Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 6084382, 10 pages
https://doi.org/10.1155/2019/6084382

Research Article

Multiview Clustering via Robust Neighboring Constraint Nonnegative Matrix Factorization

Feiqiong Chen,[1] Guopeng Li,[2] Shuaihui Wang,[3,4] and Zhisong Pan[1]

[1]*Command and Control Engineering College, Army Engineering University of PLA, Nanjing 210000, China*
[2]*College of Information and Communication, National University of Defense Technology, Xi'an 710106, China*
[3]*Graduate School, Army Engineering University of PLA, Nanjing 210000, China*
[4]*Qinhuangdao Campus, Naval Aeronautical University, Qinhuangdao 066200, China*

Correspondence should be addressed to Zhisong Pan; hotpzs@hotmail.com

Received 21 September 2019; Accepted 24 October 2019; Published 23 November 2019

Academic Editor: Rafal Zdunek

Many real-world datasets are described by multiple views, which can provide complementary information to each other. Synthesizing multiview features for data representation can lead to more comprehensive data description for clustering task. However, it is often difficult to preserve the locally real structure in each view and reconcile the noises and outliers among views. In this paper, instead of seeking for the common representation among views, a novel robust neighboring constraint nonnegative matrix factorization (rNNMF) is proposed to learn the neighbor structure representation in each view, and $L_{2,1}$-norm-based loss function is designed to improve its robustness against noises and outliers. Then, a final comprehensive representation of data was integrated with those representations of multiviews. Finally, a neighboring similarity graph was learned and the graph cut method was used to partition data into its underlying clusters. Experimental results on several real-world datasets have shown that our model achieves more accurate performance in multiview clustering compared to existing state-of-the-art methods.

1. Introduction

Clustering is a fundamental topic in machine learning and data mining tasks. Datasets often are comprised of different views, and the views often provide compatible and complementary information in real world. Thus, multiview clustering (MVC) aims to integrate those different views and uncover the consistent latent information to achieve perfect clustering performance [1]. Over the past decades, it has attracted great attention [2, 3] and has been widely used in various real applications [4].

Essentially, given the multiview inputs, the critical work in MVC is to fuse information of different views and learn the common agreement for clustering. For efficiently integrating views, many subspace clustering-based methods [5, 6] and nonnegative matrix factorization- (NMF-) [7, 8] based methods have been developed. In particular, NMF is shown to be equivalent to relaxed k-means and symmetric NMF is closely related to spectral clustering [9]. However, NMF cannot preserve the geometrical structure of the data space, which is essential for the algorithms to find the true cluster structures. Many manifold learning methods, which are motivated by the so-called locally invariant idea that the nearby points are likely to have similar embedding, have been proposed, such as locally linear embedding (LLE) [10] and locality preserving projection (LPP) [11]. In particular, Cai et al. [12] proposed graph-regularized nonnegative matrix factorization (GNMF) to find a compact representation which can uncover the hidden semantics and simultaneously respect the intrinsic geometric structure. It is well accepted that the clustering performance can be significantly enhanced when the local invariance is considered. These are single-view clustering methods.

On the other hand, many NMF-based MVC methods [13] have attracted attention, in which various constraints have been applied to the coefficient matrix to cluster the

data points. Multi-NMF [14] formulated a joint multiview NMF learning process with the constraint that encourages representation of each view toward a common consensus. Many extensions of multi-NMF methods were proposed for image clustering and other tasks [15]. In [16], two weight matrices are introduced to alleviate the issue of dataset imbalance in real applications. Ou et al. [17] explored the local geometric structure for each view under the patch alignment framework and adopted correntropy-induced metric to measure the reconstruction error of each view to improve the robustness. A deep matrix factorization model [18] aimed to seek a common representation by introducing graph regularization to guide shared representation learning in the final layer of each view. However, existing approaches are all used to exploit common information shared by multiple views but neglect the diversity among views. The diversity means that each view of the data contains some distinct information that other views do not have.

In this paper, we propose a novel MVC method, with a novel algorithm, called robust neighboring constraint NMF (rNNMF), which uses the locally neighboring structure of each view to capture the diversity features. In rNNMF, a neighboring graph is constructed and updated for each view in factorization process to obtain the underlying diversity features. Finally, these diversity features will be combined to create an integrated feature for datasets, and then, a global graph is further generated from this integrated feature and Ncut is used to partition data into its underlying groups.

In summary, the novelty and contribution of our research are as follows:

(1) A neighboring constraint NMF method is proposed to learn the diversity representation of data in each view. The proposed model only keeps the nearest relationship between a point and its nearest neighbor to maintain geometrical structure in feature learning in each view.

(2) $L_{2,1}$-norm loss function is used in rNNMF to improve the robustness of feature in each view and reduce the effect of noisy features.

The rest of this paper will be organized as follows. Section 2 will introduce some related work about NMF-based MVC algorithms. Our proposed robust NMF-based MVC model will be introduced in Sections 3, the experimental results will be introduced in Section 4, and conclusions and discussions will be introduced in Section 5.

2. Related Work

Both subspace clustering and NMF-based methods are important in MVC. For example, a robust graph can be learned with correlation consensus agreement in [5] to improve the clustering performance. A multigraph regularized low-rank representation- (LRR-) based method is proposed to achieve the data correlation consensus among all views [6]. A structured LRR was proposed by factorizing into the latent low-dimensional data-cluster representations, which characterize the data clustering structure for each view [1]. Meanwhile, NMF-based methods [19] were also proved to be useful, which enforce the constraint that the elements of the factor matrices must be nonnegative. It shows that when the Frobenius norm is used as a divergence, NMF is equivalent to a relaxed form of K-means clustering method. However, NMF fails to discover the intrinsic geometry of the data, which is essential to the real applications. To preserve the locally geometrical structure of the data space, Cai et al. imposed graph regularization on NMF (GNMF). In [20], Shang et al. proposed graph dual regularization NMF (DNMF) which simultaneously considered the geometric structures of data manifold and feature manifold. Two subspace clustering algorithms were proposed in [21]. It established connection with spectral normalized cut [22] and ratio cut clustering. It also extended the nonlinear orthogonal NMF and introduced a graph regularization to obtain a factorization that respects a local geometric structure after the nonlinear mapping.

