Graph-regularized concept factorization for multi-view document clustering

Kun Zhan[a], Jinhui Shi[a], Jing Wang[b], Feng Tian[b]

[a]*School of Information Science and Engineering, Lanzhou University, Lanzhou, China*
[b]*Faculty of Science and Technology, Bournemouth University, UK*

Abstract

We propose a novel multi-view document clustering method with the graph-regularized concept factorization (MVCF). MVCF makes full use of multi-view features for more comprehensive understanding of the data and learns weights for each view adaptively. It also preserves the local geometrical structure of the manifolds for multi-view clustering. We have derived an efficient optimization algorithm to solve the objective function of MVCF and proven its convergence by utilizing the auxiliary function method. Experiments carried out on three benchmark datasets have demonstrated the effectiveness of MVCF in comparison to several state-of-the-art approaches in terms of accuracy, normalized mutual information and purity.

Keywords: Multi-view learning, concept factorization, document clustering, manifold learning

1. Introduction

The matrix factorization-based approaches have become popular in document clustering [1, 2]. Nonnegative matrix factorization (NMF) [3] and concept factorization (CF) [1] have produced impressive results. Generally, CF mainly strives to overcome the limitations of NMF while inheriting all its strengths. CF models each concept as a linear combination of the data points, and each data point as a linear combination of the concepts. It aims to interpret the product of the two sets of linear coefficients as an approximation of the original data points. The cluster label of each data point can be easily derived from the obtained linear coefficients. However, CF does not consider the local manifold geometry but the global Euclidean geometry only. Cai *et al.* proposed LCCF [4] which uses a graph regularization term to capture the local geometry of the document sub-manifold. As data are often sparse, Liu *et al.* enforces a locality constraint onto CF and proposed LCF [5] to achieve sparsity and locality simultaneously. Taking the advantage of semi-supervised learning, Liu *et al.* incorporated additional information into CF and proposed CCF [6], which ensures the data points sharing the same label be grouped together. However, CCF is not able to handle the data points with different labels, which should be grouped into different clusters to the maximum extent. To achieve this goal, He *et al.* proposed PCCF [7] to group data which are of must-link in the same cluster and of cannot-link in the different clusters.

Essentially, all CF methods mentioned above are developed to handle a s-ingle view (feature). However, in many real world applications, data are often collected from diverse domains or obtained from different feature extractors [8–10]. For example, a document may be translated into multiple languages, a web page may be presented by its contents or a hyperlink, and a user may use heterogenous social networks. In these examples, each view alone would be insufficient for clustering without complementary information from the other views. Some methods have been proposed to deal with this situation. Bickel *et al.* proposed an Co-EM based framework [9] for multi-view clustering in mixture models. It computes expected values of hidden variables in one view and uses the values in the M-step for other views, and vice versa. This process is repeated until a suitable stopping criteria is met. The algorithm however often fails to converge. Relying on the eigen-decomposition technique, the spectral clustering methods can guarantee a global optima thus achieves better clustering performances. Kumar *et al.* proposed a multi-view spectral clustering method, CRSC [11], to ensure that corresponding data points in each view have the same cluster membership. Later, Xia *et al.* proposed RMSC [12] that explicitly handles possible noises in the multi-view input data and recovers a shared transition probability matrix via low rank and sparse decomposition. However, the limitation of the spectral clustering methods is that the negative values appearing in eigen-factorization make the factorization hard to interpret and that the obtained eigen-vectors have no direct relationship with the semantic structure of dataset [13, 14].

Recently, NMF-based multi-view clustering has received great attentions due to its better semantic interpretation [13, 14]. Liu *et al.* proposed Multi-NMF [15] to obtain a common consensus matrix, which is designed to reflect the latent clustering structure shared by different views. However, Multi-

NMF fails to preserve the locally geometrical structure of the data space. To tackle this problem, Zhang *et al.* proposed MMNMF [16]. However, the method needs to assign weights for each view individually, and it is often non-trivial to decide these weights. A graph-regularized NMF-based multi-view clustering method [17] was proposed, which extends the Liu *et al.*'s algorithm [15] with a graph regularization. In this study we take the advantage of CF and propose a novel method, called multi-view CF (MVCF). The overall approach and advantages of MVCF are as follows:

1. MVCF finds intrinsic coefficient matrices for each view, and then incorporates them with a multi-manifold regularizer to preserve the locally geometrical structure of the multi-view data space.
2. MVCF learns the weights of each view automatically. This saves the cost of setting weights individually and truly reflects the importance of each view.
3. A new updating rule is developed to efficiently and effectively solve the associated optimization problem. The proof of convergence of the rule is also provided.

The outline of the rest of the paper is as follows. In the section 2, we briefly review CF and local invariance, and then propose our MVCF. In the section 3, the optimization algorithm for solving the objective function of MVCF is proposed with proof of convergence. The experimental results on document datasets are discussed in the section 4. Finally we draw the conclusion and future work in the section 5.

2. Multi-view concept factorization

2.1. Concept factorization

Given n data $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$, each data $\boldsymbol{x}_i$ is represented by a d-dimensional feature vector. NMF aims to find a $d \times k$ matrix $\mathbf{U}$ and a $n \times k$ matrix $\mathbf{H}$ where the product of these two factors is an approximation to the original matrix, represented as $\mathbf{X} \approx \mathbf{U}\mathbf{H}^{\mathrm{T}}$. Each column vector of $\mathbf{U}$, $\boldsymbol{u}_c$, can be regarded as a basis and each data point $\boldsymbol{x}_i$ is approximated by a linear combination of these k bases, weighted by the components of data representation matrix $\mathbf{H}^{\mathrm{T}} = [h_{ic}]$: $\boldsymbol{x}_i \approx \sum_{c=1}^{k} \boldsymbol{u}_c h_{ic}$.

The speciality of NMF is that it enforces that all entries of the factor matrices must be non-negative. This brings two limitations. One is that the non-negative requirement is not applicable to applications where the data involves negative numbers. The other is that it is not clear how to effectively perform NMF in the transformed data space so that the powerful kernel method can be applied.

