Neurocomputing 138 (2014) 120–130

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Dual-graph regularized concept factorization for clustering

Jun Ye [a,b,*], Zhong Jin [a]

[a] School of Computer Science & Technology, Nanjing University of Science and Technology, Nanjing 210094, China
[b] School of Natural Sciences, Nanjing University of Posts & Telecommunications, Nanjing 210003, China

ARTICLE INFO

Article history:
Received 27 March 2013
Received in revised form
26 December 2013
Accepted 13 February 2014
Communicated by Deng Cai
Available online 13 April 2014

Keywords:
NMF
Concept factorization
Dual-graph regularized
Document clustering
Image clustering

ABSTRACT

In past decades, tremendous growths in the amount of text documents and images have become omnipresent, and it is very important to group them into clusters upon desired. Recently, matrix factorization based techniques, such as Non-negative Matrix Factorization (NMF) and Concept Factorization (CF), have yielded impressive results for clustering. However, both of them effectively see only the global Euclidean geometry, whereas the local manifold geometry is not fully considered. Recent research has shown that not only the observed data are found to lie on a nonlinear low dimensional manifold, namely data manifold, but also the features lie on a manifold, namely feature manifold. In this paper, we propose a novel algorithm, called dual-graph regularized concept factorization for clustering (GCF), which simultaneously considers the geometric structures of both the data manifold and the feature manifold. As an extension of GCF, we extend that our proposed method can also be apply to the negative dataset. Moreover, we develop the iterative updating optimization schemes for GCF, and provide the convergence proof of our optimization scheme. Experimental results on TDT2 and Reuters document datasets, COIL20 and PIE image datasets demonstrate the effectiveness of our proposed method.

1. Introduction

Clustering is one of the most important research topics in both machine learning and data mining communities. It arms at partitioning the data into groups of similar objects. An enormous number and variety of methods have been proposed over the past several decades to solve clustering problems [1]. Generally, clustering methods can be categorized as agglomerative and partitional. Agglomerative clustering methods group the data points into a hierarchical tree structure using bottom-up approaches. The procedure starts by placing each data point into a distinct cluster and then iteratively merges the two most similar clusters into one parent cluster. On the other hand, data partitioning methods decompose the data set into a given number of disjoint clusters which are usually optimal in terms of some predefined criterion functions [2]. Both of them have been well studied and investigated in previous literatures [3,4].

In the last decade, matrix factorization based approaches have attracted considerable attention for clustering. With regard to these methods, each text document or image in the corpus is often treated as a data point in the high dimensional linear space. Clustering analysis aims to look for similar data points and ensure them within the same cluster in maximum degree. Intuitively, similar samples are more likely to be grouped together than different ones, and this could be attributed to the fact that characteristics shared by similar ones in original data spaces are inherited by new representations in lower dimensional spaces, which makes the clustering more easily. There are particularly two popular matrix factorization methods widely applied to clustering analysis, i.e., Nonnegative Matrix Factorization (NMF) [5] and Concept Factorization (CF) [2]. CF mainly strives to address the limitations and meanwhile inherits all the strengths of NMF, such as better semantic interpretation and easily derived clustering results. In CF, each concept or component is modeled as a linear combination of the data points while each data point consists of a linear combination of the concepts. In general, CF is more advantageous than NMF, because of its merits that it can be applied to any data points taking both positive and negative values. However, regardless of NMF or CF, they only consider using the global Euclidean geometry to find new basis vectors, according to how the new data representation is generated [6]. However, many previous studies have shown human generated text data is probably sampled from a submanifold of the ambient Euclidean space [7–10]. In fact, the human generated text documents cannot possibly "fill up" the high dimensional Euclidean space uniformly.

* Corresponding author at: School of Computer Science & Technology, Nanjing University of Science and Technology, Nanjing 210094, China.
Tel.: +86 1360 1587 306.
E-mail address: yj8422092@163.com (J. Ye).

http://dx.doi.org/10.1016/j.neucom.2014.02.029
0925-2312/

Therefore, the intrinsic manifold structure needs to be considered while learning new data representations [11]. Inspired by this, Li et al. [12] proposed discriminative orthogonal nonnegative matrix factorization (DON), in order to obtain a good data representation that preserves both the local geometrical structure and the global discriminating information. And also in order to preserve the intrinsically geometrical structure and use the prior knowledge, Li et al. [13] proposed locally constrained a-optimal nonnegative projection (LCA). They are all NMF-based methods.

Recently, Cai et al. [11] proposed locally consistent concept factorization (LCCF) based on CF to extract the underlying concepts which are consistent with the low dimensional manifold structure. The obtained concepts can well capture the intrinsic geometrical structure and the documents associated with similar concepts can be well clustered. However, the method mentioned above focuses on one-sided clustering, i.e., clustering the data based on the similarities along the feature. Considering the duality between data points and features, several co-clustering algorithms have been proposed and shown to be superior to traditional one-sided clustering [14–20]. Gu et al. [19] proposed a Dual Regularized Co-Clustering (DRCC) method based on semi-nonnegative matrix tri-factorization. In order to discover an appropriate intrinsic manifold, Li et al. [20] proposed realtional multimanifold co-clustering based on symmetric nonnegative matrix tri-factorization. Based on NMF, Shang et al. [17] proposed Graph Dual Regularization Non-negative Matrix Factorization (DNMF) for co-clustering, which achieves an encouraging performance.

Motivated by recent progress in dual regularization [17–20] and concept factorization [2,11], we propose a novel algorithm called dual-graph regularized concept factorization for clustering (GCF), which simultaneously considers the geometric structures of the data manifold as well as the feature manifold. We encode the geometric structure information of data and feature spaces by constructing two nearest neighbor graphs, respectively. Our proposed algorithm GCF is based on the CF, it can be optimized by iterative multiplicative updating schemes, and their convergence proof of this work include:

1. We propose a novel dual-graph regularized concept factorization (GCF) algorithm which simultaneously considers the geometric structure information contained in data points as well as features.
2. We develop iterative multiplicative updating optimization schemes to solve our proposed algorithm GCF, and provide the convergence proof of the optimization scheme.

The remainder of this paper is organized as follows: Section 2 presents a brief overview of some related works. A novel GCF algorithm is proposed in Section 3. As an extension of GCF, the algorithm for negative data is described in Section 4. Experimental results on many real-world datasets are presented in Section 5. Section 6 is conclusions.

2. Related works

In this section, we briefly review some related works to our research work.

2.1. NMF

Consider a data matrix $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_N]\in\mathbf{R}^{M\times N}$, each column of $\mathbf{X}$ is a sample vector. NMF aims to decompose $\mathbf{X}$ into two low rank nonnegative matrices, basis matrix $\mathbf{U}=[u_{ik}]\in\mathbf{R}^{M\times K}$ and feature matrix $\mathbf{V}=[v_{jk}]\in\mathbf{R}^{N\times K}$, such that $\mathbf{X}\approx\mathbf{U}\mathbf{V}^T$, where $K\ll\min\{M,N\}$. Therefore, the objective optimization problem of NMF can be concluded as follows:

$$\min_{U,V}:\mathbf{J}_{\mathbf{NMF}}=||\mathbf{X}-\mathbf{U}\mathbf{V}^T||_F^2\quad s.t.\quad U,V\geq 0 \tag{1}$$

Several methods have been proposed to find a solution to this nonlinear optimization problem. The multiplicative updates rules were first investigated by Lee and Seung [21] as follows:

$$u_{ik}^{t+1}=u_{ik}^{t}\frac{(\mathbf{X}\mathbf{V})_{ik}}{(\mathbf{U}\mathbf{V}^T\mathbf{V})_{ik}};\qquad v_{jk}^{t+1}=v_{jk}^{t}\frac{(\mathbf{X}^T\mathbf{U})_{jk}}{(\mathbf{V}\mathbf{U}^T\mathbf{U})_{jk}} \tag{2}$$

Theorem 1. *[21] for $\mathbf{X},\mathbf{U},\mathbf{V}\geq 0$, the objective function $\mathbf{J}_{\mathbf{NMF}}$ in Eq. (1) is nonincreasing under each of the above multiplicative updating rules stated in Eq. (2).*

The nonnegative constraints on $\mathbf{U}$ and $\mathbf{V}$ require the combination coefficients among different basis can only be positive. This is the most significant difference between NMF and other matrix factorization methods, e.g., SVD. Unlike SVD, no subtractions can occur in NMF. For this reason, it is believed that NMF can learn a parts-based representation have been observed in many real world problems such as face analysis, document clustering.

2.2. DRCC

Gu et al. [19] proposed a dual regularized co-clustering (DRCC) method based on graph regularized (semi-)NMF, which imposes graph regularization on both the data points and features cluster assignment matrices. The objective optimization problem can be concluded as follows:

$$\min_{U,S,\mathbf{V}}:\mathbf{J}_{\mathbf{DRCC}}=||\mathbf{X}-\mathbf{U}\mathbf{S}\mathbf{V}^T||_F^2+\lambda\mathrm{Tr}(\mathbf{V}^T\mathbf{L}_{\mathbf{V}}\mathbf{V})+\mu\mathrm{Tr}(\mathbf{U}^T\mathbf{L}_{\mathbf{U}}\mathbf{U})$$
$$s.t.\quad \mathbf{U},\mathbf{V}\geq 0 \tag{3}$$

where $\lambda,\mu\geq 0$ are the regularization parameters, and $\mathbf{S}$ is a matrix whose entries can take any signs. $\mathbf{L}_{\mathbf{V}}=\mathbf{D}^{\mathbf{V}}-\mathbf{W}^{\mathbf{V}}$ is the graph Laplacian of the data graph which reflects the label smoothness of the data points, where $\mathbf{W}^{\mathbf{V}}$ is the weight matrix and $\mathbf{D}^{\mathbf{V}}$ is a diagonal matrix whose entries are column sums of $\mathbf{W}^{\mathbf{V}}$. $\mathbf{L}_{\mathbf{U}}=\mathbf{D}^{\mathbf{U}}-\mathbf{W}^{\mathbf{U}}$ is the graph Laplacian of the feature graph which reflects the label smoothness of the feature. The multiplicative updating rules minimizing Eq. (3) are given as [19].