In MVC, NMF-based methods also have received increasing attention. Let input $\mathbf{X} = \left\{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(v)}, \ldots, \mathbf{X}^{(V)}\right\}$ and $\mathbf{X}^{(v)}$ be the v-th view, then it is a $d_v \times n$ matrix, where d_v denotes the feature dimensionality in row and n is the number of data in column. $\mathbf{H}^{(v)}$ is the representation of the v-th view, it is a $p \times n$ matrix, where v denotes the feature dimensionality and $\mathbf{Z}^{(v)}$ is the corresponding basis matrix. For NMF-based methods, the overall framework is as follows:

$$\begin{aligned} &\min_{\mathbf{Z}\mathbf{H}} = \|\mathbf{X}^v - \mathbf{Z}^v\mathbf{H}^v\|_F^2 + f(\mathbf{H}^v, \mathbf{H}^s), \\ &\text{s.t.} \qquad \mathbf{Z}^v \geq \mathbf{0},\ \mathbf{H}^v \geq \mathbf{0}, \end{aligned} \tag{1}$$

where $\|\cdot\|_F^2$ is the Frobenius norm, and both $\mathbf{Z}^v$ and $\mathbf{H}^v$ should be nonnegative. By default, the input data for each view should also be nonnegative for the NMF-based methods, i.e., $\mathbf{X}^{(v)} \geq 0$. $f(\mathbf{H}^v, \mathbf{H}^s)$ is the regularization term to learn the agreement among different views. For example, MulNMF designed a constraint that encourages representation of each view toward a common consensus $\mathbf{H}^*$. DiNMF [23] introduced a constraint term $tr(\mathbf{H}^{(v)}\mathbf{H}^{(s)T})$ to guarantee the diversity among points in different views. In order to deal with mixed-sign data, based on the semi-NMF model, a deep semi-NMF method couples the output representation in the final layer of factorization and enforces views that share the same representation after layer by layer factorization.

Although good performance can be achieved in those methods by finding a common agreement among views, the consensus information cannot be explored effectively and do not make full use of information of multiple views [18]. For full use of diversity information of views, combination of views' representations is a natural method. Our work focuses on this kind of combination styles. However, original information contained among views can usually lead to poor performance in clustering. Therefore, it is necessary to design a new method which can not only get maximum preservation diversity feature of each view but also obtain the aggregated representation with good clustering performance.

3. Structural Constraint Semi-NMF

3.1. Robust Neighboring Constraint Regularization. Given $\mathbf{H}^{(v)}$ is the low-dimensional representation of $\mathbf{X}^{(v)}$, we introduce a special matrix $\mathbf{R}^{(v)}$ to indicate the constraint of the v-th view; it is defined as

$$\mathbf{R}^{(v)}(i,j) = \begin{cases} -1, & i = j, \\ 1, & i \in N^{(v)}(j), \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $N^{(v)}(j)$ is the set of the nearest neighbors of point j in the v-th view. If point i is one of the nearest neighbors of j, $\mathbf{R}^{(v)}(i,j)$ will be set as 1. We hope the difference between the points j and i is as small as possible, which can be represented by $\mathbf{H}^{(v)}(:,j) - \mathbf{H}^{(v)}(:,i)$, where $\mathbf{H}^{(v)}(:,j)$ indicates the j-th column of $\mathbf{H}^{(v)}$. So, we introduce constraint $\mathbf{H}^{(v)}\mathbf{R}^{(v)}$ to describe this diversity of the point with its neighbors. The smaller the value of $\mathbf{H}^{(v)}\mathbf{R}^{(v)}$, the more similar they are. We introduce the $L_{2,1}$ norm to penalize $\mathbf{H}^{(v)}\mathbf{R}^{(v)}$ for seeking a representation in each view:

$$\left\|\mathbf{H}^{(v)}\mathbf{R}^{(v)}\right\|_{2,1}. \tag{3}$$

3.2. Objective Function and Optimization Algorithm. In MVC, to learn the neighbor information in each view, based on ORNMF [19] which is a robust representation approach, the proposed rNNMF can be expressed as

$$\begin{aligned} \min_{\mathbf{W}^{(v)}\mathbf{H}^{(v)}} &= \sum_{v=1}^{V}\left(\left\|\mathbf{X}^{(v)} - \mathbf{W}^{(v)}\mathbf{H}^{(v)}\right\|_{2,1} + \alpha\left\|\mathbf{H}^{(v)}\mathbf{R}^{(v)}\right\|_{2,1}\right), \\ \text{s.t.} \quad & \mathbf{H}^{(v)} \geq 0, \ \mathbf{W}^{(v)} \geq 0, \end{aligned} \tag{4}$$

where the $L_{2,1}$-norm is applied to the loss function and defined as $\|\mathbf{X} - \mathbf{WH}\|_{2,1} = \sum_{i=1}^{n}\|\mathbf{X}_i - \mathbf{WH}_i\|$. With the error for each point not being squared, the impact of large errors is reduced significantly. The first term in equation (4) indicates the v-th data fidelity term, and the second term is the nearest neighboring constraint term. $\mathbf{H}^{(v)}$ is the v-th representation from the v-th view. α is the positive parameter to specify the relative importance of the factorization term and regularization term in the model.

Like the most NMF-based methods, the objective function in (4) is not convex, so we present an iterative algorithm to achieve the local minima of (4).

Computing $\mathbf{H}^{(v)}$, to update $\mathbf{H}^{(v)}$ with $\mathbf{W}^{(v)}$ fixed, we need to solve the object function as follows:

$$\min_{\mathbf{H}^{(v)}\geq 0} J\left(\mathbf{H}^{(v)}\right) = \sum_{v=1}^{V}\left\|\mathbf{X}^{(v)} - \mathbf{W}^{(v)}\mathbf{H}^{(v)}\right\|_{2,1} + \alpha\left\|\mathbf{H}^{(v)}\mathbf{R}\right\|_{2,1}, \tag{5}$$

as $\mathbf{H}^{(v)} \geq 0$ and $J(\mathbf{H}^{(v)})$ being NP-hard, a Lagrange multiplier matrix $\mathbf{\Theta} = (\mathbf{\Theta})_{ij}$ is introduced. Then, the Lagrangian function $L(\mathbf{H}^{(v)})$ is

$$\begin{aligned} L\left(\mathbf{H}^{(v)}\right) = \sum_{v=1}^{V} & tr\left(\mathbf{X}^{(v)}\mathbf{D}_1^{(v)}\mathbf{X}^{(v)T}\right) - 2tr\left(\mathbf{D}_1^{(v)}\mathbf{X}^{(v)T}\mathbf{W}^{(v)}\mathbf{H}^{(v)}\right) \\ & + tr\left(\mathbf{D}_1^{(v)}\mathbf{H}^{(v)T}\mathbf{W}^{(v)T}\mathbf{W}^{(v)}\mathbf{H}^{(v)}\right) \\ & + \alpha tr\left(\mathbf{H}^{(v)}\mathbf{R}^{(v)}\mathbf{D}_2^{(v)}\mathbf{R}^{(v)T}\mathbf{H}^{(v)T}\right) + tr\left(\mathbf{\Theta}\mathbf{H}^{(v)T}\right), \end{aligned} \tag{6}$$

where $\mathbf{D}_1^{(v)}$ and $\mathbf{D}_2^{(v)}$ are diagonal matrices, which are $n \times n$ matrices, and the elements defined as

$$\begin{aligned} \left(\mathbf{D}_1^{(v)}\right)_{ii} &= \frac{1}{\left\|\mathbf{X}_i^{(v)} - \mathbf{W}^{(v)}\mathbf{H}_i^{(v)}\right\|}, \\ \left(\mathbf{D}_2^{(v)}\right)_{ii} &= \frac{1}{\left\|\left(\mathbf{H}^{(v)}\mathbf{R}^{(v)}\right)_i\right\|}. \end{aligned} \tag{7}$$