Concept Factorization (CF) is proposed to address the above problems while inheriting NMF's strengthes. CF models each base (cluster center) $\boldsymbol{u}_c$ by a linear combination of the data points $\boldsymbol{u}_c = \sum_{i=1}^{n} w_{ic}\boldsymbol{x}_i$ where $w_{ic} \geq 0$. Let $\mathbf{W} = [w_{ic}] \in \mathbb{R}^{n\times k}$, CF tries to decompose the data matrix which satisfies the following condition

$$\mathbf{X} \approx \mathbf{U}\mathbf{H}^{\mathrm{T}} = \mathbf{X}\mathbf{W}\mathbf{H}^{\mathrm{T}}. \tag{1}$$

The basic form of CF utilizes Frobenius norm to qualify the approximation, and CF tries to optimize the following problem,

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} \|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{H}^{\mathrm{T}}\|_{\mathrm{F}}^2. \tag{2}$$

To make the solution of (2) unique [1], we require that

$$\boldsymbol{w}_c^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}_c = 1. \tag{3}$$

This requirement of normalizing $\mathbf{W}$ is given by,

$$\mathbf{W} \leftarrow \mathbf{W}[\mathrm{diag}(\mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{W})]^{-\frac{1}{2}}. \tag{4}$$

Accordingly, $\mathbf{H}$ is adjusted so that $\mathbf{W}\mathbf{H}^{\mathrm{T}}$ does no change. This is achieved by,

$$\mathbf{H} \leftarrow \mathbf{H}[\mathrm{diag}(\mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{W})]^{\frac{1}{2}}. \tag{5}$$

2.2. Local Invariance

If two data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are spatially close in the intrinsic geometry of the data distribution, the corresponding data representations $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ are also close to each other [18–21], which is called the local invariance. Recent studies on the spectral graph theory [22] and the manifold learning theory [23] have demonstrated that the local geometric structure can be effectively modeled through a nearest neighbor graph on a scattering of data points. Considering a graph with n vertices where each vertex corresponds to a document in the corpus, we can define the edge weight matrix $\mathbf{S}$ as follows:

$$(\mathbf{S})_{ij} = \begin{cases} \frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j}{\|\boldsymbol{x}_i\|\|\boldsymbol{x}_j\|}, & \text{if } \ \boldsymbol{x}_i \in N_p(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in N_p(\boldsymbol{x}_i), \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $N_p(\boldsymbol{x}_i)$ denotes a set of p-nearest neighbors of $\boldsymbol{x}_i$ [4].

If two data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close to each other, $(\mathbf{S})_{ij}$ is set approximately to one. Then, the new low-dimensional representations $\mathbf{H}$ can be obtained by minimizing the following term [18],

$$\begin{aligned} &\min_{\mathbf{H}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|\boldsymbol{h}_i - \boldsymbol{h}_j\|^2 (\mathbf{S})_{ij} \\ &= \min_{\mathbf{H}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Tr}\left[(\boldsymbol{h}_i - \boldsymbol{h}_j)(\boldsymbol{h}_i - \boldsymbol{h}_j)^{\mathrm{T}}(\mathbf{S})_{ij}\right] \\ &= \min_{\mathbf{H}} \mathrm{Tr}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\boldsymbol{h}_i(\mathbf{S})_{ij}\boldsymbol{h}_i^{\mathrm{T}} - \boldsymbol{h}_i(\mathbf{S})_{ij}\boldsymbol{h}_j^{\mathrm{T}}\right]\right\} \\ &= \min_{\mathbf{H}} \mathrm{Tr}[\mathbf{H}^{\mathrm{T}}(\mathbf{D} - \mathbf{S})\mathbf{H}] \\ &= \min_{\mathbf{H}} \mathrm{Tr}(\mathbf{H}^{\mathrm{T}}\mathbf{L}\mathbf{H}), \end{aligned} \tag{7}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is the graph Laplacian [22], $(\mathbf{D})_{ii} = \sum_j (\mathbf{S})_{ij}$ and $\mathrm{Tr}(\cdot)$ denotes the trace operator.

2.3. Objective Function

Let $\mathbf{X}^v \in \mathbb{R}^{d_v \times n}$ denote the features in v^{th} view, $\mathbf{W}^v \in \mathbb{R}^{n \times k}$ and $\mathbf{H}^v \in \mathbb{R}^{n \times k}$ be the coefficient matrices in v^{th} view, respectively. Given n_v types of heterogeneous features, $v = 1, 2, \cdots, n_v$, we integrate all these views together with the combination of the problems (2) and (7), and have the following objective function:

$$\begin{aligned} \min_{\mathbf{W}^v, \mathbf{H}^v} &\sum_{v=1}^{n_v}\{\alpha^v\|\mathbf{X}^v - \mathbf{X}^v\mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}\|_{\mathrm{F}}^2 + \\ &\quad \mathrm{Tr}[(\mathbf{H}^v)^{\mathrm{T}}\mathbf{L}^v\mathbf{H}^v]\} \\ \text{s.t. } \ &\forall v, \mathbf{W}^v \geq 0, \mathbf{H}^v \geq 0, \alpha^v \geq 0, \end{aligned} \tag{8}$$

where α^v is the weight of v^{th} view which represents the importance of the view and needs to be set separately.

Apparently, it is hard to specify the weights α^v for (8) without prior knowledge.

However, if we add a regularization term to (8) as,

$$\begin{aligned} &\min_{\alpha^v} \sum_{v=1}^{n_v} (\alpha^v)^2 \\ &\text{s.t.} \;\; \forall v, \alpha^v \geq 0, \sum_{v=1}^{n_v} \alpha^v = 1, \end{aligned} \tag{9}$$

the weights of each view can be learned adaptively, which reflects the importance of the corresponding views. Besides, the term in (9) helps avoid the situation that one view's weight may be learned and assigned to be one while the weights of other views are all zero.

Considering α^v is the parameter that controls the two terms in (8) and not limited to a fixed range, *i.e.*, $\sum_{v=1}^{n_v} \alpha^v = 1$, We extend the range to a constant λ and have

$$\begin{aligned} &\min_{\alpha^v} \sum_{v=1}^{n_v} (\alpha^v)^2 \\ &\text{s.t.} \;\; \forall v, \alpha^v \geq 0, \sum_{v=1}^{n_v} \alpha^v = \lambda. \end{aligned} \tag{10}$$

Combining (8) and (10), we propose the final objective function:

$$\begin{aligned} \mathcal{O} = \min_{\mathbf{W}^v, \mathbf{H}^v, \alpha^v} &\sum_{v=1}^{n_v} \{\alpha^v \|\mathbf{X}^v - \mathbf{X}^v \mathbf{W}^v (\mathbf{H}^v)^{\mathrm{T}}\|_{\mathrm{F}}^2 + \\ &\mathrm{Tr}[(\mathbf{H}^v)^{\mathrm{T}} \mathbf{L}^v \mathbf{H}^v]\} + \gamma \sum_{v=1}^{n_v} (\alpha^v)^2 \\ \text{s.t.} \;\; &\forall v, \mathbf{W}^v \geq 0, \mathbf{H}^v \geq 0, \alpha^v \geq 0, \sum_{v=1}^{n_v} \alpha^v = \lambda, \end{aligned} \tag{11}$$

where γ is a parameter of the last term.

In the following section, we describe a novel updating rule to obtain the local optima for solving the objective function in (11). The rule guarantees the objective function is non-increasing with each iteration.

3. Optimization

3.1. Algorithm Derivation

We optimize (11) with the following two steps.