$$\mathbf{S}=(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{X}\mathbf{V}(\mathbf{V}^T\mathbf{V})^{-1},$$

$$v_{jk}\leftarrow v_{jk}\sqrt{\frac{[\lambda\mathbf{W}^V\mathbf{V}+\mathbf{A}^+ +\mathbf{V}\mathbf{B}^-]_{jk}}{[\lambda\mathbf{D}^V\mathbf{V}+\mathbf{A}^- +\mathbf{V}\mathbf{B}^+]_{jk}}},$$

$$u_{ik}\leftarrow u_{ik}\sqrt{\frac{[\mu\mathbf{W}^U\mathbf{U}+\mathbf{P}^+ +\mathbf{U}\mathbf{Q}^-]_{ik}}{[\mu\mathbf{D}^U\mathbf{U}+\mathbf{P}^- +\mathbf{U}\mathbf{Q}^+]_{ik}}} \tag{4}$$

where $\mathbf{A}=\mathbf{X}^T\mathbf{U}\mathbf{S}=\mathbf{A}^+-\mathbf{A}^-$, $\mathbf{B}=\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S}=\mathbf{B}^+-\mathbf{B}^-$, $\mathbf{P}=\mathbf{X}\mathbf{V}\mathbf{S}^T=\mathbf{P}^+-\mathbf{P}^-$ and $\mathbf{Q}=\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}^T=\mathbf{Q}^+-\mathbf{Q}^-$, where $\mathbf{A}_{ij}^+=(|\mathbf{A}_{ij}|+\mathbf{A}_{ij})/2$, $\mathbf{A}_{ij}^-=(|\mathbf{A}_{ij}|-\mathbf{A}_{ij})/2$

Theorem 2. *[19] For $\mathbf{U},\mathbf{V}\geq 0$, the objective function $\mathbf{J}_{\mathbf{DRCC}}$ in Eq. (3) is non-increasing under each of the above updating rules stated in Eq. (4).*

Gu et al. [19] have proved that the iterative multiplicative updating scheme stated in Eq. (4) will find local minima of the objective function $\mathbf{J}_{\mathbf{DRCC}}$.

2.3. CF

NMF can only be performed in the original feature space of the data points. In the case that the data are highly non-linear distributed, it is desirable that we can kernelize NMF and apply the powerful idea of the kernel method [11]. To achieve this goal, Xu and Gong [2] proposed an extension of NMF which is called Concept Factorization (CF). Therefore, the objective optimization problem of CF can be concluded as follows:

$$\min_{W,V} : \mathbf{J_{CF}} = ||\mathbf{X} - \mathbf{XWV}^T||_F^2 \quad s.t. \quad \mathbf{W}, \mathbf{V} \geq 0 \tag{5}$$

The multiplicative updates rules were given [2] as follows:

$$w_{jk}^{t+1} = w_{jk}^t \frac{(\mathbf{KV})_{jk}}{(\mathbf{KWV}^T\mathbf{V})_{jk}}; \qquad v_{jk}^{t+1} = v_{jk}^t \frac{(\mathbf{KW})_{jk}}{(\mathbf{VW}^T\mathbf{KW})_{jk}} \tag{6}$$

where $\mathbf{K} = \mathbf{X}^T\mathbf{X}$. These multiplicative updating rules only involve the inner product of $\mathbf{x}$ and thus CF can be easily kernelized.

Theorem 3. *[2] For $\mathbf{X}, \mathbf{W}, \mathbf{V} \geq 0$, the objective function $\mathbf{J_{CF}}$ in Eq. (5) is nonincreasing under each of the above multiplicative updating rules stated in Eq. (6).*

With extensive experimental results, Xu and Gong [2] show the superiority of CF over NMF for document clustering.

2.4. LCCF

Cai et al. [11] proposed a locally consistent concept factorization (LCCF) to find concepts with respect to the intrinsic local manifold geometry structure. The objective optimization problem of LCCF can be concluded as follows:

$$\min_{W,\mathbf{V}} : \mathbf{J_{LCCF}} = ||\mathbf{X} - \mathbf{XWV}^T||_F^2 + \lambda \mathrm{Tr}(\mathbf{V}^T\mathbf{LV}) \quad s.t. \quad \mathbf{W}, \mathbf{V} \geq 0 \tag{7}$$

where $\lambda \geq 0$ is the regularization parameter. Please see [11] for details. The multiplicative updates rules were given [11] as follows:

$$w_{jk} \leftarrow w_{jk} \frac{(\mathbf{KV})_{jk}}{(\mathbf{KWV}^T\mathbf{V})_{jk}}, \qquad v_{jk} \leftarrow v_{jk} \frac{(\mathbf{KW} + \lambda\mathbf{SV})_{jk}}{(\mathbf{VW}^T\mathbf{KW} + \lambda\mathbf{DV})_{jk}} \tag{8}$$

Theorem 4. *[11] For $\mathbf{X}$, $\mathbf{W}, \mathbf{V} \geq 0$, the objective function $\mathbf{J_{LCCF}}$ in Eq. (7) is nonincreasing under each of the above multiplicative updating rules stated in Eq. (8).*

Cai et al. [11] have proved that the iterative multiplicative updating scheme stated in Eq. (8) will find local minima of the objective function $\mathbf{J_{LCCF}}$.

3. Graph dual regularization concept factorization (GCF)

In this section, we first propose a novel dual-graph regularized concept factorization (GCF) algorithm, which simultaneously considers the geometric structures of both the data manifold and the feature manifold. Then we present an optimization scheme based on the iterative updating rules of two factor matrices to solve its objective function. Finally, we present the convergence proof of our iterative updating scheme.

3.1. Data and feature graphs

A natural treatment for the data sampled from a manifold is to construct a graph to discretely approximate the manifold, whose vertices correspond to the data samples, while the edge weight represents the affinity between the data points. One common assumption about the affinity between data points is cluster assumption, which says if two samples are close to each other in the input space, then their labels (or embeddings) are also close to each other [19,20]. Furthermore, recently literatures [17,19,20] shows that not only the observed data are found to lie on a nonlinear low dimensional manifold, namely data manifold, but also from the dual view, the features are discrete samplings from another manifold, namely feature manifold. As a result, we introduce two graphs to explore the geometric structures of both the data manifold and the feature manifold, respectively. In other words, we construct two graphs: data graph and feature graph to effectively model the geometric structures of both the data manifold and the feature manifold.

We first construct a p nearest neighbor data graph whose vertices correspond to $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. And we use the 0–1 weighting scheme for constructing the p nearest neighbor graph as in [3,19], and define the data weight matrix as follows:

$$(\mathbf{S}^V)_{js} = \begin{cases} 1, & \text{if } \mathbf{x}_s \in N_p(\mathbf{x}_j); \\ 0, & \text{otherwise.} \end{cases} \quad j, s = 1, \ldots, N$$

where $N_p(\mathbf{x}_j)$ represents the set of p nearest neighbors of $\mathbf{x}_j$. The graph Laplacian of the data graph is defined as $\mathbf{L_V} = \mathbf{D^V} - \mathbf{S^V}$, where $\mathbf{D^V}$ is a diagonal degree matrix whose entries are given by $(\mathbf{D^V})_{\mathbf{jj}} = \sum_{\mathbf{s}} (\mathbf{S^V})_{\mathbf{js}}$.

And we can also use the 0–1 weighting scheme for constructing a p nearest neighbor feature graph whose vertices correspond to $\{\mathbf{x}_1^T, \ldots, \mathbf{x}_M^T\}$, and define the feature weight matrix as follows:

$$(\mathbf{S}^U)_{js} = \begin{cases} 1, & \text{if } \mathbf{x}_s^T \in N_p(\mathbf{x}_j^T); \\ 0, & \text{otherwise.} \end{cases} \quad j, s = 1, \ldots, M$$

The graph Laplacian of the feature graph is also defined as $L_{\mathbf{U}} = D^{\mathbf{U}} - S^{\mathbf{U}}$.

3.2. Objective function of GCF

Based on the graphs regularizers of both data manifold and feature manifold, we propose a novel dual-graph regularized concept factorization (GCF). By defining the original data point $\mathbf{x}_j^T$ on to the low-dimensional space $\mathbf{U} \in \mathbf{R}^{M \times K}$, the discrete approximation function of the smoothness can be computed as $\mathbf{U}^T\mathbf{L}_U\mathbf{U} \in \mathbf{R}^{K \times K}$, where $\mathbf{U}$ equals $\mathbf{XW}$ for LCCF.

GCF aims to find the nonnegative matrices $\mathbf{W} \in \mathbf{R}^{N \times K}$, $\mathbf{V} \in \mathbf{R}^{K \times N}$, which minimize the following objective function:

$$\min_{W,V} : \mathbf{J_{GCF}} = ||\mathbf{X} - \mathbf{XWV^T}||^2 + \lambda \mathrm{Tr}(\mathbf{V^T L_V V}) + \mu \mathrm{Tr}(\mathbf{W^T L_W W}) \quad s.t. \quad \mathbf{W}, \mathbf{V} \geq 0 \tag{9}$$

where $\mathbf{L_W} = \mathbf{X}^T\mathbf{L}_U\mathbf{X} = \mathbf{X}^T(\mathbf{D^U} - \mathbf{S^U})\mathbf{X} = \mathbf{D^W} - \mathbf{S^W}$. The $\lambda \geq 0, \mu \geq 0$ are the regularization parameters, which balance the reconstruction error of GCF in the first term and graph regularizations in the second and third terms. When letting $\mu = 0$, GCF degenerates to the LCCF method in Eq. (7), and when letting $\lambda = \mu = 0$, GCF degenerates to the CF in Eq. (5).

