Adaptive Structure Concept Factorization for Multiview Clustering

Kun Zhan
ice.echo@gmail.com
Jinhui Shi
shijinhui10@gmail.com
School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu 730000, China

Jing Wang
jwang@bournemouth.ac.uk
Faculty of Science and Technology, Bournemouth University, Bournemouth BH125BB, U.K.

Haibo Wang
wanghb15@lzu.edu.cn
Yuange Xie
xieyg15@lzu.edu.cn
School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu 730000, China

Most existing multiview clustering methods require that graph matrices in different views are computed beforehand and that each graph is obtained independently. However, this requirement ignores the correlation between multiple views. In this letter, we tackle the problem of multiview clustering by jointly optimizing the graph matrix to make full use of the data correlation between views. With the interview correlation, a concept factorization–based multiview clustering method is developed for data integration, and the adaptive method correlates the affinity weights of all views. This method differs from nonnegative matrix factorization–based clustering methods in that it can be applicable to data sets containing negative values. Experiments are conducted to demonstrate the effectiveness of the proposed method in comparison with state-of-the-art approaches in terms of accuracy, normalized mutual information, and purity.

1 Introduction

In data analysis, instances are often represented in heterogeneous views. For example, an image is represented by various feature extractors; a web

Neural Computation **30, 1080–1103** (2018)
doi:10.1162/NECO_a_01055

page is described by the words on the page and the words in hyperlink that point to the page; a user's information is fused and analyzed from different social networks (Jia et al., 2016); and a video includes dynamic images, sound, and subtitles (Yang et al., 2012; Yang, Zhang, & Xu, 2015; Yan et al., 2016) Multiview learning uses the correlations between views to obtain higher performance than using any single-view features (Blum & Mitchell, 1998; Bickel & Scheffer, 2004; Kakade & Foster, 2007; Zhan, Zhang, Guan, & Wang, 2017).

Multiview clustering starts with a series of works on cotraining methods. Cotraining methods train models separately on each view and iteratively learn for each model through the exploitation of disagreement between models (Blum & Mitchell, 1998); the reasons for the success of cotraining methods have been investigated by Balcan, Blum, and Yang (2004) and Wang and Zhou (2010). Spectral clustering is one of the most popular clustering approaches. Taking advantage of the well-defined mathematical framework of spectral clustering (Shi & Malik, 2000; Ng, Jordan, & Weiss, 2002; Zelnik-Manor & Perona, 2004; Von Luxburg, 2007; Yang, Xu, Nie, Yan, & Zhuang, 2010), many multiview clustering methods are proposed (Blaschko & Lampert, 2008; Kumar & Daumé, 2011; Kumar, Rai, & Daume, 2011; Cai, Nie, Huang, & Kamangar, 2011; Xia, Pan, Du, & Yin, 2014; Li, Nie, Huang, & Huang, 2015). However, the drawbacks of these spectral clustering methods are that the performance of these methods highly depends on the precomputed affinity graph matrix, involves time-consuming calculation of eigenvectors of high-dimensional matrices, and the eigenvectors obtained have no direct relationship to the semantic structure of the data sets. Nonnegative matrix factorization (NMF) methods have recently been applied to multiview clustering with impressive results (Liu, Wang, Gao, & Han, 2013; Zhang, Zhao, Zong, Liu, & Yu, 2014) because the results of NMF-based clustering approaches have better semantic interpretation (Xu, Liu, & Gong, 2003; Xu & Gong, 2004; Ding, He, & Simon, 2005) and these NMF-based methods can be implemented by novel multiplicative update rules. However, a limitation of these NMF-based methods is that they are not applicable to data sets containing negative values.

Concept factorization (CF), a variant of NMF, can be used to process arbitrary data sets even though they have negative values, and CF inherits the advantage of the multiplicative update rules of NMF. Using these two advantages of CF, we apply an adaptive CF-based method to multiview clustering in this letter. We use an adaptive graph term to capture the local intrinsic geometrical structure of the data space (Cai, He, & Han, 2011), and the similarity between the data points is measured based on the new representations. We take all the data points in each view into consideration to optimize elements of the graph matrix in a global view by assuming that there is a larger probability that data points with a small distance between them will be neighbors. Our algorithm uses novel update rules to effectively find a solution to a well-designed optimization problem. A convergence analysis is also provided. Extensive empirical results on nine data sets show that the proposed multiview clustering method achieves better clustering results than state-of-the-art approaches.

This letter makes the following contributions. First, The proposed method jointly optimizes the graph matrix to make full use of the data correlation between views for multiview clustering. The novelty lies in learning one affinity graph from multiview data to address the correlation between views and avoid exploit construction of single graphs. Second, the proposed method can process arbitrary data sets even though they contain negative values, and the CF-based method has a better semantic interpretation (Xu et al., 2003; Xu & Gong, 2004; Ding et al., 2005) than spectral clustering-based methods. Finally, we propose a multiview clustering algorithm that combines concept factorization and locality-preserving methods in a unified optimization problem and solves this hard optimization problem with alternating optimization. The effectiveness of the algorithm is evaluated on nine data sets for the multiview clustering problem.

The remainder of the letter is organized as follows. In section 2, we propose an adaptive graph-regularized multiview concept factorization algorithm. We incorporate the correlation among multiple views to improve the performance of existing concept factorization clustering algorithms by jointly optimizing the graph matrix. In section 3, we propose a novel algorithm to optimize the well-designed objective function in section 2. In section 4, we present numerical experiments and comparison results. We use nine data sets and compare them with seven state-of-the-art methods. Section 5 concludes with some discussion.

2 Multiview Concept Factorization

Let $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ denote the data matrix. Each data point $\boldsymbol{x}_i$ is represented by a d-dimensional feature vector. NMF aims to find a $d \times k$ nonnegative matrix $\mathbf{U}$ and a $k \times n$ nonnegative matrix $\mathbf{H}$ where the product of the two factors is an approximation to $\mathbf{X}$, represented as $\mathbf{X} \approx \mathbf{UH}$ (Lee & Seung, 1999). Because of the two nonnegative factors $\mathbf{U}$ and $\mathbf{H}$, only a nonnegative data matrix can be factorized by NMF. CF is proposed to address the problem. CF models each cluster as a linear combination of the data points, and each data point is a linear combination of the cluster center. CF can be used to process any data sets even if they contain negative data points, and CF can be solved quickly by multiplicative update rules. In CF, $\boldsymbol{u}_c \in \mathbb{R}^{d\times 1}$, each column vector of $\mathbf{U}$ denotes the center of the cluster c where $c \in \{1, 2, \ldots, k\}$. These centers are represented by a linear combination of the data points $\mathbf{U} = \mathbf{XW}$ where $\mathbf{W} \geq 0$ (Xu & Gong, 2004). Thus, the basic form of CF tries to optimize the following problem:

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X} - \mathbf{XWH}\|_{\mathrm{F}}^2\\ &\text{s.t. } \mathbf{W} \geq 0, \mathbf{H} \geq 0\,. \end{aligned} \tag{2.1}$$

It is straightforward to check that the objective function, equation 2.1, suffers from scale ambiguity: if $\mathbf{W}$ and $\mathbf{H}$ are the solution, then $\mathbf{WM}$ and $\mathbf{M}^{-1}\mathbf{H}$ are also a solution for any positive diagonal matrix $\mathbf{M}$. To eliminate this uncertainty, in practice, the Euclidean length of each column vector in matrix $\mathbf{U} = \mathbf{XW}$ is required to be 1 (i.e., $\boldsymbol{w}_c^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{w}_c = 1$) (Xu et al., 2003; Xu & Gong, 2004; Cai, He et al., 2011), and the matrix $\mathbf{H}$ is adjusted accordingly so that $\mathbf{XWH}$ does not change, which can be achieved by

$$\mathbf{W} \leftarrow \mathbf{W}[\mathrm{diag}(\mathbf{W}^\top \mathbf{X}^\top \mathbf{X}\mathbf{W})]^{-\frac{1}{2}}, \tag{2.2}$$

$$\mathbf{H} \leftarrow \mathbf{H}[\mathrm{diag}(\mathbf{W}^\top \mathbf{X}^\top \mathbf{X}\mathbf{W})]^{\frac{1}{2}}, \tag{2.3}$$

where the function diag$(\cdot)$ sets all of the nondiagonal elements of a matrix to zeros.