The first step is to fix α^v, and update $\mathbf{W}^v$ and $\mathbf{H}^v$ for each view independently. Then, (11) becomes,

$$\begin{aligned}\mathcal{O}_1 = \min_{\mathbf{W}^v,\mathbf{H}^v} & \ \alpha^v \|\mathbf{X}^v - \mathbf{X}^v\mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}\|_{\mathrm{F}}^2 \\ & + \mathrm{Tr}[(\mathbf{H}^v)^{\mathrm{T}}\mathbf{L}^v\mathbf{H}^v] \\ \text{s.t.} & \ \ \mathbf{W}^v \geq 0, \mathbf{H}^v \geq 0.\end{aligned} \tag{12}$$

Defining $\mathbf{K}^v = (\mathbf{X}^v)^{\mathrm{T}}\mathbf{X}^v$, we can rewrite the objective function (12):

$$\begin{aligned}& \alpha^v \|\mathbf{X}^v - \mathbf{X}^v\mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}\|_{\mathrm{F}}^2 + \mathrm{Tr}[(\mathbf{H}^v)^{\mathrm{T}}\mathbf{L}^v\mathbf{H}^v] \\ = & \ \alpha^v \mathrm{Tr}\{[\mathbf{X}^v - \mathbf{X}^v\mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}]^{\mathrm{T}}[\mathbf{X}^v - \mathbf{X}^v\mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}]\} \\ & + \mathrm{Tr}[(\mathbf{H}^v)^{\mathrm{T}}\mathbf{L}^v\mathbf{H}^v] \\ = & \ \alpha^v \mathrm{Tr}\{[\mathbf{I} - \mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}]^{\mathrm{T}}\mathbf{K}^v[\mathbf{I} - \mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}]\} \\ & + \mathrm{Tr}[(\mathbf{H}^v)^{\mathrm{T}}\mathbf{L}^v\mathbf{H}^v] \\ = & \ \alpha^v\{\mathrm{Tr}(\mathbf{K}^v) - 2\mathrm{Tr}[\mathbf{H}^v(\mathbf{W}^v)^{\mathrm{T}}\mathbf{K}^v] \\ & + \mathrm{Tr}[\mathbf{H}^v(\mathbf{W}^v)^{\mathrm{T}}\mathbf{K}^v\mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}]\} + \mathrm{Tr}[(\mathbf{H}^v)^{\mathrm{T}}\mathbf{L}^v\mathbf{H}^v],\end{aligned} \tag{13}$$

where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is an identity matrix.

Then, the Lagrangian function of (12) becomes,

$$\begin{aligned}& \mathcal{L}(\mathbf{W}^v, \mathbf{H}^v, \Psi, \Phi) \\ = & \ \alpha^v\{\mathrm{Tr}(\mathbf{K}) - 2\mathrm{Tr}[\mathbf{H}^v(\mathbf{W}^v)^{\mathrm{T}}\mathbf{K}^v] \\ + & \ \mathrm{Tr}[\mathbf{H}^v(\mathbf{W}^v)^{\mathrm{T}}\mathbf{K}^v\mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}]\} \\ + & \ \mathrm{Tr}[(\mathbf{H}^v)^{\mathrm{T}}\mathbf{L}^v\mathbf{H}^v] \\ + & \ \mathrm{Tr}[\Psi(\mathbf{W}^v)^{\mathrm{T}}] + \mathrm{Tr}[\Phi(\mathbf{H}^v)^{\mathrm{T}}],\end{aligned} \tag{14}$$

where $\Psi = [\psi_{ic}]$ and $\Phi = [\phi_{ic}]$ are the Lagrangian multipliers.

The partial derivatives of (14) with respect to $\mathbf{W}^v$ and $\mathbf{H}^v$ are

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^v} = \alpha^v[-2\mathbf{K}^v\mathbf{H}^v + 2\mathbf{K}^v\mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}\mathbf{H}^v] + \Psi, \tag{15}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{H}^v} = {} & \alpha^v[-2\mathbf{K}^v\mathbf{W}^v + 2\mathbf{H}^v(\mathbf{W}^v)^{\mathrm{T}}\mathbf{K}^v\mathbf{W}^v] \\ & + 2\mathbf{L}^v\mathbf{H}^v + \Phi. \end{aligned} \tag{16}$$

Following the Karush-Kuhn-Tucker conditions [24] $\psi_{ic}w_{ic} = 0$ and $\phi_{ic}h_{ic} = 0$, we obtain the following equations:

$$-(\mathbf{K}^v\mathbf{H}^v)_{ic}w_{ic}^v + (\mathbf{K}^v\mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}\mathbf{H}^v)_{ic}w_{ic}^v = 0, \tag{17}$$

$$\begin{aligned} & -(\alpha^v\mathbf{K}^v\mathbf{W}^v)_{ic}h_{ic}^v + (\alpha^v\mathbf{H}^v(\mathbf{W}^v)^{\mathrm{T}}\mathbf{K}^v\mathbf{W}^v)_{ic}h_{ic}^v \\ & \quad + (\mathbf{L}^v\mathbf{H}^v)_{ic}h_{ic}^v = 0. \end{aligned} \tag{18}$$

For non-negative data matrices, the equations (17) and (18) lead to the updating rules:

$$w_{ic}^v \leftarrow w_{ic}^v \frac{(\mathbf{K}^v\mathbf{H}^v)_{ic}}{(\mathbf{K}^v\mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}\mathbf{H}^v)_{ic}}, \tag{19}$$

$$h_{ic}^v \leftarrow h_{ic}^v \frac{(\alpha^v\mathbf{K}^v\mathbf{W}^v + \mathbf{S}^v\mathbf{H}^v)_{ic}}{(\alpha^v\mathbf{H}^v(\mathbf{W}^v)^{\mathrm{T}}\mathbf{K}^v\mathbf{W}^v + \mathbf{D}^v\mathbf{H}^v)_{ic}}. \tag{20}$$

The second step is to fix $\mathbf{W}^v$, $\mathbf{H}^v$, and update α^v. Then, (11) becomes:

$$\begin{aligned} & \min_{\alpha^v} \sum_{v=1}^{n_v} \alpha^v \|\mathbf{X}^v - \mathbf{X}^v\mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}\|_{\mathrm{F}}^2 + \gamma\sum_{v=1}^{n_v}(\alpha^v)^2, \\ & \text{s.t.} \;\; \forall v, \alpha^v \geq 0, \sum_{v=1}^{n_v} \alpha^v = \lambda. \end{aligned} \tag{21}$$

Denoting $f^v = \|\mathbf{X}^v - \mathbf{X}^v\mathbf{W}^v(\mathbf{H}^v)^{\mathrm{T}}\|_{\mathrm{F}}^2$, we can solve the objective function (21) as:

$$\begin{aligned} \mathcal{O}_2 = {} & \min_{\boldsymbol{\alpha}} \left\| \boldsymbol{\alpha} + \frac{1}{2\gamma}\boldsymbol{f} \right\|^2 \\ & \text{s.t.} \;\; \boldsymbol{\alpha} \geq 0, \mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha} = \lambda, \end{aligned} \tag{22}$$

where $\boldsymbol{\alpha} = [\alpha^1, \alpha^2, \cdots, \alpha^{n_v}]^{\mathrm{T}}$, and $\boldsymbol{f} = [f^1, f^2, \cdots, f^{n_v}]^{\mathrm{T}}$.