The partial derivative of Lagrangian function $L(\mathbf{H}^{(v)})$ with respect to $\mathbf{H}^{(v)}$ is computed as follows:

$$\begin{aligned} \frac{\partial L\left(\mathbf{H}^{(v)}\right)}{\partial \mathbf{H}^{(v)}} = & \left(-2\mathbf{W}^{(v)T}\mathbf{X}^{(v)}\mathbf{D}_1^{(v)T} + 2\mathbf{W}^{(v)T}\mathbf{W}^{(v)}\mathbf{H}^{(v)}\mathbf{D}_1^{(v)T}\right) \\ & + 2\alpha\left(\mathbf{H}^{(v)}\mathbf{R}^{(v)}\mathbf{D}_2^{(v)}\mathbf{R}^{(v)T}\right) + \mathbf{\Theta}. \end{aligned} \tag{8}$$

Because $\mathbf{R}^{(v)}$ is mixed-sign data, we should decompose it into two nonnegative parts $\mathbf{M}^{+}$ and $\mathbf{M}^{-}$, representing the positive part and the negative part, respectively,

$$\begin{aligned} \mathbf{M}^{+} &= \frac{|\mathbf{M}| + \mathbf{M}}{2}, \\ \mathbf{M}^{-} &= \frac{|\mathbf{M}| - \mathbf{M}}{2}. \end{aligned} \tag{9}$$

Let $\mathbf{R}_{\mathrm{a}}^{(v)} = \mathbf{R}^{(v)+}\mathbf{D}^{(v)}(\mathbf{R}^{(v)-})^T$, $\mathbf{R}_{\mathrm{b}}^{(v)} = \mathbf{R}^{(v)-}\mathbf{D}^{(v)}(\mathbf{R}^{(v)+})^T$, $\mathbf{R}_{\mathrm{c}}^{(v)} = \mathbf{R}^{(v)+}\mathbf{D}^{(v)}(\mathbf{R}^{(v)+})^T$, and $\mathbf{R}_{\mathrm{d}}^{(v)} = \mathbf{R}^{(v)-}\mathbf{D}^{(v)}(\mathbf{R}^{(v)-})^T$. The updating rule of $\mathbf{H}^{(v)}$ is formulated as follows:

$$\mathbf{H}^{(v)} \longleftarrow \mathbf{H}^{(v)} \cdot \sqrt{\frac{\mathbf{W}^{(v)T}\mathbf{X}^{(v)}\mathbf{D}_1^{(v)T} + \alpha\mathbf{H}^{(v)}\left(\mathbf{R}_{\mathrm{a}}^{(v)} + \mathbf{R}_{\mathrm{b}}^{(v)}\right)}{\mathbf{W}^{(v)T}\mathbf{W}^{(v)}\mathbf{H}^{(v)}\mathbf{D}_1^{(v)T} + \alpha\mathbf{H}^{(v)}\left(\mathbf{R}_{\mathrm{c}}^{(v)} + \mathbf{R}_{\mathrm{d}}^{(v)}\right)}}. \tag{10}$$

Computing $\mathbf{W}^{(v)}$, to update $\mathbf{W}^{(v)}$ with $\mathbf{H}^{(v)}$ fixed, the following object function should be solved:

$$\min_{\mathbf{W}^{(v)}\geq 0} J\left(\mathbf{W}^{(v)}\right) = \left\|\mathbf{X}^{(v)} - \mathbf{W}^{(v)}\mathbf{H}^{(v)}\right\|_{2,1}. \tag{11}$$

This is similar as that in [19, 24]. So we have the updating rule as

$$\mathbf{W}^{(v)} \longleftarrow \mathbf{W}^{(v)} \cdot \frac{\mathbf{X}^{(v)}\mathbf{D}_1^{(v)}\mathbf{H}^{(v)T}}{\mathbf{W}^{(v)}\mathbf{H}^{(v)}\mathbf{D}_1^{(v)}\mathbf{H}^{(v)T}}, \tag{12}$$

where $(\cdot)$ indicates the Hadamard product. For each view, the updating rule of $\mathbf{H}^{(v)}$ and $\mathbf{W}^{(v)}$ satisfies the theorem in [24], which guarantees the correctness of the rule. The correctness analysis and convergence proof of (10) and (12) are shown based on the method [19] in the following section.

3.3. Correctness and Convergence

Theorem 1. *If the updating rule of $\mathbf{H}^{(v)}$ converges, then the final solution satisfies the KKT optimality condition.*

Proof. At convergence, $\mathbf{H}^{(v)\infty} = \mathbf{H}^{(v)t+1} = \mathbf{H}^{(v)t} = \mathbf{H}^{(v)}$, where t denotes the t-th iteration, and the following formula holds:

$$\left(\mathbf{W}^{(v)T}\mathbf{X}^{(v)}\mathbf{D}_1^{(v)T} - \mathbf{W}^{(v)T}\mathbf{W}^{(v)}\mathbf{H}^{(v)}\mathbf{D}_1^{(v)T} + \alpha\left[\mathbf{H}^{(v)}\mathbf{R}^{(v)}\mathbf{D}_2^{(v)}\mathbf{R}^{(v)T}\right]^{-}\right) + \alpha\left[\mathbf{H}^{(v)}\mathbf{R}^{(v)}\mathbf{D}_2^{(v)T}\right]^{+}\Big)_{ij}\left(\mathbf{H}_{ij}^{(v)}\right)^2 = 0. \tag{13}$$

□

We now prove the convergence of the updating rule (10) using the auxiliary function approach in [19]. The definition of the auxiliary function is as follows.

Definition 1. $G(\mathbf{H}, \mathbf{H}')$ is an auxiliary function for $J(\mathbf{H})$ if $G(\mathbf{H}, \mathbf{H}') \geq J(\mathbf{H})$ and $G(\mathbf{H}, \mathbf{H}) = J(\mathbf{H})$ hold for any $\mathbf{H}$ and a constant matrix $\mathbf{H}'$.

The auxiliary function is useful because of the following Lemma 1.

Lemma 1. *J is nonincreasing under the updating rule $\mathbf{H}^{t+1} = \operatorname{argmin}_{\mathbf{H}} G(\mathbf{H}, \mathbf{H}^t)$, if G is an auxiliary function of J.*

Proof. Following the definition of $G(\mathbf{H}, \mathbf{H}')$, we have

$$J\left(\mathbf{H}^{t+1}\right) \leq G\left(\mathbf{H}^{t+1}, \mathbf{H}^t\right) \leq G\left(\mathbf{H}^t, \mathbf{H}^t\right) = J\left(\mathbf{H}^t\right). \tag{14}$$

□

The key point is to find an appropriate auxiliary function for (6). Because the learning process is independent in each view, in generally. Let $J(\mathbf{H})$ represents $J(\mathbf{H}^{(v)})$ to indicate each view's process, and let $\mathbf{B} = \mathbf{D}_1\mathbf{X}^T\mathbf{W}$, $\mathbf{A} = \mathbf{W}^T\mathbf{W}$, and $\mathbf{C} = \mathbf{R}\mathbf{D}_2\mathbf{R}^T$. We rewrite (6) as follows:

$$J(\mathbf{H}) = tr\left(\mathbf{X}^T\mathbf{D}_1\mathbf{X} - 2\mathbf{B}\mathbf{H} + \mathbf{D}_1\mathbf{H}^T\mathbf{A}\mathbf{H}\right) + \alpha tr\left(\mathbf{H}\mathbf{C}\mathbf{H}^T\right). \tag{15}$$

Since the update rules are elementwise, we should prove each $\mathbf{H}$ is nonincreasing under the update (10) by defining the auxiliary function regarding $\mathbf{H}_{ik}$ as follows.