However, CF considers only the global Euclidean geometry. The local geometric structure can be effectively modeled through a nearest-neighbor graph on a scattering of data points (Chung, 1997; Belkin & Niyogi, 2001; He & Niyogi, 2003). To preserve the local structure of the data set, following previous work on adaptive neighbors (Nie, Wang, & Huang, 2014), we exploit the local geometry of the data distribution by optimizing elements of the graph matrix in a global view. Given n_v types of heterogeneous views, $v = 1, 2, \ldots, n_v$, and instead of using precomputed graph matrices, a graph matrix can be learned by solving the following problem,

$$\min_{s_{ij}} \sum_{i,j=1}^{n} \left(\sum_{v=1}^{n_v} \alpha^v \|\boldsymbol{h}_i^v - \boldsymbol{h}_j^v\|_2^2 \right) (s_{ij})^\lambda$$
$$\text{s.t. } \forall j, \boldsymbol{s}_j \geq 0, \mathbf{1}^\top \boldsymbol{s}_j = 1, \tag{2.4}$$

where s_{ij} is an element of the affinity matrix and denotes the probability of the two data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ connecting with each other. If $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close to each other, s_{ij} is seen to be a relatively larger value, $\boldsymbol{s}_j$ is the jth column of S, parameter λ is used to control the distribution of s_{ij}, and α^v denotes the weight of vth view.

Let $\mathbf{X}^v \in \mathbb{R}^{d_v \times n}$ denote the data matrix in the vth view. We incorporate equation 2.4 as an additional term of CF:

$$\min_{\mathbf{W}^v, \mathbf{H}^v, s_{ij}} \sum_{v=1}^{n_v} \alpha^v \left(\|\mathbf{X}^v - \mathbf{X}^v \mathbf{W}^v \mathbf{H}^v\|_{\mathrm{F}}^2 + \sum_{i,j=1}^{n} \|\boldsymbol{h}_i^v - \boldsymbol{h}_j^v\|_2^2 (s_{ij})^\lambda \right)$$
$$\text{s.t. } \forall v, \mathbf{W}^v \geq 0, \mathbf{H}^v \geq 0, \forall j, \boldsymbol{s}_j \geq 0, \mathbf{1}^\top \boldsymbol{s}_j = 1. \tag{2.5}$$

However, problem 2.5 has a trivial solution with respect to α^v, since only one view weight may be learned to 1 and others are 0. If we solve the following problem,

$$\min_{\alpha^v} \sum_{v=1}^{n_v} (\alpha^v)^2$$
$$\text{s.t. } \forall v, \alpha^v \geq 0, \sum_{v=1}^{n_v} \alpha^v = 1, \tag{2.6}$$

the optimal solution is that all views can be obtained from the same view weight $\frac{1}{n^v}$, which can be seen as a prior in the view weight assignment.

It is clearly difficult to specify the weights α^v in equation 2.5 without prior knowledge. By combining equations 2.5 and 2.6, the weight of each view can be learned adaptively, which reflects the importance of the corresponding views. We then propose the following overall objective function,

$$\mathcal{O} = \min_{\mathbf{W}^v, \mathbf{H}^v, \alpha^v, s_{ij}} \sum_{v=1}^{n_v} \alpha^v \Big(\|\mathbf{X}^v - \mathbf{X}^v \mathbf{W}^v \mathbf{H}^v\|_\mathrm{F}^2$$
$$+ \sum_{i,j=1}^{n} \|\boldsymbol{h}_i^v - \boldsymbol{h}_j^v\|_2^2 (s_{ij})^\lambda \Big) + \gamma \sum_{v=1}^{n_v} (\alpha^v)^2$$
$$\text{s.t. } \forall v, \mathbf{W}^v \geq 0, \mathbf{H}^v \geq 0, \alpha^v \geq 0, \sum_{v=1}^{n_v} \alpha^v = 1,$$
$$\forall j, \boldsymbol{s}_j \geq 0, \mathbf{1}^\top \boldsymbol{s}_j = 1, \tag{2.7}$$

where γ is the regularization parameter.

As opposed to precomputing the affinity graphs, the graph in equation 2.7 is learned by globally modeling all the features from multiple views, making the multiview learning procedures mutually beneficial and reciprocal. In the following section, we describe a novel solution for obtaining the local optima to solve the objective function in equation 2.7. In equation 2.7, the first term is CF, the second term is the manifold regularization, and the third term is ℓ_2-norm regularization. The first term is used to learn the low-dimensional data representation $\mathbf{H}^v$ because most NMF-based methods are applied to data clustering. The second term is used to add a manifold regularization so that the data structure of the original space is still perserved in low-dimensional manifold. The third term is used to avoid the trivial solution of the view-weight α^v.

3 Optimization

3.1 Algorithm Derivation. We optimize equation 2.7 with the following three steps.

Step 1. First, we fix α^v and s_{ij} for all v, i, and j, updating $\mathbf{W}^v$ and $\mathbf{H}^v$ for each v independently. Then equation 2.7 becomes

$$\begin{aligned}\mathcal{O}_1 = \min_{\mathbf{W}^v,\mathbf{H}^v} \|\mathbf{X}^v - \mathbf{X}^v\mathbf{W}^v\mathbf{H}^v\|_{\mathrm{F}}^2 + \sum_{i,j=1}^{n} \|\boldsymbol{h}_i^v - \boldsymbol{h}_j^v\|_2^2 (s_{ij})^\lambda \\ \text{s.t. } \mathbf{W}^v \geq 0, \mathbf{H}^v \geq 0. \end{aligned} \tag{3.1}$$

Defining $\mathbf{K}^v = (\mathbf{X}^v)^\top \mathbf{X}^v$, we can rewrite the objective function, equation 3.1, as

$$\begin{aligned} &\|\mathbf{X}^v - \mathbf{X}^v\mathbf{W}^v\mathbf{H}^v\|_{\mathrm{F}}^2 + \sum_{i,j=1}^{n} \|\boldsymbol{h}_i^v - \boldsymbol{h}_j^v\|_2^2 (s_{ij})^\lambda \\ &= \mathrm{Tr}\big((\mathbf{X}^v - \mathbf{X}^v\mathbf{W}^v\mathbf{H}^v)^\top(\mathbf{X}^v - \mathbf{X}^v\mathbf{W}^v\mathbf{H}^v)\big) \\ &\quad + 2\left(\sum_{i=1}^{n} \boldsymbol{h}_i d_{ii} \boldsymbol{h}_i^\top - \sum_{i,j=1}^{n} \boldsymbol{h}_i (s_{ij})^\lambda \boldsymbol{h}_j^\top\right) \\ &= \mathrm{Tr}\big((\mathbf{I} - \mathbf{W}^v\mathbf{H}^v)^\top \mathbf{K}^v (\mathbf{I} - \mathbf{W}^v\mathbf{H}^v)\big) + 2\big(\mathrm{Tr}(\mathbf{H}^v\mathbf{D}(\mathbf{H}^v)^\top) \\ &\quad - \mathrm{Tr}(\mathbf{H}^v\mathbf{S}(\mathbf{H}^v)^\top)\big) \\ &= \mathrm{Tr}\big(\mathbf{K}^v - 2(\mathbf{H}^v)^\top(\mathbf{W}^v)^\top\mathbf{K}^v + (\mathbf{H}^v)^\top(\mathbf{W}^v)^\top\mathbf{K}^v\mathbf{W}^v\mathbf{H}^v\big) \\ &\quad + 2\mathrm{Tr}(\mathbf{H}^v\mathbf{L}(\mathbf{H}^v)^\top), \end{aligned} \tag{3.2}$$

where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is an identity matrix, $\mathbf{S} = [(s_{ij})^\lambda]$, $\mathbf{D}$ is a diagonal matrix and its elements are column sums of $\mathbf{S}$, $d_{ii} = \sum_j (s_{ij})^\lambda$, $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is the graph Laplacian (Chung, 1997), and $\mathrm{Tr}(\cdot)$ denotes the trace operator.