3.3. A multiplicative algorithm

The objective function $\mathbf{J_{GCF}}$ of GCF in Eq. (9) is not convex in both $\mathbf{W}$ and $\mathbf{V}$ together. Therefore, it is unrealistic to expect an algorithm to find the global minimum of $\mathbf{J_{GCF}}$. In the following, we introduce an iterative algorithm which can achieve a local minimum. Define $\mathbf{K} = \mathbf{X^T X}$, the objective function in Eq. (9) can be rewritten as:

$$\begin{aligned} \mathbf{J_{GCF}} &= \mathrm{Tr}[(\mathbf{X} - \mathbf{XWV^T})^{\mathbf{T}}(\mathbf{X} - \mathbf{XWV^T})] + \lambda \mathrm{Tr}(\mathbf{V^T L_V V}) + \mu \mathrm{Tr}(\mathbf{W^T L_W W}) \\ &= \mathrm{Tr}[(\mathbf{I} - \mathbf{WV^T})^{\mathbf{T}}\mathbf{K}(\mathbf{I} - \mathbf{WV^T})] + \lambda \mathrm{Tr}(\mathbf{V^T L_V V}) + \mu \mathrm{Tr}(\mathbf{W^T L_W W}) \\ &= \mathrm{Tr}(\mathbf{K}) - \mathbf{2}\mathrm{Tr}(\mathbf{VW^T K}) + \mathrm{Tr}(\mathbf{VW^T KWV^T}) \\ &\quad + \lambda \mathrm{Tr}(\mathbf{V^T L_V V}) + \mu \mathrm{Tr}(\mathbf{W^T L_W W}) \end{aligned} \tag{10}$$

Let $\Psi=[\psi_{jk}]$ and $\Phi=[\varphi_{jk}]$ be the Lagrange multiplier for constraints $\mathbf{W}\geq 0$ and $\mathbf{V}\geq 0$, respectively. Then the Lagrange function $\mathbf{L}$ is

$$\mathbf{L}=\mathrm{Tr}(\mathbf{K})-2\mathrm{Tr}(\mathbf{V}\mathbf{W}^{\mathbf{T}}\mathbf{K})+\mathrm{Tr}(\mathbf{V}\mathbf{W}^{\mathbf{T}}\mathbf{K}\mathbf{W}\mathbf{V}^{\mathbf{T}})+\lambda\mathrm{Tr}(\mathbf{V}^{\mathbf{T}}\mathbf{L}_{\mathbf{V}}\mathbf{V})+\mu\mathrm{Tr}(\mathbf{W}^{\mathbf{T}}\mathbf{L}_{\mathbf{W}}\mathbf{W})+\mathrm{Tr}(\Psi\mathbf{W}^{\mathbf{T}})+\mathrm{Tr}(\Phi\mathbf{V}) \quad (11)$$

The partial derivatives of $\mathbf{L}$ with respect to $\mathbf{W}$ and $\mathbf{V}$ are

$$\frac{\partial\mathbf{L}}{\partial\mathbf{W}}=-2\mathbf{K}\mathbf{V}+2\mathbf{K}\mathbf{W}\mathbf{V}^{\mathbf{T}}\mathbf{V}+2\mu\mathbf{L}_{\mathbf{W}}\mathbf{W}+\Psi,$$

$$\frac{\partial\mathbf{L}}{\partial\mathbf{V}}=-2\mathbf{K}\mathbf{W}+2\mathbf{V}\mathbf{W}^{\mathbf{T}}\mathbf{K}\mathbf{W}+2\lambda\mathbf{L}_{\mathbf{V}}\mathbf{V}+\Phi$$

Using the KKT conditions $\psi_{jk}w_{ij}=0$ and $\varphi_{jk}v_{jk}=0$, we get the following equations:

$$(-\mathbf{K}\mathbf{V}+\mathbf{K}\mathbf{W}\mathbf{V}^{T}\mathbf{V}+\mu\mathbf{L}_{W}\mathbf{W})_{\mathrm{jk}}w_{jk}=0,\quad (-\mathbf{K}\mathbf{W}+V\mathbf{W}^{T}\mathbf{K}\mathbf{W}+\lambda\mathbf{L}_{V}\mathbf{V})_{\mathrm{jk}}v_{jk}=0$$

$$(-\mathbf{K}\mathbf{V}+\mathbf{K}\mathbf{W}\mathbf{V}^{T}\mathbf{V}+\mu\mathbf{D}^{W}\mathbf{W}-\mu\mathbf{S}^{W}\mathbf{W})_{\mathrm{jk}}w_{jk}=0,$$

$$(-\mathbf{K}\mathbf{W}+\mathbf{V}\mathbf{W}^{T}\mathbf{K}\mathbf{W}+\lambda\mathbf{D}^{V}\mathbf{V}-\lambda\mathbf{S}^{V}\mathbf{V})_{\mathrm{jk}}v_{jk}=0$$

The above equations lead to the following updating rules:

$$w_{jk}^{t+1}\leftarrow w_{jk}^{t}\frac{(\mathbf{K}\mathbf{V}+\mu\mathbf{S}^{W}\mathbf{W})_{jk}}{(\mathbf{K}\mathbf{W}\mathbf{V}^{T}\mathbf{V}+\mu\mathbf{D}^{W}\mathbf{W})_{jk}} \quad (12)$$

$$v_{jk}^{t+1}\leftarrow v_{jk}^{t}\frac{(\mathbf{K}\mathbf{W}+\lambda\mathbf{S}^{V}\mathbf{V})_{jk}}{(\mathbf{V}\mathbf{W}^{T}\mathbf{K}\mathbf{W}+\lambda\mathbf{D}^{V}\mathbf{V})_{jk}} \quad (13)$$

3.4. Convergence analysis

In the following, we will investigate the convergence of the updating rules in Eqs. (12) and (13). And regarding these two updating rules, we have the following theorem:

Theorem 5. *For* $\mathbf{X}$, $\mathbf{W}\geq 0$, $\mathbf{V}\geq 0$, *the objective function* $\mathbf{J}_{\mathbf{GCF}}$ *in Eq. (9) is nonincreasing under each of the above multiplicative updating rules stated in Eqs. (12) and (13).*

Please see Appendix A for a detailed proof. Recent studies [22,23] have shown that the alternate updating rules in Eq. (2) do not guarantee the convergence to a stationary point. But a slight modification proposed in [23,24] achieves this property. Our alternate updating rules in Eqs. (12) and (13) are essentially similar to the updating rules for LCCF, so the minor modification can also be applied.

For the objective function of GCF, it is easy to check that if $\mathbf{W}$ and $\mathbf{V}$ are the solution, then, $\mathbf{WD}$, $\mathbf{VD}^{-1}$ will also form a solution for any positive diagonal matrix $\mathbf{D}$. To eliminate this uncertainty, in practice people will further require that $\mathbf{w}^{\mathbf{T}}\mathbf{K}\mathbf{w}=1$, where $\mathbf{w}$ is the column vector of $\mathbf{W}$. The matrix $\mathbf{V}$ will be adjusted accordingly so that $\mathbf{W}\mathbf{V}^{T}$ does not change. This can be achieved by

$$\mathbf{V}\leftarrow\mathbf{V}[\mathrm{diag}(\mathbf{W}^{\mathbf{T}}\mathbf{K}\mathbf{W})]^{1/2},\quad \mathbf{W}\leftarrow\mathbf{W}[\mathrm{diag}(\mathbf{W}^{\mathbf{T}}\mathbf{K}\mathbf{W})]^{-1/2}$$

Our GCF method also adopts this strategy.

3.5. Connection with gradient descent method

Intuitively, the objective function of GCF in Eq. (9) can be minimized by gradient descent algorithm. Using gradient descent method, the additive update rules for Eq. (9) problem are

$$w_{jk}\leftarrow w_{jk}+\eta_{jk}\frac{\partial\mathbf{J}_{\mathbf{GCF}}}{\partial w_{jk}},\quad v_{jk}\leftarrow v_{jk}+\delta_{jk}\frac{\partial\mathbf{J}_{\mathbf{GCF}}}{\partial v_{jk}}$$

where η_{jk} and δ_{jk} are the parameters to control the step size of gradient descent. As long as they are sufficiently small, the updates should reduce $\mathbf{J}_{\mathbf{GCF}}$. We set

$$\eta_{jk}=-\frac{w_{jk}}{2(\mathbf{K}\mathbf{W}\mathbf{V}^{T}\mathbf{V}+\mu\mathbf{D}^{W}\mathbf{W})_{jk}},\quad \delta_{jk}=-\frac{v_{jk}}{2(\mathbf{V}\mathbf{W}^{T}\mathbf{K}\mathbf{W}+\lambda\mathbf{D}^{V}\mathbf{V})_{jk}}$$

Then we can obtain

$$w_{jk}+\eta_{jk}\frac{\partial\mathbf{J}_{\mathbf{GCF}}}{\partial w_{jk}}=w_{jk}\frac{(\mathbf{K}\mathbf{V}+\mu\mathbf{S}^{W}\mathbf{W})_{jk}}{(\mathbf{K}\mathbf{W}\mathbf{V}^{T}\mathbf{V}+\mu\mathbf{D}^{W}\mathbf{W})_{jk}},$$

$$v_{jk}+\delta_{jk}\frac{\partial\mathbf{J}_{\mathbf{GCF}}}{\partial v_{jk}}=v_{jk}\frac{(\mathbf{K}\mathbf{W}+\lambda\mathbf{S}^{V}\mathbf{V})_{jk}}{(\mathbf{V}\mathbf{W}^{T}\mathbf{K}\mathbf{W}+\lambda\mathbf{D}^{V}\mathbf{V})_{jk}}$$

And we can see that the multiplicative updating rules are the special cases of gradient descent with automatically step size parameter selection. The advantage of multiplicative updating rules is the guarantee the nonnegativity of $\mathbf{W}$ and $\mathbf{V}$. Theorem 5 also guarantees the multiplicative updating rules in Eqs. (12) and (13) converge to a local optimum.

3.6. Computational complexity analysis

In this subsection, we discuss the computational cost of our proposed algorithm comparing to standard NMF, CF, DRCC and LCCF. Suppose the multiplicative updates stops after t iterations, the overall cost for NMF is $O(tMNK)$. The overall cost for CF, DRCC and LCCF is $O(tN^2K+N^2M)$, $O(N^2M+NM^2+tMNK)$ and $O(N^2M+tN^2K)$, respectively. For our proposed method, the same p nearest neighbor graph needs $O(N^2M+NM^2)$ to construct. Since two weigh matrices $\mathbf{S}^{W}$ and $\mathbf{S}^{V}$ of our proposed method are sparse, based on the updating rules of our proposed algorithm, the cost of two iterative multiplicative updating procedures is $O(tN^2K)$. The overall cost for our proposed method is $O(N^2M+NM^2+tN^2K)$.