Lemma 2. *The function*

$$G(\mathbf{H}, \mathbf{H}') = \left(\sum_{ik}\mathbf{X}_{ik}^T(\mathbf{D}_1\mathbf{X})_{ik} - 2\sum_{ik}\mathbf{B}_{ik}\mathbf{H}'_{ik}\left(1 + \log\frac{\mathbf{H}_{ik}}{\mathbf{H}'_{ik}}\right) + \sum_{ik}\left(\mathbf{A}\mathbf{H}'\mathbf{D}_1^T\right)_{ik}\frac{\mathbf{H}_{ik}^2}{\mathbf{H}'_{ik}} + \alpha\sum_{ik}\left(\mathbf{C}\mathbf{H}'\right)_{ik}\frac{\mathbf{H}_{ik}^2}{\mathbf{H}'_{ik}}\right), \tag{16}$$

is an auxiliary function of $J(\mathbf{H})$ in problem (6).

Proof. Since $G(\mathbf{H}, \mathbf{H}) = J(\mathbf{H})$ is obvious, we need to prove $G(\mathbf{H}, \mathbf{H}') \geq J(\mathbf{H})$. To this end, we compare (16) with (15). For the inequality $z \geq 1 + \log z$, which holds when $z > 0$, we have the following inequalities:

$$tr\left(\mathbf{H}^T\mathbf{B}\right) = \sum_{ik}\mathbf{B}_{ik}\mathbf{H}_{ik} \geq \sum_{ik}\mathbf{B}_{ik}\mathbf{H}'_{ik}\left(1 + \log\frac{\mathbf{H}_{ik}}{\mathbf{H}'_{ik}}\right). \tag{17}$$

□

With the lemma and proposition in [19], we have the following inequalities:

$$tr\left(\mathbf{D}_1\mathbf{H}^T\mathbf{A}\mathbf{H}\right) \leq \sum_{ik}\left(\mathbf{A}\mathbf{H}'\mathbf{D}_1^T\right)_{ik}\frac{\mathbf{H}_{ik}^2}{\mathbf{H}'_{ik}},$$
$$tr\left(\mathbf{H}\mathbf{C}\mathbf{H}^T\right) \leq \sum_{ik}\left(\mathbf{C}\mathbf{H}'\right)_{ik}\frac{\mathbf{H}_{ik}^2}{\mathbf{H}'_{ik}}. \tag{18}$$

Collecting all bounds, $G(\mathbf{H}, \mathbf{H}') \geq J(\mathbf{H})$ holds, and Lemma 2 is proven.

Theorem 2. *Problem (6) is nonincreasing under the iterative updating rule (10).*

Proof. $G(H, H')$ is a convex function. To find its minima, following the KKT condition, we let

$$\frac{\partial G(\mathbf{H}, \mathbf{H}')}{\partial \mathbf{H}_{ik}} = 2\left(-\mathbf{B}_{ik}^{+}\frac{\mathbf{H}'_{ik}}{\mathbf{H}'_{ik}} + \mathbf{B}_{ik}^{-}\frac{\mathbf{H}_{ik}}{\mathbf{H}'_{ik}} + \left(\mathbf{A}^{+}\mathbf{H}'\right)_{ik}\frac{\mathbf{H}'_{ik}}{\mathbf{H}_{ik}}\right) - \left(\mathbf{A}^{-}\mathbf{H}'\right)_{ik} \cdot \frac{\mathbf{H}'_{ik}}{\mathbf{H}'_{ik}} + 2\alpha\left(\left(\mathbf{C}^{+}\mathbf{H}'\right)_{ik}\frac{\mathbf{H}'_{ik}}{\mathbf{H}_{ik}} - \left(\mathbf{C}^{+}\mathbf{H}'\right)_{ik}\frac{\mathbf{H}_{ik}}{\mathbf{H}'_{ik}}\right) = 0. \tag{19}$$

□

This can derive the updating rule (10) under the objective function $J(\mathbf{H})$ in (6). This updating rule of $\mathbf{W}^{(v)}$ also can be derived by this method.

3.4. MVC with rNNMF

3.4.1. Representation Combining. After decomposed by the matrix factorizations model, the final representation can be obtained by combining different views' representation. The final output $\mathbf{H}$ which is a $(V^*p) \times n$ matrix can be obtained from $\mathbf{H}^{(v)}$ of multiview as

$$\mathbf{H} \longleftarrow \left[\mathbf{H}^{(1)}; \ldots; \mathbf{H}^{(v)}; \ldots; \mathbf{H}^{(V)}\right]. \tag{20}$$

3.4.2. Clustering with Similarity Graph. Because the neighboring structure is kept during the factorization, the graph-based clustering method will be chosen to cluster data in our study. The similarity graph G is built from the final representation $\mathbf{H}$ by the k-NN algorithm. Then, normalized cut (Ncut) [22] is used to obtain the final clustering results. It can achieve better performance considering the graph structure of the data. Details of our method are described in Algorithm 1.

3.5. Complexity Analysis. Suppose that the dimensions in all the input views are the same, denoted by d. The dimension of

Input: input $\mathbf{X}^{(v)}$, parameter α, parameter k
Initialize:
  for each view **do**
    $(\mathbf{W}^{(v)}, \mathbf{H}^{(v)}) \longleftarrow \mathrm{NNDSVD}(\mathbf{X}^{(v)})$
  end
design $\mathbf{R}^{(v)}$ via (2) with $\mathbf{X}^{(v)}$
while not converged **do**
  for all view **do**
    $\mathbf{W}^{(v)}$ update via (12)
    $\mathbf{H}^{(v)}$ update via (10)
    $\mathbf{R}^{(v)}$ update via (2) with $\mathbf{H}^{(v)}$
  end
end
combining: $\mathbf{H} \longleftarrow [\mathbf{H}^{(1)}; \cdots; \mathbf{H}^{(v)}; \cdots; \mathbf{H}^{(V)}]$
Similarity graph: $\mathbf{G}$ is built with $\mathbf{H}$ and k
clustering: Ncut ($\mathbf{G}$)
Output: clustering results

Algorithm 1: rNNMF-based MVC model.