The multiplicative update algorithm of equation 3.1 is based on the following theorem proposed by Sha, Lin, Saul, and Lee (2007).

Theorem 1. *The minimization of the quadratic objective function* $f(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}^\top\boldsymbol{y}$ *is*

$$\begin{aligned} &\mathit{min}_{\boldsymbol{y}} \frac{1}{2}\boldsymbol{y}^\top\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}^\top\boldsymbol{y} \\ &\mathit{s.t.}\ \ \boldsymbol{y} \geq 0, \end{aligned} \tag{3.3}$$

where $\boldsymbol{A} = [a_{ij}]$ *is an arbitrary* $n \times n$ *symmetric semipositive matrix and* $\boldsymbol{b} = [b_i]$ *is an arbitrary* $n \times 1$ *vector. The iterative solution is expressed in terms of the positive component* $\boldsymbol{A}^+$ *and negative component* $\boldsymbol{A}^-$ *of the matrix* $\boldsymbol{A}$ *in equation 3.3:*

$$a_{ij}^{+} = \begin{cases} a_{ij}, & \textit{if } a_{ij} > 0; \\ 0, & \textit{otherwise}, \end{cases} \tag{3.4}$$

$$a_{ij}^{-} = \begin{cases} |a_{ij}|, & \textit{if } a_{ij} < 0; \\ 0, & \textit{otherwise}. \end{cases} \tag{3.5}$$

It is straightforward to find that $\boldsymbol{A} = \boldsymbol{A}^{+} - \boldsymbol{A}^{-}$. The solution y_i that minimizes equation 3.3 can be obtained by the following update rule:

$$y_i^{t+1} \leftarrow y_i^t \left[\frac{-b_i + \sqrt{b_i^2 + 4(\boldsymbol{A}^{+}\boldsymbol{y}^t)_i(\boldsymbol{A}^{-}\boldsymbol{y}^t)_i}}{2(\boldsymbol{A}^{+}\boldsymbol{y}^t)_i} \right]. \tag{3.6}$$

It can be seen from equation 3.2 that $\mathcal{O}_1$ is a quadratic form of $\mathbf{W}^v$ or $\mathbf{H}^v$, so equation 3.6 can be applied to solve the objective function $\mathcal{O}_1$ and the corresponding $\mathbf{A}$ and b_i need to be identified. By fixing $\mathbf{H}^v$, the part b_i^w for the quadratic form of $\mathcal{O}_1(\mathbf{W}^v)$ can be obtained by performing its derivative with respect to $\mathbf{W}^v$ at $\mathbf{W} = 0$,

$$b_i^w = \left.\frac{\partial \mathcal{O}_1(\mathbf{W}^v)}{\partial w_{ic}}\right|_{\mathbf{W}^v=0} = -2\big(\mathbf{K}^v(\mathbf{H}^v)^{\top}\big)_{ic}, \tag{3.7}$$

and the part $\mathbf{A}^w$ for the quadratic form of $\mathcal{O}_1(\mathbf{W})$ can be obtained by performing the second-order derivative with respect to $\mathbf{W}$:

$$\mathbf{A}^w = \frac{\partial^2 \mathcal{O}_1(\mathbf{W}^v)}{\partial w_{ic} \partial w_{js}} = 2k_{ij}^v\big(\mathbf{H}^v\big(\mathbf{H}^v\big)^{\top}\big)_{cs}. \tag{3.8}$$

Substituting $\mathbf{A}^w$ and b_i^w into equation 3.6, we obtain the update rule of w_{ic},

$$w_{ic}^{t+1} \leftarrow w_{ic}^t \frac{b_i^w + \sqrt{(b_i^w)^2 + 4(\mathbf{Q}^{+})_{ic}(\mathbf{Q}^{-})_{ic}}}{2(\mathbf{Q}^{+})_{ic}} \tag{3.9}$$

where $\mathbf{Q}^{+} = (\mathbf{K}^v)^{+}\mathbf{W}^v\mathbf{H}^v(\mathbf{H}^v)^{\top}$, $\mathbf{Q}^{-} = (\mathbf{K}^v)^{-}\mathbf{W}^v\mathbf{H}^v(\mathbf{H}^v)^{\top}$, $(\mathbf{K}^v)^{+}$ and $(\mathbf{K}^v)^{-}$ denote the nonnegative matrices with elements,

$$(k_{ij}^v)^{+} = \begin{cases} k_{ij}^v, & \text{if } k_{ij}^v > 0; \\ 0, & \text{otherwise}, \end{cases} \tag{3.10}$$

$$(k_{ij}^{v})^{-} = \begin{cases} |k_{ij}^{v}|, & \text{if } k_{ij}^{v} < 0; \\ 0, & \text{otherwise.} \end{cases} \tag{3.11}$$

It is straightforward to check that $\mathbf{K}^{v} = (\mathbf{K}^{v})^{+} - (\mathbf{K}^{v})^{-}$.

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

$$h_{ic}^{t+1} \leftarrow h_{ic}^{t} \frac{b_i^h + \sqrt{(b_i^h)^2 + 4(\mathbf{P}^{+})_{ic}(\mathbf{P}^{-})_{ic}}}{2(\mathbf{P}^{+})_{ic}} \tag{3.12}$$

where $b_i^h = (\mathbf{K}^{v}\mathbf{W}^{v})_{ic}$, $\mathbf{P}^{+} = (\mathbf{W}^{v}\mathbf{H}^{v})^{\top}(\mathbf{K}^{v})^{+}\mathbf{W}^{v} + 2\mathbf{D}(\mathbf{H}^{v})^{\top}$, and $\mathbf{P}^{-} = (\mathbf{W}^{v}\mathbf{H}^{v})^{\top}(\mathbf{K}^{v})^{-}\mathbf{W}^{v} + 2\mathbf{S}(\mathbf{H}^{v})^{\top}$.

To avoid the scale ambiguity of equation 3.1, we also adopt the normalization strategy of equations 2.2 and 2.3.