The Lagrangian function of (22) is

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\alpha}, \eta, \boldsymbol{\beta}\right) = &\left\|\boldsymbol{\alpha} + \frac{1}{2\gamma}\boldsymbol{f}\right\|^2 \\ &+ \eta\left(\lambda - \mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha}\right) + \boldsymbol{\beta}^{\mathrm{T}}(-\boldsymbol{\alpha}),\end{aligned} \tag{23}$$

where η and $\boldsymbol{\beta}$ are the Lagrangian multipliers.

According to the Karush-Kuhn-Tucker condition [24], it can be verified that the optimal solution $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha} = \left(-\frac{1}{2\gamma}\boldsymbol{f} + \eta\mathbf{1}\right)_+ . \tag{24}$$

Hence, α^v for all n_v views are learned through (24) and the value of α^v is learned according to the corresponding f^v in each view.

The entire algorithm for solving the problem (11) is summarized in the Algorithm 1.

Algorithm 1 Algorithm to solve the problem (11)

```
Input: Data for $n_v$ views $\{\mathbf{X}^1, \mathbf{X}^2, \cdots, \mathbf{X}^{n_v}\}$, parameters $\gamma$, $\lambda$, and the
  number of clusters $k$.
Output: $[\alpha^1\mathbf{H}^1, \alpha^2\mathbf{H}^2, \cdots, \alpha^{n_v}\mathbf{H}^{n_v}]^{\mathrm{T}}$.
Initialize: $\forall v$, initialize $\mathbf{W}^v$, $\mathbf{H}^v$, $\alpha^v$, and $\mathbf{S}^v$.
while not converge do
    for $v \in [1, n_v]$ do
        while not converge do
            Update $\mathbf{W}^v$ and $\mathbf{H}^v$ by (19) and (20).
            Normalize $\mathbf{W}^v$ and $\mathbf{H}^v$ by (4) and (5).
        end while
    end for
    Update $\boldsymbol{\alpha}$ by solving the problem (22).
end while
```

Regarding the updating rules above, we propose the following *Theorem 1*:

Theorem 1. *The objective function $\mathcal{O}$ in* (11) *is non-increasing under the updating rules in* (19), (20) *and* (24). *The objective function is invariant* *under these updates if and only if* $\mathbf{W}^v$, $\mathbf{H}^v$ *and* α^v *are at a stationary point.*

The objective function $\mathcal{O}_2$ in (22) is a convex optimization problem. To prove *Theorem 1* we first prove the following *Theorem 2*:

Theorem 2. *The objective function* $\mathcal{O}_1$ *in* (12) *is non-increasing under the updating rules in* (19) *and* (20)*. The objective function is invariant under these updates if and only if* $\mathbf{W}^v$ *and* $\mathbf{H}^v$ *are at a stationary point.*

Notably, the updating rules of $\mathbf{W}$ and $\mathbf{H}$ in each view v do not depend on another view in (12). So we use $\mathbf{X}$, $\mathbf{W}$, $\mathbf{H}$ and α to represent $\mathbf{X}^v$, $\mathbf{W}^v$, $\mathbf{H}^v$ and α^v for brevity in the rest of this section.

We use an auxiliary function as used in the expectation maximization algorithm [25, 26] to prove the convergence of (11). The definition of the auxiliary function is given by the following *Definition 1.*

Definition 1. $G(w, w')$ *is an auxiliary function for* $F(w)$ *if the conditions*

$$G(w, w') \geq F(w), G(w, w) = F(w)$$

are satisfied.

The auxiliary function is helpful because of the following *Lemma 1*:

Lemma 1. *If* G *is an auxiliary function of* F*, then* F *is non-increasing under the update*

$$w^{t+1} = \arg\min_{w} G(w, w^t), \tag{25}$$

where t *is the number of iteration times.*

Proof.

$$F(w^{t+1}) \leq G(w^{t+1}, w^t) \leq G(w^t, w^t) = F(w^t). \tag{26}$$

□

The equality $F(w^{t+1}) = F(w^t)$ holds only if w^t is a local minimum of $G(w, w^t)$. By iterating the updates in (25), the sequence of estimates will converge to a local minimum $w_{\min} = \arg\min_{w} F(w)$.

Now that the minimum of the objective function $\mathcal{O}_1$ in (12) is exactly our update rules with *Lemma 1* and proper auxiliary functions, *Theorem 2* can be proved.

First, we prove the convergence of the update rule in (19). Given an element w_{ic} in $\mathbf{W}$, we use F_{ic} to denote the part that is only relevant to w_{ic} in $\mathcal{O}_1$. Since the update is essentially element-wise, it is sufficient to show that each F_{ic} is non-increasing under the update step of (20). Let F'_{ic} denote the first order derivative of $\mathcal{O}_1$ with respective to $\mathbf{W}$, we define the auxiliary function G for F_{ic} as follows.

Lemma 2. *The function*

$$\begin{aligned} G(w, w^t_{ic}) = F_{ic}(w^t_{ic}) + F'_{ic}(w^t_{ic})(w - w^t_{ic}) \\ + \frac{\alpha(\mathbf{KWH}^{\mathrm{T}}\mathbf{H})_{ic}}{w^t_{ic}}(w - w^t_{ic})^2 \end{aligned} \tag{27}$$

is an auxiliary function for F_{ic}, which is a part of $\mathcal{O}_1$ and relevant to w_{ic} only.

Proof. Obviously, $G(w, w) = F_{ic}(w)$. According to the definition of auxiliary function, we only need to prove that $G(w, w^t_{ic}) \geq F_{ic}(w)$. To do so, we compare (27) with the Taylor series expansion of $F_{ic}(w)$:

$$\begin{aligned} F_{ic}(w) = F_{ic}(w^t_{ic}) + F'_{ic}(w^t_{ic})(w - w^t_{ic}) \\ + \frac{1}{2}F''_{ic}(w - w^t_{ic})^2, \end{aligned} \tag{28}$$

where F''_{ic} is the second order derivative with respect to $\mathbf{W}$.