$\mathbf{H}^{(v)}$ is denoted by p. And V is the total number of views. The overall cost of $\mathbf{H}^{(v)}$ in each view is $O(n^3 + \mathrm{pnd})$. Updating $\mathbf{R}^{(v)}$ costs $O(\mathrm{pn}^2 - \mathrm{pn})$, and the cost of $\mathbf{W}^{(v)}$ is $O(\mathrm{pn}^2 + \mathrm{pnd})$. Furthermore, each of $\mathbf{R}_{\mathrm{a}}^{(v)}$, $\mathbf{R}_{\mathrm{b}}^{(v)}$, $\mathbf{R}_{\mathrm{c}}^{(v)}$, and $\mathbf{R}_{\mathrm{d}}^{(v)}$ is $O(n^3)$. Note t is the number of iterations. The overall complexity is $O(Vt(2\mathrm{pnd} + n^3 + 2\mathrm{pn}^2 - \mathrm{pn}) + 4Vn^3)$. Nevertheless, $\mathbf{R}$ (containing $2n$ non-zero elements) is a sparse matrix. The complexity will be reduced obviously.

4. Experiment

4.1. Experimental Setting

4.1.1. Datasets. Four datasets are used in the experiment.

UCI Digit (https://archive.ics.uci.edu/ml/datasets/Multiple+Features) is a dataset of handwritten digits of 0 to 9 from UCI machine learning repository. It consists of 2000 points. Similarly to the work in [21], we use 76 Fourier coefficients and 216 profile correlations.

3Sources (http://mlg.ucd.ie/datasets/3sources.html) is collected from three well-known online news sources and each is treated as one view. We select the 169 stories which are reported in all three sources.

ORL (http://cs.tju.edu.cn/faculty/zhangchangqing/code/ORL_mtv.rar) contains 400 different images of 40 subjects with three views: intensity, LBP, and Gabor. All images are resized into 48×48. LBP is a 59-dimension histogram over 9×10 pixel patches generated from cropped images. The scale parameter λ in Gabor wavelets is fixed as 4 at four orientations $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ with a cropped image of size $25^\circ \times 30$ pixels.

Washington (http://www.cs.umd.edu/projects/linqs/projects/lbc/) belongs to WebKB, which collected webpages from four universities. The webpages are distributed over five classes: student, project, course, staff, and faculty, and they are described by two views: the content view and the citation view. Each webpage is described by 1703 words in the content view and the number of citation links between other pages in the citation view. We summarize them in Table 1.

The algorithms that we employ to compare are as follows: (1) BestSV performs the best performance in each view [25]; NMF-based methods: (2) MulNMF and (3) D-SNMF; subspace clustering based methods: (4) c-LRSSC [26], (5) p-LRSSC [26], (6) RMSC [27], (7) ECMSC [28], and (8) MVGL [29].

The codes of all the baseline methods are provided by their authors. We adjust the parameters of all comparison methods according to the corresponding literature to obtain their best performance. For RMSC, its parameter is searched from 0.005 to 100 as the authors' suggestion. For all the NMF-based methods, we set the dimensionality of the new space to be the same as or bigger than the number of clusters and the initial step follows the authors' suggestion. K-means will be applied to the new representation for clustering. This process is repeated 10 times, and the average clustering performance is recorded as the final result.

For rNNMF, the dimensionality of $\mathbf{H}^{(v)}$ is 60 for UCI and ORL, and it is 20 for 3Sources and Washington. It is initialized by NNDSVD and repeated five times for average results. The hot kernel, in which the parameter σ is 2, is used to determine the distance between points to select the neighboring points as same as in [30] (https://github.com/louloupiano/PCPSNMF).

For evaluation, we use three evaluation metrics: accuracy (ACC), normalized mutual information (NMI), and adjusted Rand index (AR) [31]. For all the metrics, a high value denotes good performance.

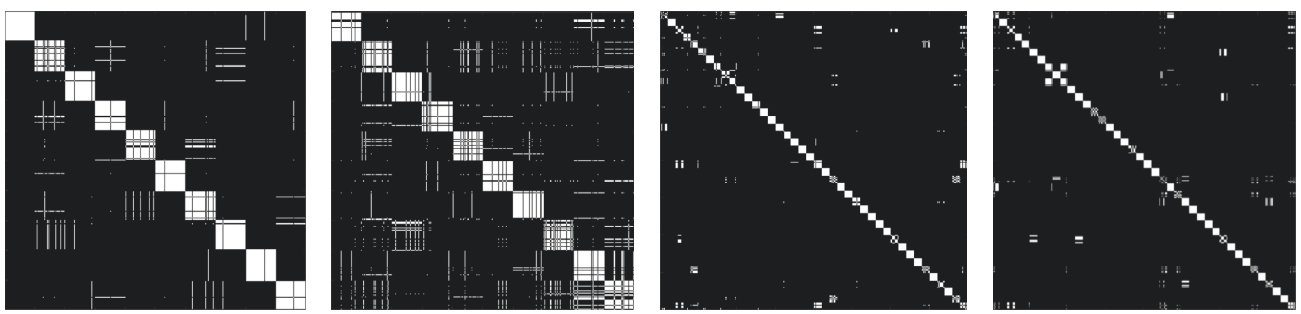
4.2. Clustering Performance. Table 2 summarizes the clustering performance. The best values are in bold. As the table shows, rNNMF achieves the highest performance. In UCI, with settings $\sigma = 0.1$ and $k = 5$, it outperforms the second best method by roughly 6.78%, 9.07%, and 12.49% in the three metrics. In 3Sources, it also obtains the best with settings $\sigma = 0.05$ and $k = 5$. In ORL, when $\sigma = 0.05$ and $k = 5$, it also outperforms the best in ACC and AR, while becomes worse slightly in NMI than ECMSC. In Washington, with $\sigma = 0.001$ and $k = 7$, it also outperforms the second best method by

Table 1: Information of datasets in the experiments.

Features	UCI	3Sources	ORL	Washington
1	Fourier	Bbc	Intensity	Content
2	Profile	Guardian	LBP	Cites
3	—	Reuters	Gabor	—
Of data	2000	169	400	230
Of classes	**10**	**6**	**40**	**5**

Table 2: Clustering results ((mean) %) on different datasets.