Step 2. We next fix s_{ij}, $\mathbf{W}^{v}$, and $\mathbf{H}^{v}$ for all v, i, and j, solving α^{v}. Then equation 2.7 becomes

$$\begin{aligned} \mathcal{O}_2 = \min_{\alpha^{v}} & \sum_{v=1}^{n_v} \alpha^{v} \left(\|\mathbf{X}^{v} - \mathbf{X}^{v}\mathbf{W}^{v}\mathbf{H}^{v}\|_{\mathrm{F}}^{2} + \sum_{i,j=1}^{n} \|\boldsymbol{h}_i^{v} - \boldsymbol{h}_j^{v}\|_2^2 (s_{ij})^{\lambda} \right) \\ & + \gamma \sum_{v=1}^{n_v} (\alpha^{v})^2 \\ & \text{s.t. } \forall v, \alpha^{v} \geq 0, \sum_{v=1}^{n_v} \alpha^{v} = 1. \end{aligned} \tag{3.13}$$

Denoting $f^{v} = \|\mathbf{X}^{v} - \mathbf{X}^{v}\mathbf{W}^{v}\mathbf{H}^{v}\|_{\mathrm{F}}^{2} + \sum_{i,j=1}^{n} \|\boldsymbol{h}_i^{v} - \boldsymbol{h}_j^{v}\|_2^2 (s_{ij})^{\lambda}$, we can solve the objective function, equation 3.13, as

$$\begin{aligned} & \min_{\boldsymbol{\alpha}} \left\| \boldsymbol{\alpha} + \frac{1}{2\gamma} \boldsymbol{f} \right\|_2^2 \\ & \text{s.t. } \boldsymbol{\alpha} \geq 0, \mathbf{1}^{\top}\boldsymbol{\alpha} = \mathbf{1}, \end{aligned} \tag{3.14}$$

where $\boldsymbol{\alpha} = [\alpha^1, \alpha^2, \ldots, \alpha^{n_v}]^{\top}$ and $\boldsymbol{f} = [f^1, f^2, \ldots, f^{n_v}]^{\top}$.

The Lagrangian function of equation 3.14 is

$$\mathcal{L}(\boldsymbol{\alpha}, \eta, \boldsymbol{\beta}) = \left\| \boldsymbol{\alpha} + \frac{1}{2\gamma} \boldsymbol{f} \right\|_2^2 + \rho\big(1 - \mathbf{1}^{\top}\boldsymbol{\alpha}\big) + \boldsymbol{\beta}^{\top}(-\boldsymbol{\alpha}), \tag{3.15}$$

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

According to the Karush-Kuhn-Tucker condition (Boyd & Vandenberghe, 2004), it can be verified that the optimal solution $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha} = \left(-\frac{1}{2\gamma}\boldsymbol{f} + \rho\mathbf{1}\right)_+ . \tag{3.16}$$

Step 3. Our last step is to fix α^v, $\mathbf{W}^v$, and $\mathbf{H}^v$ for all v, solving s_{ij}. Then equation 2.7 becomes

$$\begin{aligned} \mathcal{O}_3 = \min_{s_{ij}} \sum_{v=1}^{n_v} \alpha^v \sum_{i,j=1}^{n} \|\boldsymbol{h}_i^v - \boldsymbol{h}_j^v\|_2^2 (s_{ij})^\lambda \\ \text{s.t. } \forall\, j, \boldsymbol{s}_j \geq 0, \mathbf{1}^\top \boldsymbol{s}_j = 1. \end{aligned} \tag{3.17}$$

Denoting $p_{ij} = \sum_{v=1}^{n_v} \alpha^v ||\boldsymbol{h}_i^v - \boldsymbol{h}_j^v||_2^2$, equation 3.14 becomes

$$\begin{aligned} &\min_{s_{ij}} \sum_{i,j=1}^{n} (s_{ij})^\lambda p_{ij} \\ &\text{s.t. } \forall\, j, \boldsymbol{s}_j \geq 0, \mathbf{1}^\top \boldsymbol{s}_j = 1. \end{aligned} \tag{3.18}$$

The Lagrange function of equation 3.18 is

$$\mathcal{L}\left(s_{ij}, \eta\right) = \sum_{i,j=1}^{n} (s_{ij})^\lambda p_{ij} - \eta \left(\sum_{j=1}^{n} s_{ij} - 1\right), \tag{3.19}$$

where η is the Lagrangian multiplier.

The optimum s_{ij} can be obtained by calculating the first-order derivative, equation 3.19, with respect to s_{ij}. By setting the derivative of the function with respect to a_{ij} to zero, we have

$$\lambda (s_{ij})^{\lambda-1} p_{ij} - \eta = 0. \tag{3.20}$$

Hence,

$$s_{ij} = \left(\frac{\eta}{\lambda p_{ij}}\right)^{\frac{1}{\lambda-1}} . \tag{3.21}$$

Substituting equation 3.21 into the constraint $\sum_{j=1}^{n} s_{ij} = 1$, we obtain

$$s_{ij} = \frac{\left(\lambda p_{ij}\right)^{\frac{1}{1-\lambda}}}{\sum_{j=1}^{n} \left(\lambda p_{ij}\right)^{\frac{1}{1-\lambda}}} = \frac{\left(p_{ij}\right)^{\frac{1}{1-\lambda}}}{\sum_{j=1}^{n} \left(p_{ij}\right)^{\frac{1}{1-\lambda}}} . \tag{3.22}$$

Algorithm 1: Multiview Concept Factorization Clustering.

```
Input: Data for $n_v$ views $\{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^{n_v}\}$, parameters $\gamma$, $\lambda$, and the number of clusters $k$.
Output: $\mathbf{H}^v$.
Initialize: $\forall v$, initialize $\mathbf{W}^v$ and $\mathbf{H}^v$ randomly in the range of [0,1], and $\alpha^v$ is $\frac{1}{n_v}$.
    $s_{ij}$ is initialized by the optimal solution to the problem,
    $\min_{\sum_{j=1}^{n} s_{ij}=1, s_{ij}\geq 0} \sum_{v=1}^{n_v} \left( \sum_{i,j=1}^{n} \|\boldsymbol{x}_i^v - \boldsymbol{x}_j^v\|_2^2 \right) (s_{ij})^{\lambda}.$
while not converge do
    for $v \in [1, n_v]$ do
        while not converge do
            Update $\mathbf{W}^v$ and $\mathbf{H}^v$ by equations 3.9 and 3.12.
            Normalize $\mathbf{W}^v$ and $\mathbf{H}^v$ by equations 2.2 and 2.3.
        end while
    end for
    Update $\boldsymbol{\alpha}$ by equation 3.14.
    Update $s_{ij}$ by equation 3.22.
end while
```

The algorithm for solving the problem, equation 2.7, is summarized in algorithm 1.

3.2 Convergence Analysis.

Theorem 2. *If we set the parameter λ to a value of larger than 1, the objective function $\mathcal{O}$ in equation 2.7 is nonincreasing with respect to one variable while holding the others.*

To prove theorem 2 and because $\mathcal{O}_2$ is a convex optimization problem and $\mathcal{O}_3$ is a convex optimization problem when $\lambda > 1$, we first prove theorem 3:

Theorem 3. *The objective function $\mathcal{O}_1$ in equation 3.1 is nonincreasing under the update rules in equations 3.9 and 3.12.*

We use an auxiliary function as used in the expectation-maximization algorithm (Dempster, Laird, & Rubin, 1977; Lee & Seung, 2001) to prove the convergence of $\mathcal{O}_1$. The definition of the auxiliary function is given by definition 1.

Definition 1. *$g(h, h')$ is an auxiliary function for $f(h)$ if the conditions*

$$g(h, h') \geq f(h), g(h, h) = f(h) \tag{3.23}$$

are satisfied.

Theorem 4. *If $g(h, h^t)$ is an auxiliary function of $f(h)$, then $f(h)$ is nonincreasing under the update*

$$h^{t+1} = arg \underset{h}{min}\, g(h, h^t), \tag{3.24}$$

where t is the number of iterations.

Proof. $f(h^{t+1}) \leq g(h^{t+1}, h^t) \leq g(h^t, h^t) = f(h^t)$. □

Thus, by iterating the update in equation 3.24, we obtain a sequence of estimates that converge to a local minimum $h_{\min} = \arg\min_h f(h)$ of the objective function $f(h)$.