4. The algorithm for negative data matrices

The algorithm we introduced in Section 3.3 only works when the $\mathbf{K}$ is nonnegative. In the case that the data matrix has negative values, it is possible that the $\mathbf{K}$ has negative entries. In this section, we will introduce a general algorithm which can be applied for any case. Our approach follows [2], which is essentially based on the following theorem proposed by Sha et al. [11,25].

Theorem 6. *Define the non-negative general quadratic form as*

$$\mathbf{f}(v)=\frac{1}{2}v^{\mathbf{T}}\mathbf{A}v+b^{\mathbf{T}}v \quad (14)$$

where v is an $\mathbf{m}$ dimensional nonnegative vector, $\mathbf{A}$ is a symmetric positive definite matrix and b is an arbitrary $\mathbf{m}$ dimensional vector. Let $\mathbf{A}=\mathbf{A}^{+}-\mathbf{A}^{-}$, where $\mathbf{A}^{+}$ and $\mathbf{A}^{-}$ are two symmetric matrices whose elements are all positive. Then the solution v that minimizes $\mathbf{f}(v)$ can be obtained through the following iterative update

$$v_i\leftarrow v_i\left[\frac{-b_i+\sqrt{b_i^2+4(\mathbf{A}^{+}v)_i(\mathbf{A}^{-}v)_i}}{2(\mathbf{A}^{+}v)_i}\right] \quad (15)$$

From the Eq. (10), we can easily see that the objective function $\mathbf{J}_{\mathbf{GCF}}$ of GCF is a quadratic form of $\mathbf{W}$ (or $\mathbf{V}$) only and Theorem 6 can naturally be applied. We only need to identify the corresponding $\mathbf{A}$ and b in the objective function.

Fixing $\mathbf{V}$, the part b for the quadratic form $\mathbf{J}_{\mathbf{GCF}}(\mathbf{W})$ can be obtained by taking the first order derivative with respect to $\mathbf{W}$ at $\mathbf{W}=0$

$$\left.\frac{\partial\mathbf{J}_{\mathbf{GCF}}}{\partial w_{jk}}\right|_{\mathbf{W}=0}=\left.(-2\mathbf{K}\mathbf{V}+2\mathbf{K}\mathbf{W}\mathbf{V}^{T}\mathbf{V}+2\mu\mathbf{L}_{W}\mathbf{W})_{jk}\right|_{\mathbf{W}=0}=(-2\mathbf{K}\mathbf{V})_{jk} \quad (16)$$

The part **A** for the quadratic from $\mathbf{J_{GCF}}(\mathbf{W})$ can be obtained by taking the second order derivative with respect to **W**

$$\frac{\partial^2 \mathbf{J_{GCF}}}{\partial w_{jk}\partial w_{il}} = 2(\mathbf{K})_{ji}(\mathbf{V}^T V)_{lk} + 2\mu\delta_{lk}(\mathbf{L}_W)_{ji} \tag{17}$$

where $\delta_{lk} = \begin{cases} 1, & \text{if } l = k; \\ 0, & \text{else} \end{cases}$.. Let $\mathbf{K} = \mathbf{K}^+ - \mathbf{K}^-$, where $\mathbf{K}^+$ and $\mathbf{K}^-$ are two symmetric matrices whose elements are all positive. Substituting **A** and b in Eq. (15) using Eqs. (16) and (17), respectively, we obtain the multiplicative updating equation for each element w_{jk} of **W**

$$w_{jk} \leftarrow w_{jk}\left[\frac{(\mathbf{KV})_{jk} + \sqrt{(\mathbf{KV})_{jk}^2 + 4\mathbf{P}_{jk}^+\mathbf{P}_{jk}^+}}{2\mathbf{P}_{jk}^+}\right] \tag{18}$$

where $\mathbf{P}^+ = \mathbf{K}^+\mathbf{W}\mathbf{V}^\mathbf{T}\mathbf{V} + \mu\mathbf{D}^\mathbf{W}\mathbf{W}$ and $\mathbf{P}^- = \mathbf{K}^-\mathbf{W}\mathbf{V}^\mathbf{T}\mathbf{V} + \mu\mathbf{S}^\mathbf{W}\mathbf{W}$.

Similarly, we can get the updating equation for each element v_{jk} in **V** by applying Theorem 6 to the quadratic from $\mathbf{J_{GCF}}(\mathbf{V})$. Fixing **W**, we get

$$\left.\frac{\partial \mathbf{J_{GCF}}}{\partial v_{jk}}\right|_{\mathbf{V}=0} = (-2\mathbf{KW})_{jk}, \quad \frac{\partial^2 \mathbf{J_{GCF}}}{\partial v_{jk}\partial v_{il}} = 2\delta_{ij}(\mathbf{W}^T\mathbf{K}\mathbf{W})_{lk} + 2\lambda\delta_{lk}(\mathbf{L}_V)_{ji}$$

The updating equation for **V** is

$$v_{jk} \leftarrow v_{jk}\left[\frac{(\mathbf{KW})_{jk} + \sqrt{(\mathbf{KW})_{jk}^2 + 4\mathbf{Q}_{jk}^+\mathbf{Q}_{jk}^+}}{2\mathbf{Q}_{jk}^+}\right] \tag{19}$$

where $\mathbf{Q}^+ = \mathbf{V}\mathbf{W}^\mathbf{T}\mathbf{K}^+\mathbf{W} + \lambda\mathbf{D}^\mathbf{V}\mathbf{V}$ and $\mathbf{Q}^- = \mathbf{V}\mathbf{W}^\mathbf{T}\mathbf{K}^-\mathbf{W} + \lambda\mathbf{S}^\mathbf{V}\mathbf{V}$.

5. Experimental results

In this section, we investigate the use of our proposed GCF algorithm for data clustering. To examine the performance of GCF, we compare it with several stat-of-the-art clustering algorithms on document and image corpora. First, we give the descriptions of the datasets and evaluation metrics. Then the performance comparisons and results analysis are presented.

5.1. Data corpora

Our empirical studies on clustering were accomplished on different kinds of datasets: two real world document corpora (i.e., TDT2 and Reuters) and two images databases (i.e., COIL20 and PIE). The statistics of them are summarized in Table 1. In particular, the documents in both of TDT2 and Reuters have been manually clustered based on their topics, which are ideal for examining the clustering performance. Detailed descriptions about the four datasets are shown as follows.

- TDT2: This corpus is comprised of 11,201 on-topic documents with 96 semantic topics. It collects various documents mainly from six predominant news agencies, including two radio programs (VOA and PRI), two television programs (CNN and ABC), and two newswires (APW and NYT). We remove the stop words and delete the terms that appear too few times for the whole corpus [11] (e.g., the terms covered in less than ten documents). Those documents containing multiple topics are eliminated, and the topics with more than or equal to ten documents are kept, thus in total leaving us with 10,021 documents that can be grouped into 56 clusters.
- Reuters: This corpus contains 21,578 documents which are grouped into 135 clusters. Compared with TDT2 corpus, the Reuters corpus is more difficult for clustering. In TDT2, each document has a unique category label, and the content of each cluster is narrowly defined, whereas in Reuters, many documents have multiple category labels, and documents in each cluster have a broader variety of content. In our test, we discarded documents with multiple category labels, and only select the categories with more than 10 documents. This lefts us with 8213 documents in total [11].

  In both of the two document corpora, the stop words are removed and each document is treated as a term-frequency vector in the term-space.
- COIL20: The COIL20 database is composed of 20 subjects with 1440 images altogether. And each subject has 72 images. All the images are scaled to 32×32 pixel, and the images of each subjects were taken 5 degrees apart as the subject is rotated on a turntable, thus each image is represented by a 1024 dimensional feature vector.
- PIE: The CMU-PIE database is composed of 68 subjects with 41,368 face images altogether. Original images are normalized in scale and orientation to make the two eyes be aligned at the same position. The size of each cropped image is 32×32, with 256 Gy levels per pixel. Each person is under 13 different poses, 43 different illumination conditions, and with four different expressions. In our test, we fixed the pose and expression and then have 11,554 images under different lighting conditions.

Table 1
Statistics of the datasets.

Datasets	Domain	Instances	Features	Classes
TDT2	Document	10,021	36,771	56
Reuters	Document	8,213	18,933	41
COIL20	Image	1,440	1,024	20
PIE	Image	11,554	1,024	68

5.2. Evaluation metrics

To examine the clustering performance of our method, we choose two popular evaluation metrics: the accuracy (AC) and the normalized mutual information (NMI) [2]. Usually, the clustering result is evaluated by comparing the gained cluster label of each document with its ground truth label. Given a document $\mathbf{x}_i$, let r_i and s_i be the obtained cluster label and the label provided by the corpus in respective, and then AC is defined as

$$\mathrm{AC} = \frac{\sum_{i=1}^{n} \delta(s_i, map(r_i))}{N} \tag{20}$$

where N is the total number of documents, and $\delta(x, y)$ is the function that equals 1 when $x = y$ and is 0 otherwise. $map(r_i)$ is a mapping function which maps each cluster label to an equivalent given label. The Kuhn–Munkres algorithm [27] is used for the best mapping. The greater the accuracy, the better the clustering quality.