		BestSV	MulNMF	D-SNMF	c-LRSSC	p-LRSSC	RMSC	ECMSC	MVGL	rNNMF
UCI	ACC	65.40	86.78	74.91	83.54	83.43	78.48	80.95	82.10	**93.56**
	NMI	61.34	78.20	80.86	85.14	85.12	71.70	83.61	83.39	**88.13**
	AR	50.19	74.11	71.26	81.79	81.09	64.23	76.63	76.70	**86.60**
3Sources	ACC	53.49	65.68	42.90	70.36	67.10	60.12	66.86	75.74	**76.21**
	NMI	45.23	60.00	29.48	60.02	58.71	51.47	61.46	60.49	**62.76**
	AR	33.47	50.76	19.51	57.09	53.12	39.33	54.20	53.93	**63.93**
ORL	ACC	71.56	77.50	74.22	74.31	70.60	72.55	81.50	72.75	**84.80**
	NMI	88.09	86.58	86.41	89.24	85.52	87.24	**92.71**	85.14	92.57
	AR	64.48	64.46	63.96	68.17	60.47	65.41	76.49	47.01	**80.27**
Washington	ACC	59.00	57.39	41.74	45.78	47.15	58.30	57.39	64.35	**66.78**
	NMI	22.45	23.52	10.62	32.72	32.58	28.80	30.71	33.00	**33.77**
	AR	27.86	32.21	8.96	24.84	25.44	33.47	33.00	31.57	**44.34**



(a) (b) (c) (d)

Figure 1: Visual comparison of the similarity matrices produced on UCI and ORL by 2 high ACC performance methods. (a) rNNMF on UCI. (b) MulNMF on UCI. (c) rNNMF on ORL. (d) ECMSC on ORL.

roughly 2.43%, 0.77%, and 12.87%. In all, rNNMF achieves more accurate performance than others obviously.

We present Figure 1 to show more details on clustering results with the similarity matrix yielded from two high-performance MVC methods in UCI and ORL. For UCI, we can see that diagonal blocks of rNNMF are whiter than MulNMF, and the surrounding nondiagonal black blocks are blacker than MulNMF. For ORL, the similar conclusions also hold, which is clearer in the similarity matrix of rNNMF.

Hence, with the representation combining process, rNNMF can fuse multiple views efficiently. And neighboring constraint plays an important role in discovering the underlying structure of points. It is observed that a comprehensive graph structure is important to discover the cluster structure.

4.3. Influence of the Parameters. σ plays an important role for representation learning process in rNNMF. We test σ within the setting [0, 0.0001, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1] for all datasets. Among them, $\sigma = 0$ means there is no neighboring constraint in our model. The ACC and NMI results are shown in Figure 2.

It shows that ACC performance and NMI performance improve with increasing σ in UCI, and they will reach their best performance when $\sigma = 0.1$. Then, the performance drops obviously. This tendency can also be observed in 3Sources. The only difference is that the best ACC and NMI can be obtained when $\sigma = 0.05$. However, in ORL dataset, performances improve slightly with increasing σ, and they drop obviously after reaching the best value. ACC and NMI in Washington will drop with increasing σ. This shows that too large σ can destroy the similarity structure in each view, which can lead to worse performance in final clustering process. And the suitable σ can strengthen the cluster structure in learning process. Furthermore, when $\sigma = 0$, ACC and NMI performances in our model are better than those in some methods, such as D-SNMF and RMSC,

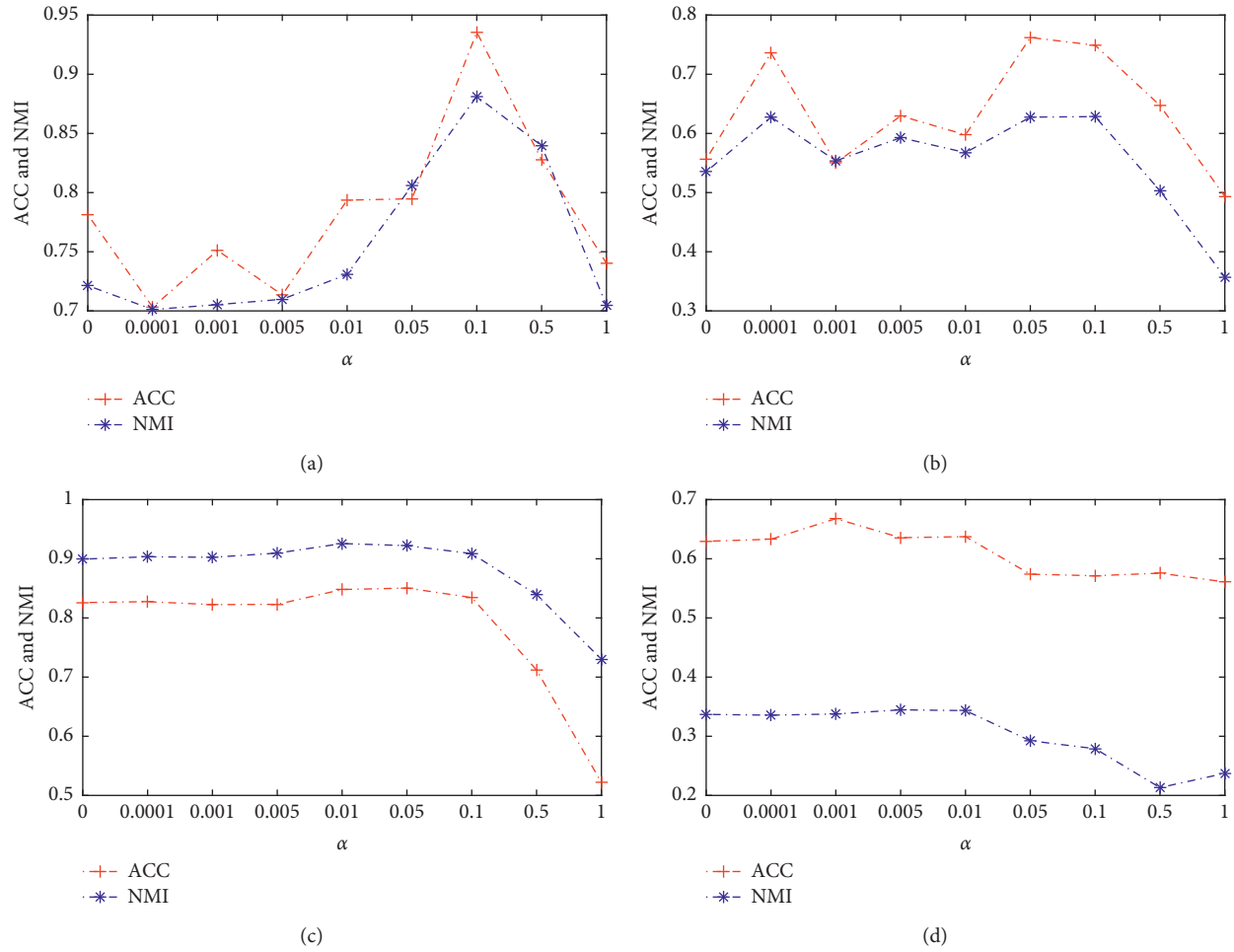


Figure 2: ACC and NMI with parameter σ on four datasets. (a) UCI. (b) 3Sources. (c) ORL. (d) Washington.