It is not difficult to check,

$$F'_{ic} = 2\alpha(-\mathbf{KH} + \mathbf{KWH}^{\mathrm{T}}\mathbf{H})_{ic}, \tag{29}$$

$$F''_{ic} = 2\alpha(\mathbf{K})_{i'i}(\mathbf{H}^{\mathrm{T}}\mathbf{H})_{cc'}. \tag{30}$$

Subsituting eqreff1 and (30) into (28) and with (27) to find that $G(h, h^t_{ic}) \geq F_{ic}(h)$ is equivalent to prove

$$\frac{\alpha(\mathbf{KWH}^{\mathrm{T}}\mathbf{H})_{ic}}{w^t_{ic}} \geq \frac{1}{2}F''_{ic} = \alpha(\mathbf{K})_{i'i}(\mathbf{H}^{\mathrm{T}}\mathbf{H})_{cc'}. \tag{31}$$

To prove the inequality above, we have

$$\begin{aligned}&(\mathbf{K}\mathbf{W}\mathbf{H}^{\mathrm{T}}\mathbf{H})_{ic}\\&=\sum_{i=1}^{n}(\mathbf{K})_{i'i}\big[\sum_{c=1}^{k}w_{ic}(\mathbf{H}^{\mathrm{T}}\mathbf{H})_{cc'}\big]_{ic'}\\&\geq w_{ic}(\mathbf{K})_{i'i}(\mathbf{H}^{\mathrm{T}}\mathbf{H})_{cc'}.\end{aligned} \tag{32}$$

Thus, (31) holds and $G(w, w_{ic}^t) \geq F_{ic}(w)$. □

According to (25), the optimum w^{t+1} can be obtained by calculating the first order derivative (27) with respect to w, i.e., $\frac{\partial G(w,w_{ic}^t)}{\partial w} = 0$.

Thus,

$$\begin{aligned}w_{ic}^{t+1}&=w_{ic}^{t}-w_{ic}^{t}\frac{F_{ic}'(w_{ic}^{t})}{2\alpha(\mathbf{K}\mathbf{W}\mathbf{H}^{\mathrm{T}}\mathbf{H})_{ic}}\\&=w_{ic}^{t}\frac{(\mathbf{K}\mathbf{H})_{ic}}{(\mathbf{K}\mathbf{W}\mathbf{H}^{\mathrm{T}}\mathbf{H})_{ic}}.\end{aligned} \tag{33}$$

Then we define an auxiliary function for the update rule in (20). Similarly, for any element h_{ic} in $\mathbf{H}$, let $\mathcal{F}_{ic}$ denote the part of $\mathcal{O}_1$ that is relevant to h_{ic} only. The auxiliary function for the objective function with regard to variable h_{ic} is defined as follows.

Lemma 3. *The function*

$$\begin{aligned}\mathcal{G}(h,h_{ic}^{t})=&\mathcal{F}_{ic}(h_{ic}^{t})+\mathcal{F}_{ic}'(h_{ic}^{t})(h-h_{ic}^{t})\\&+\frac{(\alpha\mathbf{H}\mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{W}+\mathbf{D}\mathbf{H})_{ic}}{h_{ic}^{t}}(h-h_{ic}^{t})^{2}\end{aligned} \tag{34}$$

is an auxiliary function for $\mathcal{F}_{ic}$, which is a part of $\mathcal{O}_1$ and relevant to h_{ic} only.

Proof. It is equivalent to prove

$$\frac{(\alpha\mathbf{H}\mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{W}+\mathbf{D}\mathbf{H})_{ic}}{h_{ic}^{t}}\geq\frac{1}{2}\mathcal{F}_{ic}''=\alpha(\mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{W})_{cc}+(\mathbf{L})_{ii}. \tag{35}$$

To prove the above inequality, we have

$$\begin{aligned}\alpha(\mathbf{H}\mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{W})_{ic} &= \alpha\sum_{s=1}^{k} h_{is}^{t}(\mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{W})_{sc} \\ &\geq \alpha h_{ic}^{t}(\mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{W})_{cc}\end{aligned} \tag{36}$$

and

$$\begin{aligned}(\mathbf{D}\mathbf{H})_{ic} = \sum_{j=1}^{n}(\mathbf{D})_{ij}h_{jc}^{t} &\geq (\mathbf{D})_{ii}h_{ic}^{t} \\ \geq (\mathbf{D}-\mathbf{S})_{ii}h_{ic}^{t} &= (\mathbf{L})_{ii}h_{ic}^{t}.\end{aligned} \tag{37}$$

Thus, (35) holds and $\mathcal{G}(h, h_{ic}^{t}) \geq \mathcal{F}_{ic}(h)$ with (36) and (37). □

Similarly, Substituting $\mathcal{G}(h, h_{ic}^{t})$ of (34) into (25), we get

$$\begin{aligned}h_{ic}^{t+1} &= \arg\min_{h}\mathcal{G}(h, h_{ic}^{t}) \\ &= h_{ic}^{t}\frac{(\alpha\mathbf{K}\mathbf{W}+\mathbf{S}\mathbf{H})_{ic}}{(\alpha\mathbf{H}\mathbf{W}^{\mathrm{T}}\mathbf{K}\mathbf{W}+\mathbf{D}\mathbf{H})_{ic}}.\end{aligned}$$

Since (35) is an auxiliary function, $\mathcal{F}_{ic}$ is non-increasing under this update rule according to *Lemma 3*.

Until now, we have proved *Theorem 2*. Again, since the objective function $\mathcal{O}_2$ in (22) is a convex optimization problem, the *Theorem 1* is also proved.

3.2. Algorithm for Data with Negative Values

For data matrices which contains negative values, our multiplicative updating algorithm is based on the following *Theorem 3* proposed by Sha *et al.* [27].

Theorem 3. *Define the general problem of nonnegative quadratic programming. The minimization of the quadratic objective function is as following:*

$$\begin{aligned}&\min_{\boldsymbol{y}} f(\boldsymbol{y}) = \min_{\boldsymbol{y}} \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\mathbf{A}\boldsymbol{y} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}, \\ &\text{s.t.}\quad \boldsymbol{y} \geq 0,\end{aligned} \tag{38}$$

where $\mathbf{A}$ *is an arbitrary* $n \times n$ *symmetric semi-positive matrix and* $\boldsymbol{b}$ *is an arbitrary* $n \times 1$ *vector. The iterative solution is expressed in terms of the positive component* $\mathbf{A}^+$ *and negative component* $\mathbf{A}^-$ *of the matrix* $\mathbf{A}$ *in* (38)*,*

$$(\mathbf{A}^+)_{ij} = \begin{cases} (\mathbf{A})_{ij}, & \textit{if } (\mathbf{A})_{ij} > 0 \\ 0 & \textit{otherwise} \end{cases} \tag{39}$$

$$(\mathbf{A}^-)_{ij} = \begin{cases} |(\mathbf{A})_{ij}|, & \textit{if } (\mathbf{A})_{ij} < 0 \\ 0 & \textit{otherwise} \end{cases} \tag{40}$$

It is easy to find that $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$*. The solution* $\boldsymbol{y}$ *that minimizes* (38) *can be obtained by the following updating rule,*

$$y_i^{t+1} = y_i^t \left[\frac{-b_i + \sqrt{b_i^2 + 4(\mathbf{A}^+\boldsymbol{y}^t)_i(\mathbf{A}^-\boldsymbol{y}^t)_i}}{2(\mathbf{A}^+\boldsymbol{y}^t)_i} \right]. \tag{41}$$

Proof. The function

$$\begin{aligned} g(y_i, y_i^t) = {} & \frac{1}{2}\sum_i \frac{(\mathbf{A}^+\boldsymbol{y}^t)_i}{y_i^t} y_i^2 \\ & - \frac{1}{2}\sum_{ij}(\mathbf{A}^-)_{ij} y_i^t y_j^t \left(1 + \log \frac{y_i y_j}{y_i^t y_j^t}\right) + \sum_i b_i y_i \end{aligned} \tag{42}$$

is an auxiliary function for $f(y_i)$.

