–2038.
- [7] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, M. Sugiyama, Co-teaching: robust training of deep neural networks with extremely noisy labels, in: *Advances in Neural Information Processing Systems*, 2018, pp. 8536–8546.
- [8] W. Zhan, M.-L. Zhang, Inductive semi-supervised multi-label learning with co-training, in: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2017, pp. 1305–1314.
- [9] S. Yu, B. Krishnapuram, R. Rosales, R.B. Rao, Bayesian co-training, *J. Mach. Learn. Res.* 12 (3) (2011) 2649–2680.
- [10] W. Wang, Z.H. Zhou, A new analysis of co-training, in: *Proceedings of the International Conference on International Conference on Machine Learning*, 2010, pp. 1135–1142.
- [11] J. Du, C.X. Ling, Z.-H. Zhou, When does cotraining work in real data? *IEEE Trans. Knowl. Data Eng.* 23 (5) (2011) 788–799.
- [12] B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Q. Chen, K.-C. Chou, Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection, *Bioinformatics* 30 (4) (2013) 472–479.
- [13] D. Ramazzotti, A. Lal, B. Wang, S. Batzoglou, A. Sidow, Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival, *Nat. Commun.* 9 (1) (2018) 4453.
- [14] W. Li, F. Abtahi, Z. Zhu, A deep feature based multi-kernel learning approach for video emotion recognition, in: *Proceedings of the ACM on International Conference on Multimodal Interaction*, ACM, 2015, pp. 483–490.
- [15] S.S. Bucak, R. Jin, A.K. Jain, Multiple kernel learning for visual object recognition: a review, *IEEE Trans. Pattern Anal. Mach. Intell.* 36 (7) (2014) 1354–1369.
- [16] X. Cao, C. Zhang, H. Fu, S. Liu, H. Zhang, Diversity-induced multi-view subspace clustering, in: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, IEEE, 2015, pp. 586–594.
- [17] H. Gao, F. Nie, X. Li, H. Huang, Multi-view subspace clustering, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4238–4246.
- [18] Z. Ding, Y. Fu, Robust multi-view subspace learning through dual low-rank decompositions, in: *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016, pp. 1181–1187.
- [19] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, X. Huang, Robust subspace clustering for multi-view data by exploiting correlation consensus, *IEEE Trans. Image Process.* 24 (11) (2015) 3939–3949.
- [20] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (1) (2016) 188–194.
- [21] C. Zhang, Q. Hu, H. Fu, P. Zhu, X. Cao, Latent multi-view subspace clustering, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4279–4287.
- [22] C. Zhang, H. Fu, S. Liu, G. Liu, X. Cao, Low-rank tensor constrained multiview subspace clustering, in: *Proceedings of the IEEE international conference on Computer Vision*, 2015, pp. 1582–1590.
- [23] X. Wang, X. Guo, Z. Lei, C. Zhang, S.Z. Li, Exclusivity-consistency regularized multi-view subspace clustering, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 923–931.
- [24] W. Zhuge, C. Hou, Y. Jiao, J. Yue, H. Tao, D. Yi, Robust auto-weighted multi-view subspace clustering with common subspace representation matrix, *PLoS One* 12 (5) (2017) e0176769.
- [25] B. Mohar, Y. Alavi, G. Chartrand, O.R. Oellermann, A.J. Schwenk, The laplacian spectrum of graphs, in: *Proceedings of the Graph Theory, Combinatorics, and Applications*, 2001, pp. 871–898.
- [26] K. Fan, On a theorem of Weyl concerning eigenvalues of linear transformations i, *Proc. Natl. Acad. Sci.* 35 (11) (1949) 652–655.
- [27] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: *Proceedings of the 14th Advance Neural Information Processing System*, 2002, pp. 849–856.
- [28] K. Zhan, C. Zhang, J. Guan, J. Wang, Graph learning for multiview clustering, *IEEE Trans. Cybern.* PP (99) (2017) 1–9.
- [29] B. Wang, A.M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale, *Nat. Methods* 11 (3) (2014) 333.
- [30] A. Kumar, P. Rai, H. Daume, Co-regularized multi-view spectral clustering, in: *Proceedings of the 24th Advance Neural Information Processing System*, 2011, pp. 1413–1421.
- [31] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, *Int. J. Comput. Vis.* 42 (3) (2001) 145–175.
- [32] X. Aodan, C. Jiazhou, P. Hong, H. Guoqiang, C. Hongmin, Simultaneous Interrogation of Cancer Omics to Identify Subtypes With Significant Clinical Differences, *Front. Genet.* 10 (2019) 236.
- [33] C. Jiazhou, P. Hong, H. Guoqiang, C. Hongmin, HOGMMNC: a higher order graph matching with multiple network constraints model for gene–drug regulatory modules identification, *Bioinform.* 35 (4) (2019) 602–610.



Wanlin Weng is currently pursuing the Graduate degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. She received the B.S. degree in digital media technology from Shandong University, Weihai, China. Her current research interests include machine learning and pattern recognition.



Weiwei Zhou is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He received the B.S. degree in mathematics and applied mathematics from China University of Petroleum, Tsingtao, China. His current research interests include machine learning and pattern recognition.



Jiazhou Chen is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He received the B.S. degree in computer science and technology from Jiangyin University, Meizhou, China and M.S. degree in software engineering from Guangdong University of Technology, Guangzhou, China. His current research interests include bioinformatics, machine learning, and image processing.



Hong Peng is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. She received the B.S. degree in mathematics and applied mathematics from Jiangsu University, Zhenjiang, China. Her current research interests include machine learning and image processing.



Hongmin Cai is a Professor at the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He received the B.S. and M.S. degrees in mathematics from the Harbin Institute of Technology, Harbin, China, in 2001 and 2003, respectively, and the Ph.D. degree in applied mathematics from Hong Kong University in 2007. His areas of research interests include biomedical image processing, machine learning and omics data integration.