For NMI, let C and C' denote the sets of clusters from the ground truth and the algorithms respectively, then the mutual information between them is expressed by

$$MI(C, C') = \sum_{c_i \in C, c_j' \in C'} p(c_i, c_j') \times \log_2 \frac{p(c_i, c_j')}{p(c_i) \times p(c_j')} \tag{21}$$

where the probabilities $p(c_i)$, $p(c_j')$ denote that to what extent a document arbitrarily selected from the corpus belongs to the clusters c_i and c_j', respectively. And $p(c_i, c_j')$ is the joint probability that the arbitrarily chosen document belongs to the cluster c_i and c_j' at the same time. To simplify comparisons among different cluster sets, we use the normalized mutual information (NMI),

Table 2
Clustering accuracy result on TDT2 corpus.

k	2	3	4	5	6	7	8	9	10	Avg.
KM	91.83	84.23	89.43	77.68	76.32	73.82	71.58	70.87	70.13	78.43
NMF	85.69	79.51	76.58	69.34	71.43	69.12	66.80	67.23	66.48	72.46
CF	86.13	79.67	78.29	73.63	74.82	70.53	67.04	67.35	68.65	74.06
DRCC	95.54	89.21	86.83	83.07	83.76	79.45	74.56	72.13	70.57	81.68
LCCF	94.23	88.53	86.65	83.41	81.26	79.12	74.28	**75.18**	69.24	81.32
GCF	**96.21**	**91.04**	**89.45**	**87.66**	**84.61**	**80.39**	**75.20**	73.28	**71.86**	**83.30**
KM_{NCW}	93.28	85.32	91.24	84.39	83.67	83.12	82.23	80.14	77.78	84.57
NMF_{NCW}	98.43	95.91	95.05	87.73	91.03	87.72	89.24	88.74	89.02	91.43
CF_{NCW}	98.34	96.64	96.01	89.27	92.32	89.46	87.24	87.84	86.64	91.53
$DRCC_{NCW}$	98.69	97.52	97.10	96.35	97.31	94.54	95.38	94.15	95.37	96.27
$LCCF_{NCW}$	98.47	97.49	97.67	97.38	97.02	96.47	96.01	96.95	95.74	97.02
GCF_{NCW}	**99.12**	**97.63**	**98.51**	**98.40**	**98.11**	**97.94**	**97.36**	**97.71**	**96.25**	**97.89**

Table 3
Clustering normalized mutual information result on TDT2 corpus.

k	2	3	4	5	6	7	8	9	10	Avg.
KM	80.28	76.65	79.71	70.39	73.62	72.83	71.29	71.83	70.67	74.14
NMF	65.72	63.68	68.70	62.35	67.51	65.87	67.89	67.73	66.96	66.27
CF	67.28	68.91	69.78	64.65	67.86	66.54	68.31	69.43	69.81	68.06
DRCC	84.67	78.33	78.71	73.53	75.25	71.40	72.23	71.76	70.47	75.15
LCCF	84.18	78.23	77.58	72.04	74.86	73.32	69.73	70.42	69.59	74.43
GCF	**85.94**	**81.33**	**80.64**	**76.56**	**77.41**	**75.17**	**72.58**	**72.20**	**71.85**	**77.08**
KM_{NCW}	93.12	82.24	84.56	79.39	75.41	74.78	77.54	75.49	74.63	79.68
NMF_{NCW}	90.56	88.12	91.36	84.43	87.71	83.62	85.47	83.64	82.28	86.35
CF_{NCW}	92.97	87.53	90.34	82.72	85.86	81.97	84.54	83.65	81.39	85.66
$DRCC_{NCW}$	94.56	93.85	94.03	91.34	92.43	90.16	92.51	92.03	91.28	92.47
$LCCF_{NCW}$	94.10	93.87	93.82	90.65	93.27	92.78	93.14	93.25	92.46	93.04
GCF_{NCW}	**95.82**	**94.25**	**94.14**	**92.36**	**94.45**	**93.68**	**93.50**	**93.73**	**93.32**	**93.92**

Table 4
Clustering accuracy result on Reuter corpus.

k	2	3	4	5	6	7	8	9	10	Avg.
KM	81.54	70.26	64.35	59.87	59.29	55.46	47.58	45.17	46.32	58.87
NMF	83.15	72.34	69.27	59.32	58.76	54.85	46.87	45.78	47.40	59.75
CF	83.34	72.59	70.34	61.64	61.74	55.92	47.21	46.26	49.37	60.93
DRCC	86.31	77.94	76.42	70.30	68.17	62.71	58.65	56.14	57.53	68.24
LCCF	85.28	76.62	76.18	71.35	67.95	61.47	59.98	57.31	58.86	68.33
GCF	**87.36**	**78.97**	**78.43**	**72.16**	**68.45**	**63.21**	**61.79**	**58.32**	**59.11**	**69.76**
KM_{NCW}	88.74	83.87	79.52	71.47	71.35	64.58	59.22	57.83	58.16	70.53
NMF_{NCW}	88.84	83.65	78.76	74.34	72.61	68.58	61.21	59.14	61.37	72.06
CF_{NCW}	88.78	84.52	79.87	74.40	72.83	69.84	63.45	61.74	60.45	72.88
$DRCC_{NCW}$	89.03	84.86	79.94	75.25	74.10	72.52	69.16	65.30	66.41	75.17
$LCCF_{NCW}$	88.82	84.76	81.13	76.07	76.68	74.86	71.58	71.06	65.74	76.74
GCF_{NCW}	**89.77**	**85.40**	**81.91**	**78.26**	**79.04**	**77.19**	**74.53**	**73.78**	**67.37**	**78.58**

Table 5
Clustering normalized mutual information result on Reuter corpus.

k	2	3	4	5	6	7	8	9	10	Avg.
KM	42.53	41.39	46.47	43.16	48.62	46.76	39.84	39.83	46.58	43.91
NMF	43.67	42.10	48.37	43.24	48.76	45.27	38.24	39.74	45.93	43.92
CF	44.20	42.36	51.61	43.57	49.64	46.71	38.73	40.26	46.83	44.88
DRCC	49.87	47.73	53.80	48.14	51.64	48.15	43.07	44.36	48.79	48.39
LCCF	50.16	46.34	54.47	49.25	51.96	49.38	46.48	45.83	50.43	49.37
GCF	**52.28**	**48.30**	**58.51**	**52.02**	**53.50**	**53.02**	**49.34**	**46.49**	**51.78**	**51.69**
KM_{NCW}	**62.57**	64.20	65.15	53.27	57.68	53.16	45.76	45.62	53.30	55.63
NMF_{NCW}	61.28	64.45	63.57	55.24	57.96	55.61	46.51	46.96	54.62	56.24
CF_{NCW}	61.21	64.87	63.78	56.12	58.87	55.98	47.24	47.43	54.35	56.65
$DRCC_{NCW}$	58.88	64.43	63.52	56.46	59.71	57.07	53.15	52.87	55.44	57.95
$LCCF_{NCW}$	60.23	65.42	65.78	57.85	63.15	61.40	55.27	53.29	56.67	59.89
GCF_{NCW}	61.35	**67.28**	**67.80**	**59.21**	**65.56**	**62.85**	**56.49**	**54.31**	**57.89**	**61.42**

Table 6
Clustering accuracy result on COIL20 dataset.

k	2	3	4	5	6	7	8	9	10	Avg.
KM	92.71	79.35	73.19	71.67	67.78	68.34	66.13	66.23	64.60	72.22
NMF	89.84	77.80	73.01	70.36	65.20	64.64	65.16	64.87	65.37	70.69
CF	89.72	79.34	73.04	71.33	75.21	63.85	64.64	62.86	62.15	71.34
DRCC	91.04	83.42	80.36	75.15	77.74	70.13	71.67	67.42	**68.97**	76.21
LCCF	90.74	84.22	78.14	74.46	79.59	70.08	71.64	67.87	65.71	75.82
GCF	**92.48**	**85.36**	**82.69**	**79.23**	**82.90**	**73.62**	**75.51**	**70.02**	68.44	**78.91**
KM_{NCW}	91.65	78.28	74.11	70.49	66.42	65.95	66.07	62.33	64.82	71.12
NMF_{NCW}	89.61	76.60	72.30	72.21	66.37	68.25	68.15	66.78	66.59	71.87
CF_{NCW}	88.39	81.51	75.97	72.48	76.74	65.33	64.26	64.35	61.12	72.24
$DRCC_{NCW}$	90.32	84.21	81.76	76.45	78.84	73.38	72.65	69.27	70.54	77.49
$LCCF_{NCW}$	89.90	84.53	81.12	75.60	81.63	72.66	71.45	69.91	67.20	77.11
GCF_{NCW}	**91.67**	**86.55**	**83.89**	**80.15**	**85.31**	**77.09**	**74.78**	**73.10**	**72.17**	**80.52**

Table 7
Clustering normalized mutual information result on COIL20 dataset.

k	2	3	4	5	6	7	8	9	10	Avg.
KM	79.64	66.11	67.56	68.95	71.51	72.17	71.32	72.39	70.57	71.13
NMF	71.25	63.42	67.87	66.07	68.34	70.14	70.40	71.65	**71.89**	69.00
CF	71.13	63.21	66.38	67.67	65.33	66.67	67.28	66.40	66.27	66.70
DRCC	77.29	74.57	75.14	72.26	72.86	73.42	73.89	70.38	69.40	73.25
LCCF	74.51	68.69	70.63	72.22	68.81	70.57	70.67	69.86	68.69	70.52
GCF	**80.40**	**76.35**	**77.43**	**78.56**	**74.89**	**75.31**	**76.45**	**72.71**	70.63	**75.86**
KM_{NCW}	77.27	65.46	66.41	65.85	66.28	65.75	67.15	66.23	68.24	67.63
NMF_{NCW}	70.85	62.62	66.23	70.03	67.56	70.48	71.85	71.89	71.24	69.19
CF_{NCW}	67.63	64.35	63.77	65.28	65.09	67.72	69.20	68.39	67.13	66.51
$DRCC_{NCW}$	76.92	74.28	73.57	76.86	72.67	74.63	75.49	73.64	71.32	74.38
$LCCF_{NCW}$	73.46	71.25	69.82	73.88	70.61	72.35	73.58	72.31	70.37	71.95
GCF_{NCW}	**78.58**	**76.35**	**75.77**	**78.01**	**73.53**	**75.90**	**77.02**	**74.19**	**72.48**	**75.76**

Table 8
Clustering accuracy result on PIE dataset.