We can obtain $g(y_i, y_i^t) \geq f(y_i)$ according to [27]. Then, the (42) is minimized by setting its derivative to zero with respect to y_i, leading to the updating rule in (41). □

It can be seen that $\mathcal{O}_1$ is a quadratic form of $\mathbf{W}$ or $\mathbf{H}$ from (13), so (41) can be applied to solve the objective function $\mathcal{O}_1$ and the corresponding $\mathbf{A}$ and $\boldsymbol{b}$ need to be identified. By fixing $\mathbf{H}$, the part $\boldsymbol{b}$ for quadratic form of $\mathcal{O}_1(\mathbf{W})$ can be obtained by (29) at $\mathbf{W} = 0$, and the part $\mathbf{A}$ for quadratic form of $\mathcal{O}_1(\mathbf{W})$ can be obtained by (30).

Substituting $\mathbf{A}$ and $\boldsymbol{b}$ into (41), we obtain the update rule of w_{ic},

$$w_{ic}^{t+1} = w_{ic}^t \frac{(\mathbf{KH})_{ic} + \sqrt{(\mathbf{KH})_{ic}^2 + 4(\mathbf{Q}^+)_{ic}(\mathbf{Q}^-)_{ic}}}{2(\mathbf{Q}^+)_{ic}}, \tag{43}$$

where $\mathbf{Q}^+ = \mathbf{K}^+\mathbf{W}\mathbf{H}^{\mathrm{T}}\mathbf{H}$, $\mathbf{Q}^- = \mathbf{K}^-\mathbf{W}\mathbf{H}^{\mathrm{T}}\mathbf{H}$, $\mathbf{K}^+$ and $\mathbf{K}^-$ denotes the non-negative matrices with elements,

$$(\mathbf{K}^+)_{ij} = \begin{cases} (\mathbf{K})_{ij}, & \text{if } (\mathbf{K})_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{44}$$

and

$$(\mathbf{K}^-)_{ij} = \begin{cases} |(\mathbf{K})_{ij}|, & \text{if } (\mathbf{K})_{ij} < 0 \\ 0 & \text{otherwise} \end{cases} \tag{45}$$

It is easy to derive that $\mathbf{K} = \mathbf{K}^+ - \mathbf{K}^-$.

Similarly, we can obtain the updating rule of h_{ic},

$$h_{ic}^{t+1} = h_{ic}^{t} \frac{(\mathbf{KW})_{ic} + \sqrt{(\mathbf{KW})_{ic}^2 + 4(\mathbf{P}^+)_{ic}(\mathbf{P}^-)_{ic}}}{2(\mathbf{P}^+)_{ic}}, \tag{46}$$

where $\mathbf{P}^+ = \alpha\mathbf{H}\mathbf{W}^{\mathrm{T}}\mathbf{K}^+\mathbf{W} + \mathbf{D}\mathbf{H}$ and $\mathbf{P}^- = \alpha\mathbf{H}\mathbf{W}^{\mathrm{T}}\mathbf{K}^-\mathbf{W} + \mathbf{S}\mathbf{H}$.

4. Experimental results

4.1. Datasets

In this paper, we test our method, MVCF, on three benchmark multi-view datasets.

3-Sources[1] is constructed from three well-known online news sources: BBC, Reuters, and Guardian. In total there are 948 news articles covering 416 distinct news stories from the period February to April 2009. Of these stories, 169 were reported in all three sources. Each story was manually annotated with one of the six topical labels: business, entertainment, health, politics, sport and technology.

Cora[2] contains 2708 documents over seven labels (neural networks, rule learning, reinforcement learning, probabilistic methods, theory, genetic algorithms, and case based). In this paper, two views, content and cites, are used. The documents are described by 1433 words in the content view, and by 5429 links between them in the citations views.

[1] `http://mlg.ucd.ie/datasets`

[2] `http://lig-membres.imag.fr/grimal/data.html`

Cornell[28] contains 195 documents over five labels (student, project, course, staff, faculty). In this paper, we use two views, content and cites. The documents are described by 1703 words in the content view, and by the 569 links between them in the citations views.

4.2. Baseline algorithms

We compared MVCF with the state-of-the-art methods, including

1. CTSC [29]: It is multi-view spectral clustering approach using the idea of co-training. Under the assumption that the true underlying clustering would assign a point to the same cluster irrespective of the view, CTSC learns the clustering result in one view and then use the results to label the data in other views so as to modify the graph structure (similarity matrix).
2. CRSC [11]: It applies the centroid based co-regularization scheme to the multi-view spectral clustering. To uncover the data structure shared by different views, CRSC enforces the view-specific eigenvectors to look similar by regularizing them towards a common consensus, and then optimizes individual clusterings as well as the consensus by utilizing a joint cost function.
3. MultiNMF [15]: It aims to search for a factorization that gives compatible clustering solutions across multiple views, requiring coefficient matrices learnt from factorizations of different views to be regularized towards a common consensus.
4. RMKMC [30]: It simultaneously performs the clustering using each view of features and unifies their results based on their importance to the clustering task. $\ell_{2,1}$-norm is employed to improve the robustness.
5. RMSC [12]: For each view, it constructs a corresponding transition probability matrix, which is then used for recovering a low-rank transition probability matrix. The standard Markov chain method is then utilized for processing before clustering is conducted.
6. GRNMF [17]: This is a graph-regularized NMF-based multi-view clustering method. The GRNMF extends the Liu *et al.*'s algorithm [15] with a graph regularization.

4.3. Evaluation metric

Three metrics, the clustering accuracy (ACC), the normalized mutual information (NMI) [31] and the purity [32] are used to evaluate the performances in this work. For each metric, a higher value indicates better clustering quality. These measurements are widely used by comparing the obtained label of each sample with that provided by the datasets in different clustering approaches.

ACC measures the percentage of correct labels obtained, which is defined as

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} \delta(l_i, \mathrm{map}(r_i))}{n} \tag{47}$$

where n denotes the total number of documents, l_i denotes the ground-truth label, r_i denotes the obtained cluster label, δ denotes the delta function

$$\delta(x, y) = \begin{cases} 1, & \text{if } \ x = y \\ 0, & \text{otherwise} \end{cases} \tag{48}$$

and $\mathrm{map}(r_i)$ is the optimal mapping function that permutes clustering labels to match the ground-truth labels. The best mapping can be found by using the Kuhn-Munkres algorithm [33].