Note that the minimum of the objective function $\mathcal{O}_1$ in equation 3.1 is our update rules in equations 3.9 and 3.12 with theorem 4 and proper auxiliary functions. As the two update rules are based on theorem 1, we need the proof of theorem 1:

Proof. The function

$$g(y_i, y_i^t) = \frac{1}{2}\sum_i \frac{(\mathbf{A}^+\boldsymbol{y}^t)_i}{y_i^t} y_i^2 - \frac{1}{2}\sum_{ij} (\mathbf{A}^-)_{ij} y_i^t y_j^t \left(1 + \log \frac{y_i y_j}{y_i^t y_j^t}\right) + \sum_i b_i y_i \tag{3.25}$$

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

$f(y_i)$ can be decomposed as the combination of three terms:

$$f(y_i) = \frac{1}{2}\sum_{ij} y_i a_{ij}^+ y_j - \frac{1}{2}\sum_{ij} y_i a_{ij}^- y_j + \sum_i b_i y_i. \tag{3.26}$$

Clearly, $g(y_i, y_i) = f(y_i)$. According to definition 1, we need to prove $g(y_i, y_i^t) \geq f(y_i)$. Comparing equation 3.26 to 3.25 is equivalent to proving the following inequalities:

$$\frac{1}{2}\sum_{ij} y_i a_{ij}^{+} y_j \leq \frac{1}{2}\sum_{i} \frac{(\mathbf{A}^{+}\boldsymbol{y}^{t})_i}{y_i^t} y_i^2, \tag{3.27}$$

$$-\frac{1}{2}\sum_{ij} y_i a_{ij}^{-} y_j \leq -\frac{1}{2}\sum_{ij} a_{ij}^{-} y_i^t y_j^t \Big(1+\log\frac{y_i y_j}{y_i^t y_j^t}\Big) \tag{3.28}$$

As the left of equation 3.27 is the quadratic function $\boldsymbol{y}^\top \mathbf{A}^{+}\boldsymbol{y}$, we suppose that the right of that equation has a similar form $\boldsymbol{y}^\top \mathbf{K}\boldsymbol{y}$. We only need to prove $\boldsymbol{y}^\top(\mathbf{K}-\mathbf{A}^{+})\boldsymbol{y} \geq 0$. Defining δ_{ij} as the Kronecker delta function, K denotes the diagonal matrix with elements $k_{ij} = \delta_{ij}\frac{(\mathbf{A}^{+}\boldsymbol{y}^{t})_i}{y_i^t}$:

$$\begin{aligned}\boldsymbol{y}^\top(\mathbf{K}-\mathbf{A}^{+})\boldsymbol{y} &= \sum_{ij} y_i y_i^t (k_{ij} - a_{ij}^{+}) y_j^t y_j \\ &= \sum_{ij} a_{ij}^{+} y_i^t y_j^t y_i^2 - \sum_{ij} a_{ij}^{+} y_i^t y_j^t y_i y_j \\ &= \frac{1}{2}\sum_{ij} a_{ij}^{+} y_i^t y_j^t (y_i - y_j)^2 \geq 0. \end{aligned} \tag{3.29}$$

To prove equation 3.28, we use a simple inequality: $x \geq 1 + \log x$. Substituting $x = \frac{y_i y_j}{y_i^t y_j^t}$ into the inequality,

$$y_i y_j \geq y_i^t y_j^t \Big(1+\log\frac{y_i y_j}{y_i^t y_j^t}\Big); \tag{3.30}$$

thus, $g(y_i, y_i^t) \geq f(y_i)$ holds with equations 3.29 and 3.30.

The minimization of equation 3.25 is performed by setting its derivative to zero with respect to y_i, leading to the update rule in equation 3.6. □

As $\mathcal{O}_1$ is a quadratic form of $\mathbf{W}^v$ or $\mathbf{H}^v$, we have proved theorem 3. Again, since objective functions $\mathcal{O}_2$ and $\mathcal{O}_3$ are convex optimization problems, theorem 2 is also proved.

3.3 Computational Complexity Analysis. The overall computational complexity of the proposed algorithm is $O(n^2)$, where n is the number of data points. The complexity of the first step in the algorithm is $O(t_1 k n^2)$, where t_1 is the number of iterations and k is the number of clusters. The second step is $O(t_2 n_v)$, where t_2 is the number of iterations and n_v is the number of views. The third step is $O(kn^2 n_v)$. Since $n >> t_1, n >> k, n >> n_v$, and $n >> t_2$, the overall complexity is $O(n^2)$.

4 Experimental Results

4.1 Data Sets. 3-Sources is constructed from three well-known online news sources: BBC, Reuters, and the *Guardian*. In total there are 948 news articles covering 416 distinct news stories from February 2009 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.

WebKB contains four subsets (Cornell, Texas, Washington, and Wisconsin) of documents and is described by two views (content and citations). Cornell contains 195 documents over five labels (student, project, course, staff, and faculty). The documents are described by 1703 words in the content view and by the 569 links between them in the citations view. Texas, Washington, and Wisconsin have the same structure and contain 187, 230, and 265 documents, respectively.

Animals with Attributes (AwA) is an animal data set. We use four published features for 500 images belonging to five classes. The features are SIFT, Local Self-Similarity, PyramidHOG, and SURF, respectively.

Caltech101 is an object recognition data set. We select seven widely used classes: Faces, Motorbikes, Dolla-Bill, Garfield, Snoopy, Stop-Sign, and Windsor-Chair. We sample 441 data points from the data set in our experiment.

Handwritten Numerals (Numerals) consists of 2000 data points for 0 to 9 10-digit classes. We use the four published visual features extracted from each image: Fourier coefficients of the character shapes, profile correlations, pixel averages in 2×3 windows, and Zernike moment.

Outdoor Scene (Scene) is an outdoor scene data set. This data set contains 2150 data points corresponding to 2150 color images, which belong to eight outdoor scene categories: coast, mountain, forest, open county, street, inside city, tall buildings, and highways.

4.2 Experimental Setup. We evaluate the performance of the proposed multiview concept factorization (MVCF) method on the nine data sets. MVCF is compared with state-of-the-art multiview clustering methods, including multimodal spectral clustering (MMSC) (Cai, Nie et al., 2011), cotrained spectral clustering (CTSC) (Kumar & Daumé, 2011), coregularized spectral clustering (CRSC) (Kumar et al., 2011), multiview NMF clustering (MultiNMF) (Liu et al., 2013), robust multiview *k*-means clustering (RMKMC) (Cai, Nie, & Huang, 2013), robust multiview spectral clustering (RMSC) (Xia et al., 2014), and large-scale multiview spectral clustering (MVSC) (Li et al., 2015), to demonstrate its effectiveness.

We compare MVCF with the following methods:

1. Single-view CF. We apply the proposed concept factorization framework into each single view of all data sets to obtain the clustering

results. To achieve this, we solve the optimization problem,

$$\min_{\mathbf{W},\mathbf{H},s_{ij}} \|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{H}\|_{\mathrm{F}}^2+\sum_{i,j=1}^{n}\|\boldsymbol{h}_i-\boldsymbol{h}_j\|_2^2(s_{ij})^{\lambda}$$
$$\text{s.t. } \mathbf{W}\geq 0, \mathbf{H}\geq 0, \forall\, j, \boldsymbol{s}_j\geq 0, \mathbf{1}^{\top}\boldsymbol{s}_j=1, \tag{4.1}$$

where **X** denotes the data of each single view and the optimal **H** is used for clustering.