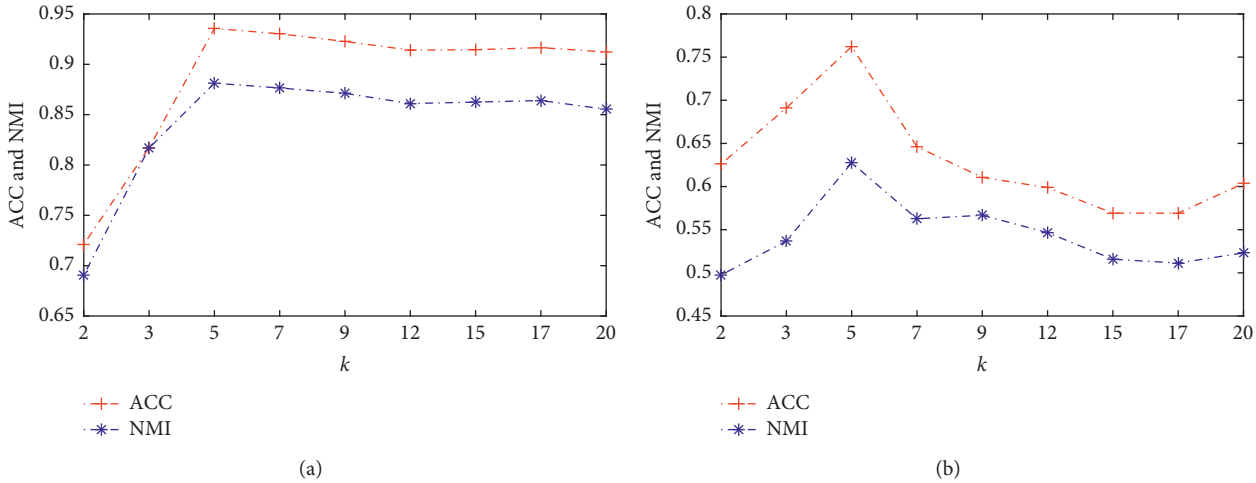


Figure 3: Continued.

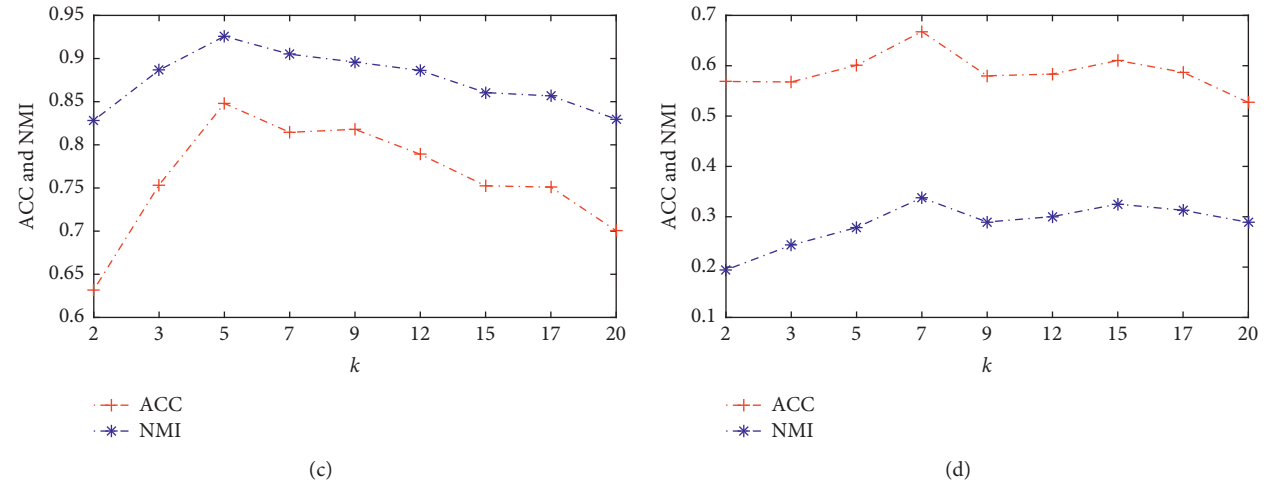


Figure 3: ACC and NMI with parameter k on four datasets. (a) UCI. (b) 3Sources. (c) ORL. (d) Washington.

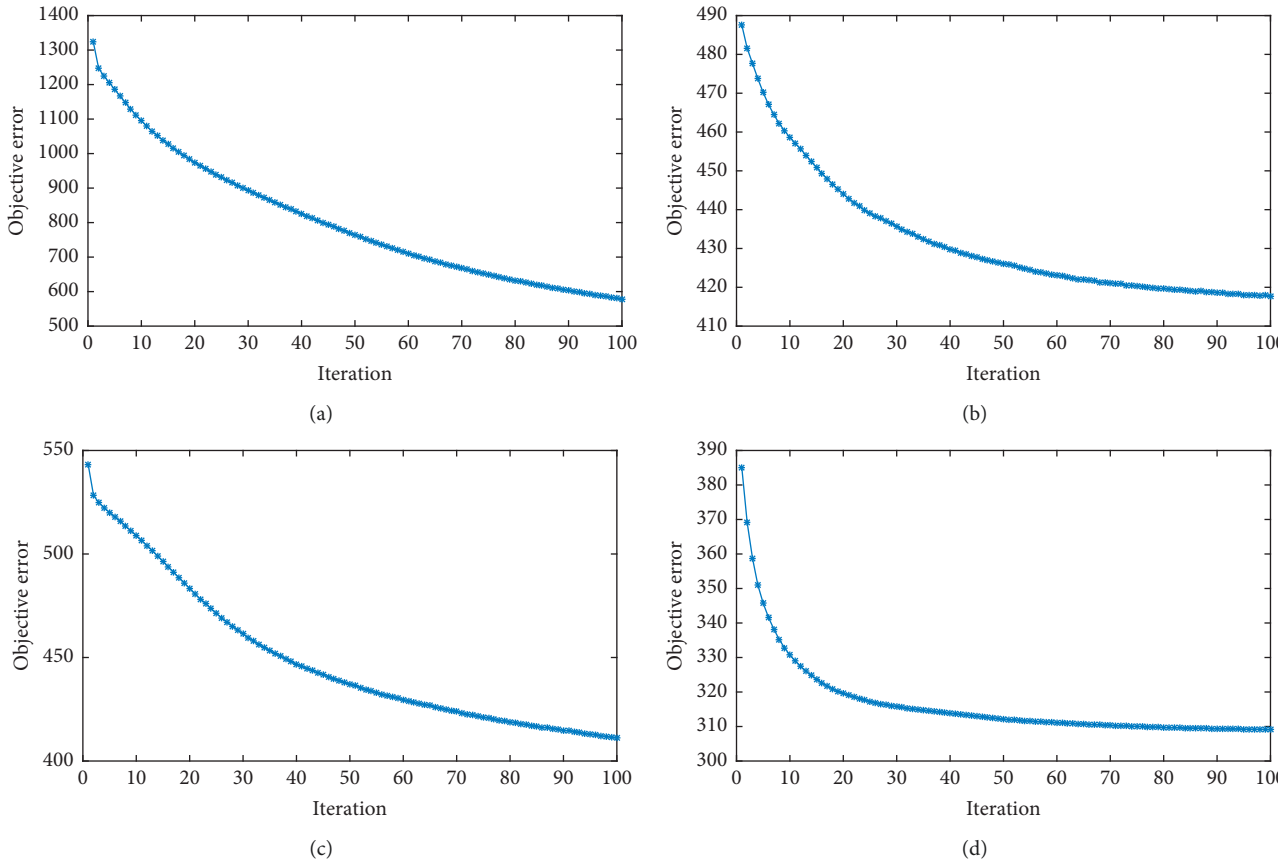


Figure 4: Objective error values on four datasets. (a) UCI. (b) 3Sources. (c) ORL. (d) Washington.

demonstrating that combining multiviews and building graph is useful in MVC.