k	2	3	4	5	6	7	8	9	10	Avg.
KM	61.34	48.57	49.76	44.96	44.58	45.16	45.90	42.74	43.25	47.36
NMF	56.24	53.37	54.53	52.38	51.59	52.96	53.50	51.16	51.83	53.06
CF	57.23	58.14	58.36	58.89	57.63	57.80	55.97	57.26	56.85	57.57
DRCC	66.38	63.40	65.65	62.04	61.67	64.50	63.79	61.71	62.15	63.48
LCCF	62.89	58.42	60.12	60.61	59.01	59.36	56.12	57.45	57.07	59.01
GCF	**67.14**	**69.25**	**71.71**	**72.86**	**69.67**	**70.27**	**64.03**	**65.39**	**64.10**	**68.27**
KM_{NCW}	62.76	53.24	51.78	45.68	43.16	46.38	42.28	44.03	42.34	47.96
NMF_{NCW}	59.85	56.61	57.18	54.89	52.50	53.64	55.90	54.38	54.67	55.51
CF_{NCW}	59.06	58.74	58.30	58.57	57.21	57.24	55.19	54.53	51.25	56.68
$DRCC_{NCW}$	72.36	68.53	69.56	65.20	64.13	63.40	64.47	62.81	64.29	66.08
$LCCF_{NCW}$	60.86	59.81	59.23	61.42	60.21	61.57	58.23	60.95	60.14	60.27
GCF_{NCW}	**77.06**	**70.74**	**71.29**	**74.88**	**72.60**	**73.75**	**68.75**	**69.80**	**67.52**	**71.82**

Table 9
Clustering normalized mutual information result on PIE dataset.

k	2	3	4	5	6	7	8	9	10	Avg.
KM	17.64	26.78	33.63	35.87	41.25	46.74	49.27	50.86	51.27	39.26
NMF	27.57	29.96	28.36	32.84	34.21	35.68	37.78	34.54	34.83	32.86
CF	21.27	37.14	47.62	51.86	56.68	59.37	60.72	63.61	65.38	51.52
DRCC	37.23	44.08	44.79	51.54	56.57	64.40	65.15	63.31	64.24	54.59
LCCF	30.62	40.53	51.50	54.45	57.06	61.11	62.61	64.83	67.40	54.46
GCF	**38.41**	**48.36**	**54.75**	**58.39**	**62.94**	**67.66**	**67.20**	**70.24**	**75.62**	**60.39**
KM_{NCW}	19.14	30.87	41.65	38.73	43.96	50.59	49.90	52.63	53.14	42.29
NMF_{NCW}	30.47	32.56	31.12	38.98	45.14	50.78	55.69	55.13	56.24	44.01
CF_{NCW}	18.83	34.67	47.48	52.46	58.16	60.35	61.19	62.86	63.98	51.11
$DRCC_{NCW}$	38.13	44.67	44.42	51.85	59.67	63.67	67.23	64.87	65.50	55.56
$LCCF_{NCW}$	31.24	38.37	52.19	55.07	58.46	60.86	61.53	63.49	65.77	54.11
GCF_{NCW}	**42.94**	**46.52**	**55.81**	**57.39**	**65.73**	**70.38**	**71.08**	**73.12**	**77.64**	**62.29**

which is formulated as follows:

$$NMI(C,C') = \frac{MI(C,C')}{\max(H(C),H(C'))} \tag{22}$$

where $H(C)$ and $H(C')$ denote the entropies of C and C', respectively. The normalized value varies between 0 and 1. The greater the NMI, the better the clustering quality.

5.3. Performance evaluations and comparisons

To show the data clustering performance, we compare our algorithm with other related methods on four datasets. The algorithms that we evaluated are listed below:

- Traditional k-means clustering method (KM in short).
- Nonnegative Matrix Factorization based clustering (NMF in short) [28].
- Concept Factorization based clustering (CF in short) [2].
- Dual regularized co-clustering method based on semi-nonnegative matrix tri-factorization (DRCC in short) [19].
- Locally Consistent Concept Factorization based clustering (LCCF in short) [11].
- Our proposed Dual-graph regularized Concept Factorization (GCF in short).

In addition to the original form of all the above algorithms, we also implement the normalized-cut weighted form (NCW) suggested by [2,11,28]. When the data set is unbalanced, the NCW weighting can automatically reweight the samples which lead to

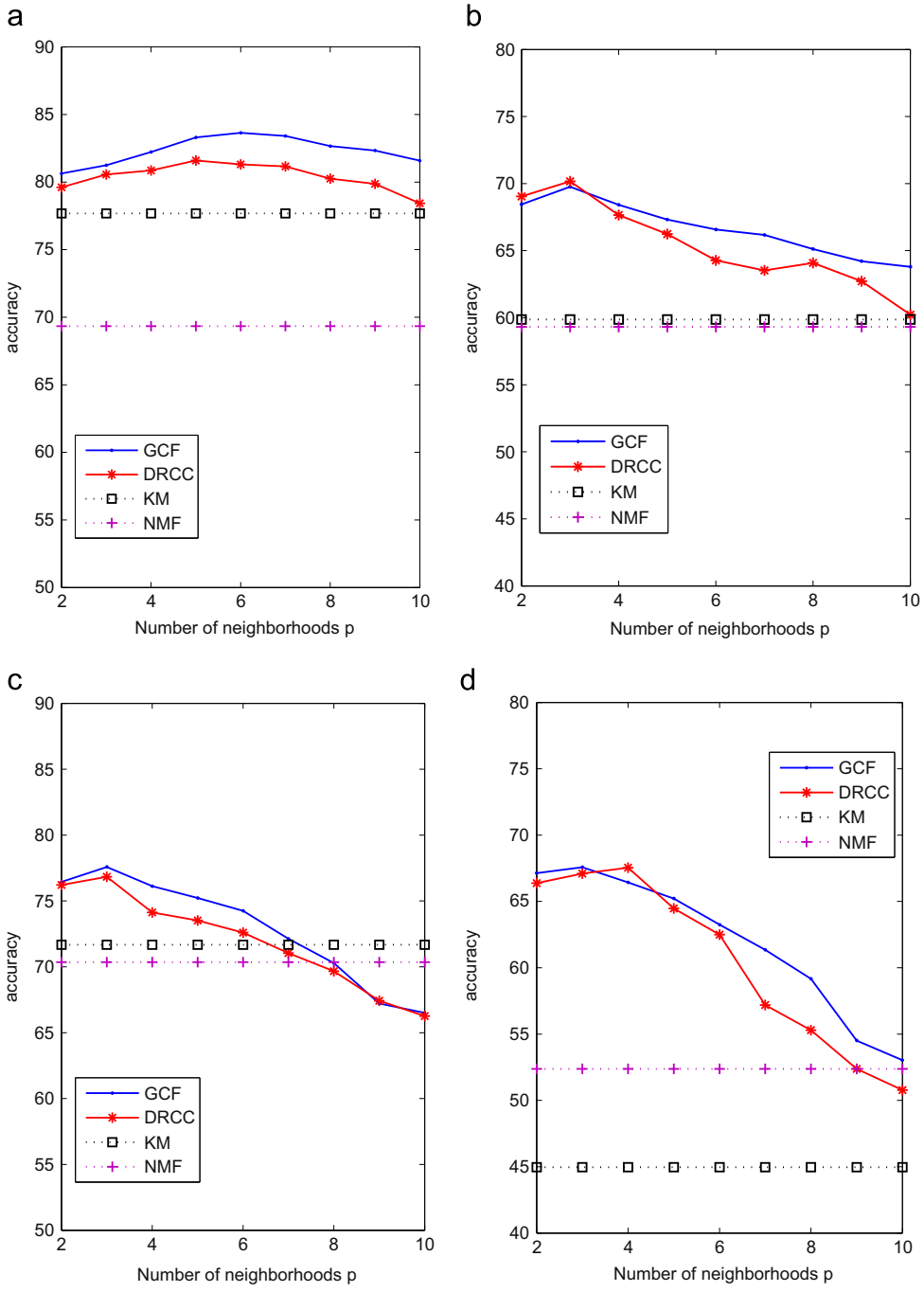

Fig. 1. The performance of our proposed algorithm varies with the size of neighborhood.

better clustering performance [2,11,28]. Using the original data matrix $\mathbf{X}$ to conduct document clustering can be considered as the original form. And using the weighted data matrix $\mathbf{X}' = \mathbf{X}\mathbf{D}^{-1/2}$, where $\mathbf{D} = \mathrm{diag}(\mathbf{X}^T\mathbf{X}\mathbf{e})$ to conduct document clustering can be considered as the NC weighted variation. The weighted form of GCF is derived in Appendix B. Essentially, we have compared six approaches (KM, NMF, CF, DRCC, LCCF, and GCF) and their NCW versions in the experiment. For the algorithms to which the kernel trick can be applied (i.e., KM, CF, LCCF, GCF and their NCW versions), we also implement their kernelized versions with degree 2 polynomial kernel following the literature [11].

The evaluations were conducted for the cluster numbers ranging from 2 to 10. For each given cluster number k, 20 test runs were conducted on different randomly chosen clusters from the corpus, the final performance scores were obtained by averaging the scores from the 20 test runs in Tables 2–9 for the TDT2, Reuter, COIL20 and PIE datasets, respectively.

We set the two parameters in the method of LCCF following the literature [11]. And in the DRCC and GCF methods, each method has three parameters: the number of nearest neighbors p and the regularization parameters λ and μ. Throughout our experiments, we empirically set the number of nearest neighbors p to 5. Furthermore, the regularization parameter λ is also set to be the same value as the regularization parameter μ for simplicity, and the value of the regularization parameter is set to 100. From the results shown in Tables 2–9, we can observe the following:

- DRCC, LCCF and our proposed method GCF consider the geometrical structure information contained in the data and commonly achieve good performance. This suggests that the underlying manifold structure of the data is useful in data clustering. Also, co-clustering the features and data points together, the clustering of features can lead to improvement in the clustering of data points in the method of DRCC and GCF. Our proposed method GCF can achieve better clustering performance than DRCC. The reason for this can be explained that GCF method is based on CF which mainly strives to address the limitations and meanwhile inherits all the strengths of the NMF-based method, such as better semantic interpretation and easily derived clustering results.