2. MMSC (Cai, Nie et al., 2011). In the MMSC algorithm, each type of feature is considered as one modal. The MMSC algorithm aims to learn a commonly shared graph Laplacian matrix by unifying different modals. In addition, a nonnegative relaxation is added in this method to improve the robustness and efficiency of clustering.
3. CTSC (Kumar & Daumé, 2011). This is a multiview spectral clustering approach using the idea of cotraining. Under the assumption that the true underlying clustering would assign a point to the same cluster regardless of the view, it learns the clustering in one view and then uses it to label the data in the other view so as to modify the graph structure (similarity matrix).
4. CRSC (Kumar et al., 2011). This applies a centroid-based coregularization scheme to multiview spectral clustering. To make the clusterings in different views agree with each other, CRSC enforces view-specific eigenvectors to look similar by regularizing them toward a common consensus and then optimizes individual clusterings as well as the consensus by using a joint cost function.
5. MultiNMF (Liu et al., 2013). This aims to search for a factorization that gives compatible clustering solutions across multiple views, requiring coefficient matrices learned from factorizations of different views to be regularized toward a common consensus.
6. RMKMC (Cai et al., 2013). This simultaneously performs clustering using each view of features and unifies their results based on their importance to the clustering task. $\ell_{2,1}$-norm is also employed to improve the robustness.
7. RMSC (Xia et al., 2014). For each view, this constructs a corresponding transition probability matrix, which is then used for recovering a low-rank transition probability matrix. Based on this, the standard Markov chain method is utilized for processing, and then clustering is conducted.
8. MVSC (Li et al., 2015). This large-scale multiview spectral clustering approach is based on the bipartite graph. MVSC uses local manifold fusion to integrate heterogeneous features and approximates the similarity graphs using bipartite graphs to improve efficiency. Furthermore, this method can be easily extended to handle the out-of-sample problem.

The parameters of the eight baseline algorithms are tuned to obtain the best results, as suggested by the respective authors. Our method has two parameters: λ and γ. For all experiments, λ is empirically fixed as 10 for all data sets. γ controls the weight distribution of different views, and we obtain the best γ^* by searching $\log_{10}\gamma$ in the range of $[-4.8, -2.6]$ with interval 0.2. We obtain the optimal data representations by adding the product of the data representation matrix $\mathbf{H}^v$ and its weight α_v in each view together (Wang et al., 2017). Because each learned $\mathbf{H}^v$ represents diverse information of an intrinsic data structure, we can integrate them with the weighted sum rule. Following Li et al. (2015), we obtain the clustering labels by running k-means on the optimal data. Without loss of generality, we run each method 10 times and report the mean performance as well as the standard deviation. In each experiment, we run k-means clustering processing 30 times and obtain the best result to reduce the randomness of k-means.

4.3 Evaluation Metric. Three metrics—the clustering accuracy (ACC), the normalized mutual information (NMI) (Strehl & Ghosh, 2002) and the Purity (Ievgen & Younes, 2014)—are used to evaluate the performance in this work. These measurements are widely used, and they can be calculated by comparing the obtained label of each sample with the ideal label provided by the data set. For each metric, a larger value indicates better clustering performance.

ACC is used to measure the clustering accuracy of the clustering result, which is defined as

$$\mathrm{ACC} = \frac{\sum_{i=1}^{N} \delta(g_i, \mathrm{map}(r_i))}{N}, \tag{4.2}$$

where N denotes the total number of samples, g_i denotes the ground truth label of the ith sample, r_i denotes the corresponding obtained clustering label, δ denotes the Dirac delta function,

$$\delta(x, y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases}, \tag{4.3}$$

and $\mathrm{map}(r_i)$ is the optimal mapping function that permutes the obtained labels to match the ground truth labels. The best mapping can be found by using the Kuhn-Munkres algorithm (Lovász & Plummer, 2009).

NMI is used to measure the similarity between the preexisting input label G and the clustering assignment R, which is defined as

$$\mathrm{NMI}(G, R) = \frac{I(G, R)}{\sqrt{E(G)E(R)}} \tag{4.4}$$

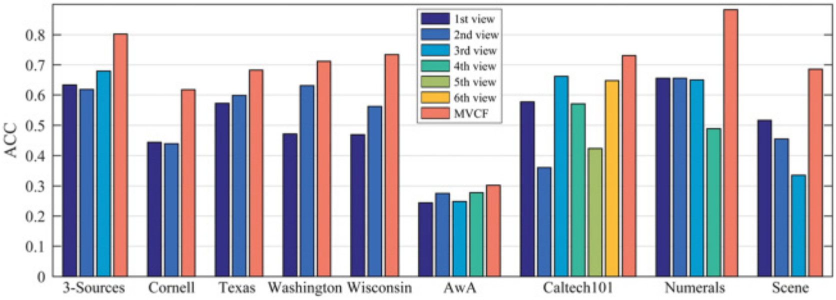

Figure 1: ACC results of multiview and single-view clustering using concept factorization.

where $I(G, R)$ denotes the mutual information between G and R, and $E(\cdot)$ returns the entropy.

Let n_c be the number of objects in the cth cluster $(1 \leq c \leq k)$ obtained by using the clustering algorithm and $\tilde{n}_s$ be the object number of the sth cluster $(1 \leq s \leq k)$ in the ground truth label. Then NMI is defined as

$$\mathrm{NMI} = \frac{\sum_{c=1}^{k} \sum_{s=1}^{k} n_{c,s} \log\left(\frac{N \cdot n_{c,s}}{n_c \tilde{n}_s}\right)}{\sqrt{\left(\sum_{c=1}^{k} n_c \log \frac{n_c}{N}\right)\left(\sum_{s=1}^{k} \tilde{n}_s \log \frac{\tilde{n}_s}{N}\right)}}, \tag{4.5}$$

where $n_{c,s}$ is the number of objects in the intersection of the cth cluster and the sth cluster. NMI varies from zero for a totally wrong clustering result to one for a perfect clustering result.

Purity gives the percentage of correct labels obtained. Assuming total N data points belong to k clusters, the definition of purity is

$$\mathrm{Purity} = \frac{1}{N} \sum_{i=1}^{k} \max_{1 \leq j \leq k} |p_i \cap q_j|, \tag{4.6}$$

where p_i represents the ith obtained cluster and q_i implies the ith ground truth cluster.

4.4 Performance Comparison. First, we apply the proposed concept factorization framework into each single view of all data sets to obtain the clustering results and then compare the results with MVCF's results, which are shown in Figures 1, 2, and 3. From the three bar graphs, it is obvious that MVCF outperforms any single view's result, which means that the multiview framework can learn and integrate all of the useful information from complementary views, consequently obtaining better clustering results.

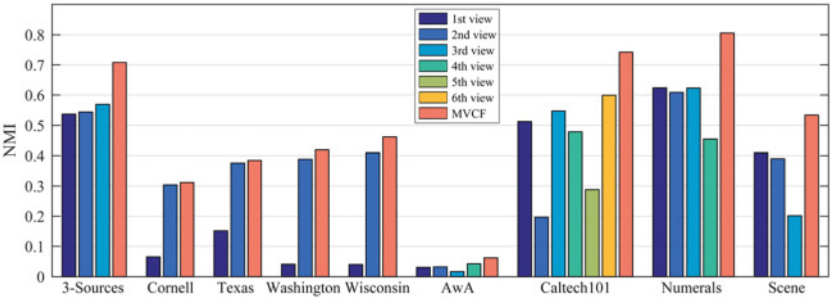

Figure 2: NMI results of multiview and single-view clustering using concept factorization.