Parameter k is important to create the final similarity graph. Figure 3 shows the ACC and NMI results with different k values. As we can see, the final results are sensitive with k in integrated representation. In UCI, 3Sources, and ORL, obviously, the best performance can be obtained when $k = 5$, and then, the tendency of performance will drop as k increases. In Washington, ACC and NMI reach their best when $k = 7$. This shows that the influence of k mainly focuses on similarity graph building from integrated representation, and it is important to build the graph from representation.

4.4. Convergence Analysis. Figure 4 shows the convergence property of rNNMF by computing the objective error in

each iteration. It is clear that the objective value decreases steadily in all datasets. All of NMI finally keeps the rough stability around the convergence point. So the maximum number of iterations is set to 100 for all the experiments.

5. Conclusion

In this paper, we proposed a novel NMF-based MVC model, named "rNNMF." In this model, the neighbor structure representation can be learned in each view, and $L_{2,1}$-norm-based loss function is designed to improve its robustness against noises and outliers. Then, a final representation of data was integrated with those representations of all views, and a graph was learned from this representation. Finally, the graph cut method was used to partition data into its underlying clusters. Unlike existing methods, rNNMF can well encode the local structure from each view feature space and achieve the structure agreement via combining fusion. Experiments show that the rNNMF-based model yields higher performance. One of the important further works is to find a better graph structure to obtain more clear representation. In addition, the weight for each view is also worth studying to deal with varying levels of quality. The future work can also be done to extend rNNMF model and optimization strategy to handle dynamic data and achieve online multiview clustering.

Data Availability

All the data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Key Research Development Program of China under Grant no. 2017YFB0802800 and in part by the National Natural Science Foundation of China under Grant no. 61473149.

References

[1] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4833–4843, 2018.

[2] S. Wang, E. K. Wang, X. Li, Y. Ye, R. Y. K. Lau, and X. Du, "Multi-view learning via multiple graph regularized generative model," *Knowledge-Based Systems*, vol. 121, pp. 153–162, 2017.

[3] K. Zhan, C. Niu, C. Chen, F. Nie, C. Zhang, and Y. Yang, "Graph structure fusion for multiview clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1984–1993, 2019.

[4] C. Li, J. Bai, and G. D. Hager, "A unified framework for multi-view multi-class object pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 254–269, Munich, Germany, September 2018.

[5] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, and X. Huang, "Robust subspace clustering for multi-view data by exploiting correlation consensus," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3939–3949, 2015.

[6] Y. Wang, W. Zhang, W. Lin et al., "Iterative views agreement: an iterative low-rank based structured optimization method to multi-view spectral clustering," in *Proceedings of the IJCAI*, pp. 2153–2159, New York, NY, USA, July 2016.

[7] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 556–562, Vancouver, BC, Canada, December 2001.

[8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[9] P. Luo, J. Peng, Z. Guan, and J. Fan, "Dual regularized multi-view non-negative matrix factorization for clustering," *Neurocomputing*, vol. 294, pp. 1–11, 2018.

[10] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[11] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using Laplacian faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[12] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[13] K. Zhan, J. Shi, J. Wang, H. Wang, and Y. Xie, "Adaptive structure concept factorization for multiview clustering," *Neural Computation*, vol. 30, no. 4, pp. 1080–1103, 2018.

[14] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proceedings of the 2013 SIAM International Conference on Data Mining*, pp. 252–260, Austin, TX, USA, January 2013.

[15] W. Han, G. Li, and X. Zhang, "Accurately detecting community with large attribute in partial networks," in *Lecture Notes in Computer Science*, pp. 643–657, Springer Science+Business Media, Berlin, Germany, 2018.

[16] M. M. Kalayeh, H. Idrees, M. Shah, and NMF-KNN, "Image annotation using weighted multi-view non-negative matrix factorization," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 184–191, Columbus, OH, USA, June 2014.

[17] W. Ou, S. Yu, G. Li, J. Lu, K. Zhang, and G. Xie, "Multi-view non-negative matrix factorization by patch alignment framework with view consistency," *Neurocomputing*, vol. 204, pp. 116–124, 2016.

[18] H. Zhao, Z. Ding, and Y. Fu, "Multi view clustering via deep matrix factorization," in *Proceedings of the AAAI*, pp. 2921–2927, San Francisco, CA, USA, February 2017.

[19] J. Wang, F. Tian, C. H. Liu, H. Yu, X. Wang, and X. Tang, "Robust nonnegative matrix factorization with ordered structure constraints," in *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 478–485, Anchorage, AK, USA, May 2017.

[20] F. Shang, L. C. Jiao, and F. Wang, "Graph dual regularization non-negative matrix factorization for co-clustering," *Pattern Recognition*, vol. 45, no. 6, pp. 2237–2250, 2012.

[21] D. Tolic´, N. Antulov-Fantulin, and I. Kopriva, "A nonlinear orthogonal non-negative matrix factorization approach to subspace clustering," *Pattern Recognition*, vol. 82, pp. 40–55, 2018.

[22] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[23] J. Wang, F. Tian, H. Yu, C. H. Liu, K. Zhan, and X. Wang, "Diverse non-negative matrix factorization for multiview data representation," *IEEE Transactions on Cybernetics*, vol. 48, no. 9, pp. 2620–2632, 2018.

[24] C. H. Q. Ding, T. Tao Li, M. I. Jordan et al., "Convex and semi-nonnegative matrix factorizations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 45–55, 2010.

[25] N. Jordan and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 849–856, Denver, CO, USA, November 2001.

[26] M. Brbi and I. Kopriva, "Multi-view low-rank sparse subspace clustering," *Pattern Recognition*, vol. 73, pp. 247–258, 2018.

[27] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *Proceedings of the AAAI*, pp. 2149–2155, Québec City, Québec, Canada, July 2014.

[28] X. Wang, X. Guo, Z. Lei et al., "Exclusivity-Consistency regularized multi-view subspace clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 923–931, Honolulu, HI, USA, June 2017.

[29] K. Zhan, C. Zhang, J. Guan, and J. Wang, "Graph learning for multiview clustering," *IEEE Transactions on Cybernetics*, vol. 48, no. 10, pp. 2887–2895, 2018.

[30] W. Wu, Y. Jia, S. Kwong, and J. Hou, "Pairwise constraint propagation-induced symmetric nonnegative matrix factorization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 12, pp. 6348–6361, 2018.

[31] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao, "Low-rank tensor constrained multiview subspace clustering," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1582–1590, Santiago, Chile, December 2015.