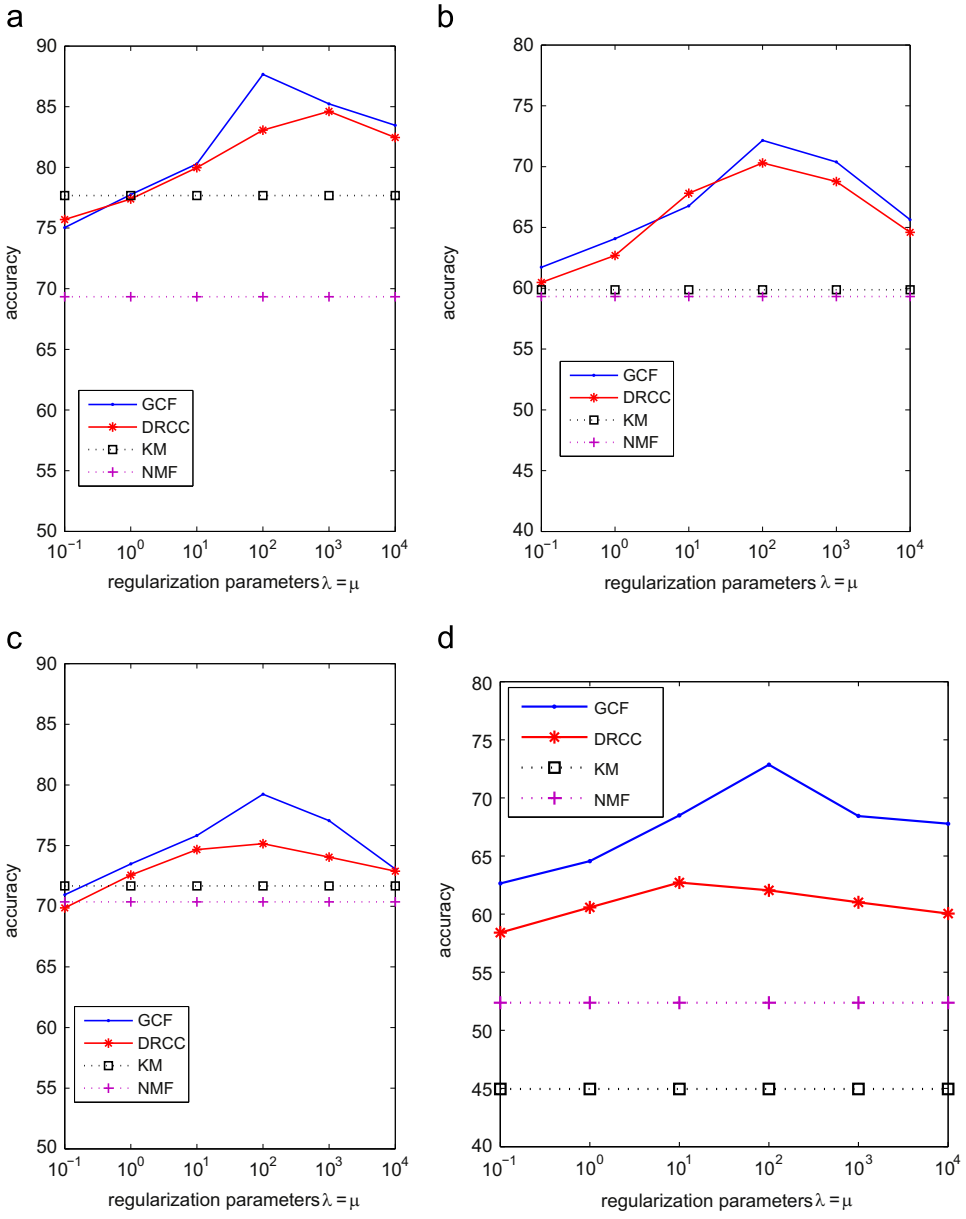

Fig. 2. The performance of our proposed algorithm varies with the regularization parameters.

- Regardless of the data, we can see that GCF_{NCW} always has the best performance. This shows that by simultaneously considering the intrinsic geometrical structure of both data manifold and feature manifold, GCF have more discriminating power than other methods.
- Clustering performances on TDT2 and Reuters document datasets are almost enhanced using NCW weighting scheme for all the algorithm. However, NCW weighting method does little help to improve the clustering performance on COIL20 and PIE. This phenomenon reveals that NCW weighting is beneficial to document clustering rather than image clustering.

5.4. Parameters selection

Our GCF algorithm has three essential parameters which are the same with the DRCC method: the number of nearest neighbors p and the regularization parameters λ and μ. We will investigate the sensitivity with respect to the regularization parameter $\lambda(=\mu)$. When we vary the value of λ, we keep the parameter p fixed at 5 in the methods of DRCC and GCF. Also when we vary the value of p, we keep the regularization parameter $\lambda(=\mu)$ fixed at 100 in the methods of DRCC and GCF. We conduct four experiments on the TDT2 and Reuters, COIL20 and PIE datasets to test the sensitivity of our proposed algorithm to the selection of these parameters, and the results are shown in Figs. 1 and 2. From these figures, we can clearly see that:

- From the results shown in Fig. 1, we can observe that the clustering results of our proposed algorithm decreases as the size of neighborhood p increases. Since the graph constructed with relatively large size of neighborhood cannot reflect the underlying manifold structures of datasets.
- The performance of our proposed algorithm is very stable with respect to the value of two regularization parameters λ and μ.

6. Conclusion

In this paper, we proposed a novel algorithm, called dual-graph regularized concept factorization (GCF), which simultaneously considers the geometric structures of both data manifold and feature manifold. As an extension of GCF, we extend that our proposed method can also be apply to the negative dataset. Since our proposed algorithm: GCF can effectively make use of the structure information contained in data as well as features, it has more discriminating power than NMF, CF and LCCF. Then we developed the iterative updating optimization schemes for GCF, and provided the convergence proof of the optimization scheme. Finally, we provided a variety of experiments on TDT2 and Reuters document datasets, COIL20 and PIE image datasets to demonstrate the effectiveness of our proposed algorithm, from which we also find that our proposed method have high parameter stability.

Acknowledgment

This work is partially supported by the National Natural Science Foundation of China under Grant nos. 61373063, 61233011, 61125305, 61375007, 61220301, and by National Basic Research Program of China under Grant no. 2014CB349303. Also this work is supported in part by the Natural Science Foundation of Jiangsu Province (BK20130868), the Natural Science Research Foundation for Jiangsu Universities (13KJB510022), and the Talent Introduction Foundation and Natural Science Foundation of Nanjing University of Posts and Telecommunications (NY212014, NY212039).

Appendix A. (proof of Theorem 5)

To prove Theorem 5, we need to show that the objective function $\mathbf{J_{GCF}}$ in Eq. (9) is nonincreasing under the updating rules stated in Eqs. (12) and (13). Since the third term of $\mathbf{J_{GCF}}$ is only related to $\mathbf{W}$, we have exactly the same update formula for $\mathbf{V}$ in (our proposed method) as the LCCF. Thus, we can use the convergence proof of LCCF to show that $\mathbf{J_{GCF}}$ is nonincreasing under the update step in Eq. (13). Please see [11] for details. Now, we make use of an auxiliary function similar to that used in the EM algorithm [26] to prove the convergence of Theorem 5. We begin with the definition of the auxiliary function.

Definition 1. The function $G(w,w')$ is an auxiliary function for $F(w)$, if the $G(w,w')\geq F(w)$ and $G(w,w)=F(w)$ are satisfied.

The auxiliary function is very useful because of the following lemma.

Lemma 1. *if G is an auxiliary function of F, then F is nonincreasing under the update*

$$w^{(K+1)}=\arg\min_{w} G(w,w^{(K)}) \tag{23}$$

Proof. $F(w^{(K+1)})\leq G(w^{(K+1)},w^{(K)})\leq G(w^{(K)},w^{(K)})=F(w^{(K)})$.

Next we will show that the updating rule for $\mathbf{W}$ in Eq. (12) is exactly the update in Eq. (23) with a proper auxiliary function.

Considering any element w_{ab} in $\mathbf{W}$, we use F_{ab} to denote the part of $\mathbf{J_{GCF}}$ which is only relevant to w_{ab}. It is easy to check that

$$\mathbf{F'_{ab}}=\left(\frac{\partial \mathbf{J_{GCF}}}{\partial \mathbf{W}}\right)_{\mathbf{ab}}=(-2\mathbf{KV}+2\mathbf{KWV^TV}+2\mu\mathbf{L_W W})_{\mathbf{ab}},$$

$$\mathbf{F''_{ab}}=2(\mathbf{K})_{\mathbf{aa}}(\mathbf{V^TV})_{\mathbf{bb}}+2\mu(\mathbf{L_W})_{\mathbf{aa}}$$

Since our update is essentially element-wise, it is sufficient to show that each F_{ab} is nonincreasing under the update step of Eq. (12). □

Lemma 2. *Function*

$$G(w,w_{ab}^{(K)})=F_{ab}(w_{ab}^{(K)})+F'_{ab}(w_{ab}^{(K)})(w-w_{ab}^{(K)}) + \frac{(\mathbf{KWV}^T\mathbf{V})_{ab}+\mu(\mathbf{D}^W\mathbf{W})_{ab}}{w_{ab}^{(K)}}(w-w_{ab}^{(K)})^2 \tag{24}$$

is an auxiliary function for F_{ab}.