NMI is used to measure the similarity between the cluster assignments and the pre-existing input labeling of the classes. Let n_c be the number of objects in cluster $m_c(1 \leq c \leq k)$ obtained by using the clustering algorithms and $\tilde{n}_s$ be the object number of cluster $g_s(1 \leq s \leq k)$ in the ground-truth labels. NMI is defined as

$$\mathrm{NMI} = \frac{\sum_{c=1}^{k} \sum_{s=1}^{k} n_{c,s} \log \left(\frac{n \cdot n_{c,s}}{n_c \tilde{n}_s} \right)}{\sqrt{\left(\sum_{c=1}^{k} n_c \log \frac{n_c}{n} \right) \left(\sum_{s=1}^{k} \tilde{n}_s \log \frac{\tilde{n}_s}{n} \right)}} \tag{49}$$

where $n_{c,s}$ is the number of object which are in the intersection of cluster m_c and g_s. NMI varies from 0 for a totally wrongly clustering dataset to 1 for a perfectly clustering dataset.

Purity is given by

$$\mathrm{purity} = \frac{1}{n} \sum_{i=1}^{k} \max_{1 \leq c \leq k} |m_i \cap g_j|. \tag{50}$$

where n is the number of data points belonging to k clusters. m_i represents the i^{th} obtained clusters, and g_i implies the i^{th} ground-truth clusters.

4.4. Experimental setup

The parameters of all comparing methods to be compared with are tuned to achieve the best results, according to the parameter settings in original papers where the approaches were first proposed. For MVCF, two parameters involved, γ and λ, are fixed as $\gamma = 5$ and $\lambda = 4.5$ for all datasets. We initialize $\mathbf{W}^v$ and $\mathbf{H}^v$ randomly within the range [0,1], and α^v is initialized by $\frac{\lambda}{n_v}$. Then, we run the experiment for t times until the objective function converges to obtain the new data representation $\mathbf{H}^v$. The convergence criteria applied is

$$\left|\frac{\mathcal{O}_{t+1} - \mathcal{O}_t}{\mathcal{O}_t}\right| < 10^{-5} \tag{51}$$

where $\mathcal{O}_t$ is the objective function value in the t-th iteration of each algorithm. Finally, we obtain the optimal data representations by adding the product of the data representation matrix $\mathbf{H}^v$ and its weight in each view together. k-means is applied to the optimal data representation for document clustering, which is repeated 10 times. The average result in terms of the cost function of k-means is noted. Finally, we compare the obtained clusters with the grouth truth to compute the ACC, NMI, and purity.

4.5. Comparisons of performance

The average performance for the three datasets are shown in Tables 1, 2 and 3. In each column of the tables, the best results are highlighted in **boldface**, and the second best are highlighted in *italic*.

For all three datasets, the performances of the proposed MVCF are better than other methods. Specifically, Table 1 shows that ACC by MVCF are 4.57%, 7.16% and 7.18% higher than the second best results on the three datasets, respectively. Table 2 presents MVCF produces the highest NMI, outperforming the second best results with a significant margin, especially on 3-Sources and Cornell. In addition, the corresponding purity is increased with 15.98%, 8.20%, and 4.61% with MVCF as shown in Table 3. Note that MVCF achieves the largest improvements on 3-Sources Dataset, which contains the most views. This is due to a fact that MVCF utilizes the multiple views efficiently according to the importance of each view and the results of NMF-based clustering approaches have better semantic interpretation [1, 2, 13, 14].

Table 1: Clustering performance on three datasets-**ACC** (%)

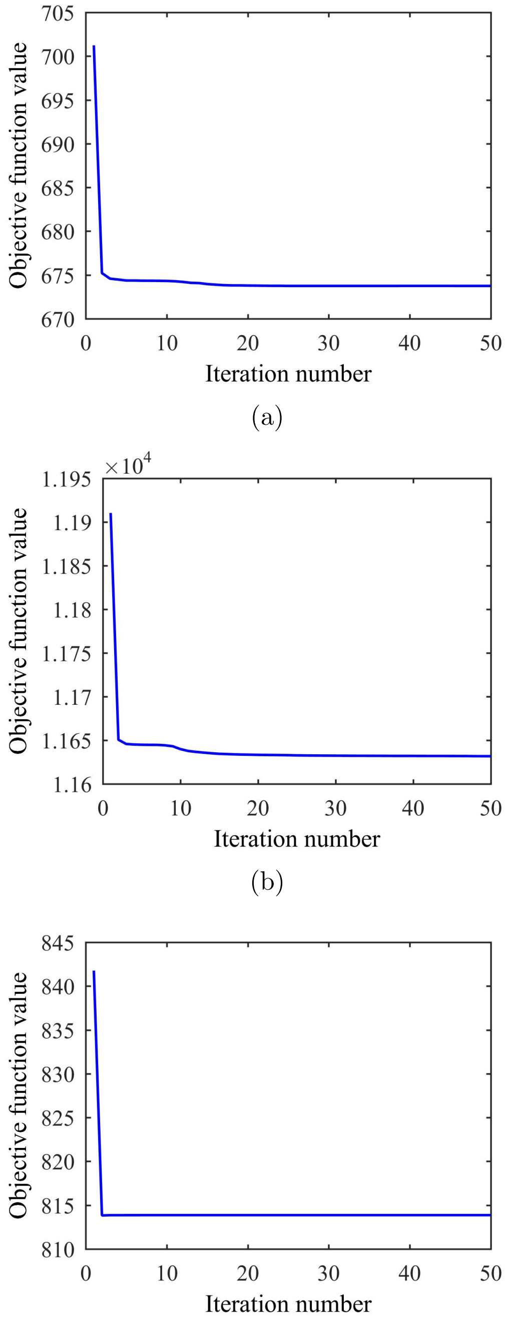
	3-Sources	Cora	Cornell
CTSC	60.83	48.42	44.51
CRSC	63.31	35.60	57.44
MultiNMF	54.44	*48.60*	43.59
RMKMC	56.21	46.34	*60.51*
RMSC	54.44	34.08	38.46
GRNMF	*76.50*	34.42	42.05
MVCF	**81.07**	**55.76**	**67.69**

Table 2: Clustering performance on three datasets-**NMI**(%)

	3-Sources	Cora	Cornell
CTSC	54.65	27.88	23.19
CRSC	54.13	14.52	*33.03*
MultiNMF	48.24	*28.54*	14.16
RMKMC	27.81	22.44	26.83
RMSC	43.15	16.36	14.27
GRNMF	*67.04*	21.01	11.86
MVCF	**75.53**	**31.89**	**45.87**

Table 3: Clustering performance on three datasets-**purity**(%)

	3-Sources	Cora	Cornell
CTSC	66.86	43.50	54.87
CRSC	*68.64*	39.36	*63.08*
MultiNMF	63.31	*49.00*	47.69
RMKMC	58.58	46.34	60.51
RMSC	65.68	41.06	47.69
GRNMF	65.09	39.59	46.15
MVCF	**84.62**	**57.20**	**67.69**



(a)

(b)

(c)

Figure 1: Convergence curves over different datasets (a) 3-Sources, (b) Cora, (c) Cornell.