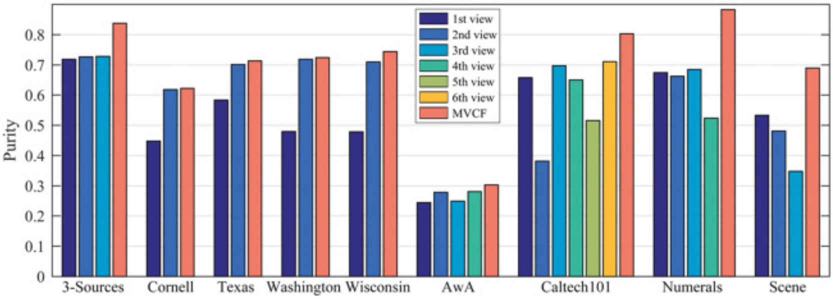

Figure 3: Purity results of multiview and single-view clustering using concept factorization.

After comparing the proposed method with other baseline algorithms, we show the clustering results in terms of ACC, NMI, and Purity in Tables 1, 2, and 3, respectively. In each row of the tables, the best and second best results are highlighted in bold. Note that AwA, Caltech101, Numerals, and Scene contain negative data, so the MultiNMF method is not applicable to these data sets.

Clearly, MVCF achieves the best performance in most cases; for the remaining ones, it surprisingly still produces competitive results. Compared with the second-best performance on the 3-Sources data set, the proposed MVCF method significantly improves the clustering performance significantly by more than 10%. In addition, we calculate the mean performance of the different methods on all data sets, shown in the last row of each table. An interesting point is that CRSC is then demonstrated to be the second-best method and the MVCF method performs the best. The quantitative result fully demonstrates the superiority of the proposed method because MVCF better captures the geometrical structure of the data space.

Table 1: Clustering Performance Measured by ACC(%).

	MMSC	CTSC	CRSC	RMKMC	RMSC	MVSC	MultiNMF	MVCF
3-Sources	52.37 ± 3.02	59.97 ± 1.29	**62.72** ± 0.93	45.92 ± 3.57	52.78 ± 1.30	62.54 ± 3.47	54.20 ± 0.75	**80.24** ± 1.58
Cornell	46.97 ± 0.26	43.71 ± 0.73	**57.03** ± 1.15	48.62 ± 6.42	38.62 ± 0.64	50.62 ± 3.49	42.87 ± 0.50	**61.74** ± 4.97
Texas	**64.01** ± 2.05	56.26 ± 1.10	61.23 ± 1.64	58.45 ± 2.28	42.46 ± 1.19	59.63 ± 5.20	63.90 ± 1.08	**68.24** ± 4.34
Washington	59.00 ± 1.25	59.24 ± 0.30	59.61 ± 0.25	60.61 ± 2.41	42.09 ± 1.77	59.78 ± 3.18	**63.09** ± 0.86	**71.26** ± 1.17
Wisconsin	51.36 ± 0.38	49.43 ± 0.60	58.87 ± 0.69	57.55 ± 3.12	36.11 ± 1.56	**59.28** ± 2.96	45.85 ± 3.47	**73.47** ± 2.28
AwA	28.12 ± 0.44	28.23 ± 0.37	26.92 ± 0.58	28.80 ± 0.72	28.52 ± 0.76	**29.72** ± 0.49	—	**30.20** ± 0.71
Caltech101	67.82 ± 0.41	71.24 ± 1.29	**73.83** ± 1.55	69.21 ± 2.13	71.00 ± 0.80	71.07 ± 1.75	—	**73.11** ± 2.02
Numerals	77.31 ± 1.60	79.08 ± 1.02	86.35 ± 3.04	70.81 ± 2.68	83.52 ± 2.10	**88.41** ± 1.11	—	**88.30** ± 1.48
Scene	43.46 ± 1.83	63.80 ± 1.07	**65.44** ± 1.00	55.99 ± 2.65	61.71 ± 0.12	64.28 ± 1.20	—	**68.56** ± 3.93
Average	54.49	56.77	**61.33**	55.11	50.76	60.59	—	**68.35**

Table 2: Clustering Performance Measured by NMI(%).

	MMSC	CTSC	CRSC	RMKMC	RMSC	MVSC	MultiNMF	MVCF
3-Sources	45.28 ± 5.40	**54.55** ± 0.59	54.09 ± 1.17	33.57 ± 8.16	43.23 ± 1.28	47.31 ± 6.30	48.23 ± 0.87	**70.84** ± 2.04
Cornell	17.26 ± 1.71	22.67 ± 0.48	**32.54** ± 0.79	23.92 ± 9.42	14.93 ± 0.36	27.48 ± 4.12	13.21 ± 0.96	**31.18** ± 5.89
Texas	**31.95** ± 5.06	25.39 ± 0.38	28.89 ± 2.02	26.08 ± 5.32	18.36 ± 0.34	26.48 ± 6.03	25.37 ± 1.10	**38.42** ± 4.54
Washington	22.49 ± 2.20	**30.89** ± 0.53	30.67 ± 0.49	29.79 ± 5.09	18.79 ± 1.00	27.52 ± 6.04	24.49 ± 1.25	**41.99** ± 1.57
Wisconsin	15.61 ± 1.17	23.02 ± 0.95	**41.81** ± 0.00	35.32 ± 6.64	14.40 ± 0.33	35.45 ± 3.93	12.64 ± 3.76	**46.26** ± 3.40
AwA	4.65 ± 0.46	4.33 ± 0.34	3.16 ± 0.28	4.37 ± 0.92	4.55 ± 0.26	**5.16** ± 0.63	—	**6.20** ± 0.57
Caltech101	57.27 ± 0.60	69.99 ± 0.49	**70.84** ± 1.78	61.82 ± 1.92	68.09 ± 1.89	64.97 ± 3.89	—	**74.26** ± 2.97
Numerals	71.30 ± 0.30	76.42 ± 0.31	77.65 ± 1.58	67.21 ± 3.00	75.94 ± 1.07	**80.52** ± 1.26	—	**80.53** ± 1.54
Scene	35.50 ± 2.95	50.22 ± 0.46	50.11 ± 0.72	47.81 ± 2.92	47.78 ± 0.27	**50.45** ± 0.81	—	**53.37** ± 2.13
Average	33.48	39.72	**43.31**	36.65	34.01	40.59	—	**49.23**

Table 3: Clustering Performance Measured by Purity(%).

	MMSC	CTSC	CRSC	RMKMC	RMSC	MVSC	MultiNMF	MVCF
3-Sources	62.25 ± 4.90	**75.15** ± 3.46	70.41 ± 2.33	58.22 ± 4.59	66.33 ± 0.98	68.93 ± 4.87	63.91 ± 1.12	**83.79** ± 1.58
Cornell	51.54 ± 1.46	54.87 ± 0.48	**62.05** ± 1.67	57.13 ± 7.28	50.26 ± 1.51	59.44 ± 3.56	46.51 ± 0.84	**62.26** ± 4.94
Texas	**69.84** ± 3.23	67.01 ± 1.84	65.61 ± 2.28	65.78 ± 3.86	59.47 ± 2.32	66.26 ± 3.86	64.87 ± 1.43	**71.34** ± 2.63
Washington	63.65 ± 1.45	66.87 ± 0.67	**68.04** ± 1.33	67.61 ± 2.37	62.74 ± 1.79	67.43 ± 2.09	65.17 ± 0.56	**72.39** ± 0.83
Wisconsin	54.15 ± 0.89	60.42 ± 1.95	**74.08** ± 1.82	69.32 ± 4.69	56.26 ± 1.30	70.83 ± 2.47	52.91 ± 2.11	**74.42** ± 2.51
AwA	28.56 ± 0.60	27.52 ± 0.14	27.18 ± 0.58	29.16 ± 0.83	29.14 ± 0.33	**29.98** ± 0.69	—	**30.30** ± 0.74
Caltech101	69.23 ± 0.40	**80.63** ± 3.04	79.46 ± 2.95	71.32 ± 3.02	73.38 ± 2.62	76.87 ± 2.96	—	**80.32** ± 2.39
Numerals	77.56 ± 1.04	72.62 ± 0.16	86.37 ± 2.98	72.65 ± 2.61	83.52 ± 2.10	**88.41** ± 1.11	—	**88.30** ± 1.48
Scene	43.62 ± 1.73	43.47 ± 2.03	**65.44** ± 1.00	56.35 ± 3.24	61.71 ± 0.12	64.39 ± 1.10	—	**68.94** ± 3.10
Average	57.82	60.95	**66.52**	60.84	60.31	65.84	—	**70.23**