Proof. Since $G(w,w)=F_{ab}(w)$ is obvious, we need show that $G(w,w_{ab}^{(K)})\geq F_{ab}(w)$. To do this, we compare the Taylor series expansion of $F_{ab}(w)$

$$F_{ab}(w)=F_{ab}(w_{ab}^{(K)})+F'_{ab}(w_{ab}^{(K)})(w-w_{ab}^{(K)}) +[(\mathbf{K})_{aa}(\mathbf{V}^T\mathbf{V})_{bb}+\mu(\mathbf{L}_W)_{ab}](w-w_{ab}^{(K)})^2$$

with Eq. (23) to find that $G(w,w_{ab}^{(K)})\geq F_{ab}(w)$ is equivalent to

$$\frac{(\mathbf{KWV}^T\mathbf{V})_{ab}+\mu(\mathbf{D}^W\mathbf{W})_{ab}}{w_{ab}^{(K)}}\geq(\mathbf{K})_{aa}(\mathbf{V}^T\mathbf{V})_{bb}+\mu(\mathbf{L}_W)_{aa} \tag{25}$$

We have

$$\begin{aligned}(\mathbf{KWV}^T\mathbf{V})_{ab}&=\sum_{l=1}^{k}(\mathbf{KW})_{al}(\mathbf{V}^T\mathbf{V})_{lb}\\ &\geq(\mathbf{KW})_{ab}(\mathbf{V}^T\mathbf{V})_{bb}\\ &\geq\sum_{l=1}^{k}(\mathbf{K})_{al}w_{lb}^{(K)}(\mathbf{V}^T\mathbf{V})_{bb}\geq w_{ab}^{(K)}(\mathbf{K})_{aa}(\mathbf{V}^T\mathbf{V})_{bb}\end{aligned}$$

and

$$\mu(\mathbf{D}^W\mathbf{W})_{ab} = \mu\sum_{j=1}^{M}\mathbf{D}^W_{aj}w^{(K)}_{jb} \geq \mu\mathbf{D}^W_{aa}w^{(K)}_{ab} \geq \mu(\mathbf{D}^W-\mathbf{S}^W)_{aa}w^{(K)}_{ab} = \mu(\mathbf{L}_W)_{aa}w^{(K)}_{ab}$$

Thus, Eq. (25) holds and $G(w,w^{(K)}_{ab}) \geq F_{ab}(w)$.

We can now demonstrate the convergence of Theorem 5. □

Proof of Theorem 5. Replacing $G(w,w^{(K)}_{ab})$ in Eq. (23) by Eq. (24), we get

$$w^{(K+1)}_{ab} = w^{(K)}_{ab} - w^{(K)}_{ab}\frac{F'_{ab}(w^{(K)}_{ab})}{2(\mathbf{KWV}^T\mathbf{V})_{ab}+2\mu(\mathbf{D}^W\mathbf{W})_{ab}} = w^{(K)}_{ab}\frac{(\mathbf{KV}+\mu\mathbf{S}^W\mathbf{W})_{ab}}{(\mathbf{KWV}^T\mathbf{V}+\mu\mathbf{D}^W\mathbf{W})_{ab}}$$

Since Eq. (24) is an auxiliary function, F_{ab} is nonincreasing under this updating rule. □

Appendix B. (Weighted GCF)

In this appendix, we give the solution to the weighted GCF. Let each data point has weight γ_j and $\mathbf{z_j^T}$ is the jth row vector of $\mathbf{V}$, the weighted objective function is

$$\begin{aligned}\mathbf{J_{GCF}} &= \sum_{j=1}^{N}\gamma_{\mathbf{j}}(\mathbf{x_j}-\mathbf{XWz_j})^{\mathbf{T}}(\mathbf{x_j}-\mathbf{XWz_j})+\lambda\mathrm{Tr}(\mathbf{V^TL_VV})+\mu\mathrm{Tr}(\mathbf{W^TL_WW})\\ &= \mathrm{Tr}[(\mathbf{X}-\mathbf{XWV^T})\Gamma(\mathbf{X}-\mathbf{XWV^T})^{\mathbf{T}}]+\lambda\mathrm{Tr}(\mathbf{V^TL_VV})+\mu\mathrm{Tr}(\mathbf{W^TL_WW})\\ &= \mathrm{Tr}[(\mathbf{X}\Gamma^{1/2}-\mathbf{XWV^T}\Gamma^{1/2})(\mathbf{X}\Gamma^{1/2}-\mathbf{XWV^T}\Gamma^{1/2})^{\mathbf{T}}]\\ &\quad+\lambda\mathrm{Tr}(\mathbf{V^TL_VV})+\mu\mathrm{Tr}(\mathbf{W^TL_WW})\\ &= \mathrm{Tr}[(\mathbf{X}\Gamma^{1/2}-\mathbf{XWV^T}\Gamma^{1/2})^{\mathbf{T}}(\mathbf{X}\Gamma^{1/2}-\mathbf{XWV^T}\Gamma^{1/2})]\\ &\quad+\lambda\mathrm{Tr}(\mathbf{V^TL_VV})+\mu\mathrm{Tr}(\mathbf{W^TL_WW})\\ &= \mathrm{Tr}[(\mathbf{I}-\mathbf{W'V'^T})^{\mathbf{T}}\mathbf{K'}(\mathbf{I}-\mathbf{W'V'^T})]+\lambda\mathrm{Tr}(\mathbf{V'^TL'_VV'})+\mu\mathrm{Tr}(\mathbf{W'^TL'_{W}W'})\end{aligned}$$

where Γ is the diagonal matrix consists of γ_j, $\mathbf{W'}=\Gamma^{-1/2}\mathbf{W}$, $\mathbf{V'}=\Gamma^{1/2}\mathbf{V}$,$\mathbf{L'_V}=\Gamma^{-1/2}\mathbf{L_V}\Gamma^{-1/2}$,

$\mathbf{L'_W}=\Gamma^{1/2}\mathbf{L_W}\ \Gamma^{1/2}$ and $\mathbf{K'}=\Gamma^{1/2}\mathbf{K}\Gamma^{1/2}$. Notice that the above equation has the same form as Eq. (10), so the same algorithm can be used to find the solution.

References

[1] A. Jain, M. Murty, P. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.

[2] W. Xu, Y. Gong. Document clustering by concept factorization, in: Proceedings of International Conference on Research and Development in Information Retrieval (SIGIR'04), Sheffield, UK, July 2004, pp. 202–209.

[3] D. Cai, X. He, J. Han, T. Huang, Graph regularized nonnegative matrix factorization for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 1548–1560.

[4] Y. Yang, H. Shen, F. Nie, et al. Nonnegative spectral clustering with discriminative regularization, in: Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI' 11), pp. 555–560.

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

[6] Ping Li, Chun Chen, JiaJun Bu, Clustering analysis using manifold kernel concept factorization, Neurocomputing 87 (2012) 120–131.

[7] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Inf. Process. Syst. 14 (2001) 585–591.

[8] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.

[9] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.

[10] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from examples, J. Mach. Learn. Res. 7 (2006) 2399–2434.

[11] Deng Cai, Xiaofei He, Han Jiawei, Locally consistent concept factorization for document clustering, IEEE Trans. Knowl. Data Eng. 23 (6) (2011) 902–913.

[12] Ping Li, Jiajun Bu, Yi Yang, Rongrong Ji, Chun Chen, Cai Deng, Discriminative orthogonal nonnegative matrix factorization with flexibility for data representation, Expert Syst. Appl. 41 (4) (2014) 1283–1293 (Part 1).

[13] Ping Li, Jiajun Bu, Chun Chen, Can Wang, Deng Cai.Subspace learning via locally constrained A-optimal nonnegative projection, Neurocomputing 115 (2013) 49–62.

[14] I.S. Dhillon, Co-clustering documents and words using bipartite spectral graph partitioning, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2001, pp. 269–274.

[15] I.S. Dhillon, S. Mallela, D.S. Modha, Information-theoretic co-clustering, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2003, pp. 89–98.

[16] C.H.Q. Ding, T. Li, W. Peng, H. Park, Orthogonal nonnegative matrix tri-factorization for clustering, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2006, pp. 126–135.

[17] Fanhua Shang, L.C. Jiao, Fei Wang, Graph dual regularization non-negative matrix factorization for co-clustering, Pattern Recognit. 45 (2012) 2237–2250.

[18] V. Sindhwani, J. Hu, A. Mojsilovic, Regularized co-clustering with dual supervision, Adv. Neural Inf. Process. Syst. 21 (2009) 1505–1512.

[19] Q. Gu, J. Zhou, Co-clustering on manifolds, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2009, pp. 359–368.

[20] Jiajun Bu Ping Li, Chun Chen, Zhanying He, Cai Deng, Relational multi-manifold co-clustering, IEEE Trans. Cybern. 43 (6) (2013) 1871–1881.

[21] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, Adv. Neural Inf. Process. Syst. 13 (2001) 556–562.

[22] M. Catral, L. Han, M. Neumann, R. Plemmons, On reduced rank nonnegative matrix factorization for symmetric nonnegative matrices, Linear Algebra Appl. 393 (2004) 107–126.

[23] C.-J. Lin, On the convergence of multiplicative update algorithms for non-negative matrix factorization, IEEE Trans. Neural Networks 18 (6) (2007) 1589–1596.

[24] Quanquan Gu, Chris Ding, Jiawei Han, On Trivial Solution and Scale Transfer Problems in Graph Regularized NMF, IJCAI'11, pp. 1288–1293.

[25] F. Sha, Y. Lin, L.K. Saul, D.D. Lee, Multiplicative updates for nonnegative quadratic programming, Neural Comput. 19 (8) (2007) 2004–2031.

[26] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. Ser. B: Methodological 39 (1) (1977) 1–38.

[27] L. Lovasz, M. Plummer, Matching Theory, Akad´emiai Kiad´o, North Holland, Budapest, 1986.

[28] W. Xu, X. Liu, Y. Gong. Document clustering based on nonnegative matrix factorization, in: Proceedings of International Conference on Research and Development in Information Retrieval (SIGIR'03), Toronto, Canada, Aug. 2003, pp. 267–273.

Jun Ye received the BS degree in Mathematics from XuZhou Normal University in 2003 and the MS degree in Applied Mathematics from Nanjing University of Science and Technology (NUST) in 2005. Now he is pursuing the Ph.D. degree in Pattern Recognition and Intelligent Systems from Nanjing University of Science and Technology (NUST). His current research interests include pattern classification, face recognition and image processing.

Zhong Jin received the BS degree in Mathematics, the MS degree in Applied Mathematics and the PhD degree in Pattern Recognition and Intelligence System from Nanjing university of Science and Technology (NUST), China, in 1982, 1984, and 1999, respectively. He is a professor in the Department of Computer Science, NUST. He had a stay of 15 months as research assistant at the Department of Computer Science and engineering, the Chinese University of Hong Kong from 2000 to 2001. He visited the Laboratoire HEUDIASYC, Universite de Technologie de Compiegne, France, from October 2001 to July 2002. He visited the Centrede Visioper Computador, Universit at Autonomade Barcelona, Spain, as the Ramony Cajal program Research Fellow from September 2005 to October 2005. His current interests are in the areas of pattern recognition, computer vision, face recognition, facial expression analysis and content-based image retrieval.