4.6. Study of convergence

To demonstrate the convergence of MVCF, Fig. 1 illustrates the convergence speed on all three datasets. In each sub-figure, the x-axis and the y-axis denote the iteration number and the corresponding objective function value, respectively. We can see that the value of the objective function decreases sharply within five iterations and then becomes steadily afterwards. This indicates that MVCF converges efficiently sufficient.

5. Conclusion and future work

In this paper, we proposed a multi-view document clustering method using the MVCF. The method fully exploits multi-view feature information and reduces the data dimension to achieve better clustering performance. With MVCF, high dimensional data points from different views are reduced to low dimensional data representations with more locally consistent structure. Both the new representation matrices and view weights are learned by MVCF. The clustering labels are obtained by running k-means on the low-dimensional data. We theoretically proved the convergence of our algorithm, and this is in accordance with our experiments. The experimental results showed that MVCF achieves higher performance than the state-of-the-art methods in terms of clustering accuracy, normalized mutual information and clustering purity. In the future, we will bring the sparse regulation into MVCF to obtain a more accurate data representation matrix, with which better clustering performance can be expected.

Acknowledgments

Jing Wang is the corresponding author. This work has been supported by the National Science Foundation of China under the Grant No. 61201422, the Specialized Research Fund for the Doctoral Program of Higher Education under the Grant No. 20120211120013, and the Fundamental Research Funds for the Central Universities under the Grant No. lzujbky-2016-239.

References

[1] W. Xu, Y. Gong, Document clustering by concept factorization, in: Proc. of SIGIR, Vol. 27, ACM, 2004, pp. 202–209.

[2] W. Xu, X. Liu, Y. Gong, Document clustering based on non-negative matrix factorization, in: Proc. of SIGIR, Vol. 26, ACM, 2003, pp. 267–273.

[3] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.

[4] D. Cai, X. He, J. Han, Locally consistent concept factorization for document clustering, IEEE Transactions on Knowledge and Data Engineering 23 (6) (2011) 902–913.

[5] H. Liu, Z. Yang, Z. Wu, Locality-constrained concept factorization, in: Proc. of IJCAI, Vol. 22, 2011, pp. 1378–1383.

[6] H. Liu, G. Yang, Z. Wu, D. Cai, Constrained concept factorization for image representation, IEEE Transactions on Cybernetics 44 (7) (2014) 1214–1224.

[7] Y. He, H. Lu, L. Huang, S. Xie, Pairwise constrained concept factorization for data representation, Neural Networks 52 (2014) 1 – 17.

[8] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proc. of COLT, Vol. 11, ACM, 1998, pp. 92–100.

[9] S. Bickel, T. Scheffer, Multi-view clustering., in: Proc. of ICDM, Vol. 4, 2004, pp. 19–26.

[10] X. Song, L. Nie, L. Zhang, M. Akbari, T.-S. Chua, Multiple social network learning and its application in volunteerism tendency prediction, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2015, pp. 213–222.

[11] A. Kumar, P. Rai, H. Daume, Co-regularized multi-view spectral clustering, in: Proc. of NIPS, Vol. 24, 2011, pp. 1413–1421.

[12] R. Xia, Y. Pan, L. Du, J. Yin, Robust multi-view spectral clustering via low-rank and sparse decomposition, in: Proc. of AAAI, Vol. 28, 2014, pp. 2149–2155.

[13] C. H. Ding, X. He, H. D. Simon, On the equivalence of nonnegative matrix factorization and spectral clustering, in: Proc. of SDM, Vol. 5, SIAM, 2005, pp. 606–610.

[14] F. Shahnaz, M. W. Berry, V. Pauca, R. J. Plemmons, Document clustering using nonnegative matrix factorization, Information Processing and Management 42 (2) (2006) 373 – 386.

[15] J. Liu, C. Wang, J. Gao, J. Han, Multi-view clustering via joint nonnegative matrix factorization, in: Proc. of SDM, Vol. 13, SIAM, 2013, pp. 252–260.

[16] X. Zhang, L. Zhao, L. Zong, X. Liu, H. Yu, Multi-view clustering via multi-manifold regularized nonnegative matrix factorization, in: Proc. of ICDM, Vol. 14, 2014, pp. 1103–1108.

[17] Z. Wang, X. Kong, H. Fu, M. Li, Y. Zhang, Feature extraction via multiview non-negative matrix factorization with local graph regularization, in: Proc. of ICIP, IEEE, 2015, pp. 3500–3504.

[18] X. He, P. Niyogi, Locality preserving projections, in: Proc. of NIPS, Vol. 16, 2003, pp. 153–160.

[19] D. Cai, X. Wang, X. He, Probabilistic dyadic data analysis with local and global consistency, in: Proc. of ICML, Vol. 26, ACM, 2009, pp. 105–112.

[20] Y. Yang, D. Xu, F. Nie, S. Yan, Y. Zhuang, Image clustering using local discriminant models and global integration, IEEE Transactions on Image Processing 19 (10) (2010) 2761–2773.

[21] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, Y. Pan, A multimedia retrieval framework based on semi-supervised ranking and relevance feedback, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (4) (2012) 723–742.

[22] F. R. Chung, Spectral Graph Theory, American Mathematical Soc., 1997.

[23] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering., in: Proc. of NIPS, Vol. 14, 2001, pp. 585–591.

[24] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[25] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B (Methodological) 39 (1) (1977) 1–38.

[26] D. D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, in: Proc. of NIPS, Vol. 13, 2001, pp. 556–562.

[27] F. Sha, L. K. Saul, D. D. Lee, Multiplicative updates for nonnegative quadratic programming in support vector machines, in: Proc. of NIPS, Vol. 15, 2002, pp. 1041–1048.

[28] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: Proc. of ACM MM, ACM, 2010, pp. 251–260.

[29] A. Kumar, H. Daumé, A co-training approach for multi-view spectral clustering, in: Proc. of ICML, Vol. 28, ACM, 2011, pp. 393–400.

[30] X. Cai, F. Nie, H. Huang, Multi-view k-means clustering on big data, in: Proc. of AAAI, Vol. 23, 2013, pp. 2598–2604.

[31] A. Strehl, J. Ghosh, Cluster ensembles-a knowledge reuse framework for combining partitionings, in: Proc. of AAAI, Vol. 12, 2002, pp. 93–99.

[32] R. Ievgen, B. Younes, Random subspaces NMF for unsupervised transfer learning, in: Proc. of IJCNN, Vol. 24, IEEE, 2014, pp. 3901–3908.

[33] L. Lovász, M. D. Plummer, Matching Theory, Vol. 367, American Mathematical Soc., 2009.