5 Conclusion and Future Work

In this letter, we have proposed a multiview clustering model that can address the negative data issue under nonnegativity constraints and the interview correlation issue of most existing models. The first issue is tackled by adopting concept factorization, and the second is addressed by learning a single affinity graph from the multiple views. We have proposed a novel CF-based algorithm that not only inherits the strengths of NMF, such as fast multiplicative iteration and parts-based representation in accordance with human brain intuition (Lee & Seung, 1999) but also is applicable to data sets containing negative values. We have taken the great impact of local manifold geometry structure into consideration and extended the proposed algorithm to a multiview clustering to effectively use the complementary information of the data. The experiments demonstrate the superiority of our algorithm over other state-of-the-art methods.

MVCF exploits the data structure by using manifold regularization without the requirement of eigenvalue decomposition, which renders MVCF effective. However, the time complexity is still high. In the future, we will use the active Riemannian subspace search for maximum margin matrix factorization (Yan et al., 2015) to reduce the complexity and obtain high accuracy in large-scale data sets.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China under grant 61201422, the Specialized Research Fund for the Doctoral Program of Higher Education under grant 20120211120013, and the Fundamental Research Funds for the Central Universities under grants lzujbky-2017-it73 and No. lzujbky-2017-it76.

References

Balcan, M. F., Blum, A., & Yang, K. (2004). Co-training and expansion: Towards bridging theory and practice. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17* (pp. 89–96). Cambridge, MA: MIT Press.

Belkin, M., & Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14* (pp. 585–591). Cambridge, MA: MIT Press.

Bickel, S., & Scheffer, T. (2004). Multi-view clustering. In *Proceedings of the IEEE International Conference on Data Mining* (pp. 19–26). Piscataway, NJ: IEEE.

Blaschko, M. B., & Lampert, C. H. (2008). Correlational spectral clustering. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (pp. 1–8). Piscataway, NJ: IEEE.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Annual Conference on Computational Learning Theory*. New York: ACM.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

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

Cai, X., Nie, F., & Huang, H. (2013). Multi-view *k*-means clustering on big data. In *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 2598–2604). Cambridge, MA: AAAI.

Cai, X., Nie, F., Huang, H., & Kamangar, F. (2011). Heterogeneous image feature integration via multi-modal spectral clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 11977–1984). Piscataway, NJ: IEEE.

Chung, F. R. (1997). *Spectral graph theory*. Providence, RI: American Mathematical Society.

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

Ding, C. H., He, X., & Simon, H. D. (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the SIAM International Conference on Data Mining* (pp. 606–610). Philadelphia: SIAM.

He, X., & Niyogi, P. (2003). Locality preserving projections. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information proceesing systems* (pp. 153–160). Cambridge, MA: MIT Press.

Ievgen, R., & Younes, B. (2014). Random subspaces NMF for unsupervised transfer learning. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 3901–3908). Piscataway, NJ: IEEE.

Jia, Y., Song, X., Zhou, J., Liu, L., Nie, L., & Rosenblum, D. S. (2016). Fusing social networks with deep learning for volunteerism tendency prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 165–171). Cambridge, MA: AAAI.

Kakade, S. M., & Foster, D. P., (2007). Multi-view regression via canonical correlation analysis. In *Proceedings of the Annual Conference on Learning Theory* (pp. 82–96). New York: Springer.

Kumar, A., & Daumé, H. (2011). A co-training approach for multi-view spectral clustering. In *Proceedings of the International Conference on Machine Learning* (pp. 393–400). New York: ACM.

Kumar, A., Rai, P., & Daume, H. (2011). Co-regularized multi-view spectral clustering. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information proceesing systems, 24* (pp. 1413–1421). Red Hook, NY: Curran.

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

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information proceesing systems, 13* (pp. 556–562). Cambridge, MA: MIT Press.

Li, Y., Nie, F., Huang, H., & Huang, J. (2015). Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Cambridge, MA: AAAI.

Liu, J., Wang, C., Gao, J., & Han, J. (2013). Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the SIAM International Conference on Data Mining* (pp. 252–260). Philadelphia: SIAM.

Lovász, L., & Plummer, M. D. (2009). *Matching theory*. Providence, RI: American Mathematical Society.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information proceesing systems, 14* (pp. 849–856). Cambridge, MA: MIT Press.

Nie, F., Wang, X., & Huang, H. (2014). Clustering and projected clustering with adaptive neighbors. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 977–986). New York: ACM.

Sha, F., Lin, Y., Saul, L. K., & Lee, D. D. (2007). Multiplicative updates for nonnegative quadratic programming. *Neural Computation*, *19*, 2004–2031.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 888–905.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—A knowledge reuse framework for combining partitionings. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 93–99). Cambridge, MA: AAAI.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*, 395–416.

Wang, J., Tian, F., Yu, H., Liu, C. H., Zhan, K., & Wang, X. (2017). Diverse non-negative matrix factorization for multiview data representation. *IEEE Transactions on Cybernetics*.

Wang, W., & Zhou, Z. H. (2010). A new analysis of co-training. In *Proceedings of the International Conference on Machine Learning* (pp. 1135–1142). New York: ACM.

Xia, R., Pan, Y., Du, L., & Yin, J., (2014). Robust multi-view spectral clustering via low-rank and sparse decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 2149–2155). Cambridge, MA: AAAI.

Xu, W., & Gong, Y. (2004). Document clustering by concept factorization. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 202–209). New York: ACM.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 267–273). New York: ACM.

Yang, X., Zhang, T., & Xu, C. (2015). Cross-domain feature learning in multimedia. *IEEE Transactions on Multimedia*, *17*, 64–78.

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

Yan, Y., Nie, F., Li, W., Gao, C., Yang, Y., & Xu, D. (2016). Image classification by cross-media active learning with privileged information. *IEEE Transactions on Multimedia*, *18*, 2494–2502.

Yan, Y., Tan, M., Tsang, I. W., Yang, Y., Zhang, C., & Shi, Q. (2015). Scalable maximum margin matrix factorization by active riemannian subspace search. In *Proceedings* *of the International Joint Conference on Atrificial Intelligence* (pp. 3988–3994). Cambridge, MA: AAAI.

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

Zelnik-Manor, L., & Perona, P. (2004). Self-tuning spectral clustering. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information proceesing systems, 14* (pp. 1601–1608). Cambridge, MA: MIT Press.

Zhan, K., Zhang, C., Guan, J., & Wang, J. (2017). Graph learning for multiview clustering. *IEEE Transactions on Cybernetics* 1–9.

Zhang, X., Zhao, L., Zong, L., Liu, X., & Yu, H. (2014). Multi-view clustering via multi-manifold regularized nonnegative matrix factorization. In *Proceedings of the IEEE International Conference on Data Mining* (pp. 1103–1108). Piscataway, NJ: IEEE.

Received August 16, 2017; accepted October 28, 2017.