Pattern Recognition 55 (2016) 172–182

Contents lists available at ScienceDirect

Pattern Recognition

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

Global discriminative-based nonnegative spectral clustering

Ronghua Shang*, Zhu Zhang, Licheng Jiao, Wenbing Wang, Shuyuan Yang

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 20 December 2014
Received in revised form
17 October 2015
Accepted 29 January 2016
Available online 8 February 2016

Keywords:
Spectral clustering
Nonnegative matrix factorization (NMF)
Global discrimination information

ABSTRACT

Based on spectral graph theory, spectral clustering is an optimal graph partition problem. It has been proven that the spectral clustering is equivalent to nonnegative matrix factorization (NMF) under certain conditions. Based on the equivalence, some spectral clustering methods are proposed, but the global discriminative information of the dataset is neglected. In this paper, based on the equivalence between spectral clustering and NMF, we simultaneously maximize the between-class scatter matrix and minimize the within-class scatter matrix to enhance the discriminating power. We integrate the geometrical structure and discriminative structure in a joint framework. With a global discriminative regularization term added into the nonnegative matrix factorization framework, we propose two novel spectral clustering methods, named global discriminative-based nonnegative and spectral clustering (GDBNSC-Ncut and GDBNSC-Rcut) These new spectral clustering algorithms can preserve both the global geometrical structure and global discriminative structure. The intrinsic geometrical information of the dataset is detected, and clustering quality is improved with enhanced discriminating power. In addition, the proposed algorithms also have very good abilities of handling out-of-sample data. Experimental results on real word data demonstrate that the proposed algorithms outperform some state-of-the-art methods with good clustering qualities.

1. Introduction

Cluster analysis is an important part of data mining and pattern recognition [1,2], which is the problem of portioning the dataset into several categories according to certain similarity measure, so that data points belonging to the same class share high similarity, while the data points belonging to different classes have low similarity [3]. Clustering algorithms have been widely applied in many fields, such as image segmentation [4], genetic information analysis [5], document analysis [6], image retrieval [7], image compression [8], and so on.

Over the past decades, spectral clustering [9–14] has gained considerable attention from both the academic and the industrial communities. Compared with conventional clustering algorithms, spectral clustering has obvious advantages. It can converge to global optimum and that it performs well for the sample space of arbitrary shape, especially suitable for non-convex dataset [15]. Spectral clustering is based on algebraic graph theory, which treats data clustering problem as a graph partitioning problem [16]. It constructs an undirected weighted graph with each node corresponds to a data point, and the weight of the edge connecting the two nodes being the similarity value between the two points [17]. Then, using certain graph cut method, we divide the graph into connected components, which are called clusters. Typical graph cut methods include normalized cut (Ncut) [18], ratio cut (Rcut) [19], minimum cut (Mcut) [20] and min–max cut (MMcut) [21]. The optimal solution of graph partition can be obtained by minimizing or maximizing the objective function of the graph cut methods [22]. However, seeking the optimal solution of graph partition criteria is often NP-hard. Spectral clustering seeks to get the relaxation solution of graph cut objective function, which is an approximate optimal solution for graph partition. The basic idea is considering a continuous relaxation form of the original problem, turning to solve the eigenvalues and eigenvectors of the graph Laplacian matrix. In this paper, we only focus on spectral clustering approaches using Ncut and Rcut as objective functions.

Nonnegative matrix factorization (NMF) [23,24] is a typical method for dimensionality reduction and matrix factorization. NMF obtains a low-dimensional approximation of the original data matrix and gets a part-based representation of the data. The biggest difference between NMF and other matrix decomposition methods (such as SVD) is that the nonnegative constraints lead to the iterative multiplicative updating rules. By biological knowledge, we know that our brain has a part-based approach for recognition and understanding. The idea of NMF is consistent with our cognitive rules of the objective world [25,26]. Therefore, NMF has a clear physical meaning and strong interpretability.

* Corresponding author. Tel.: +86 29 88202279.
E-mail address: rhshang@mail.xidian.edu.cn (R. Shang).

http://dx.doi.org/10.1016/j.patcog.2016.01.035
0031-3203/

NMF is closely related to some algorithms in machine learning and pattern recognition communities. Probabilistic latent semantic indexing (PLSI) and NMF have been proven to be equivalent [27], although they are different methods, they optimize the same objective function. Ding et al. proved that kernel k-means can be treated as an NMF problem for symmetric matrix decomposition, and that NMF equals to Ncut spectral clustering [28]. It has also been pointed out that Laplace embedding is equivalent to Rcut spectral clustering [29]. In [29], the nonnegativity constraint is rigorously enforced, a nonnegative Laplacian embedding (NLE) approach is proposed and its links with NMF algorithm are demonstrated. In [28] and [29], symmetric NMF are involved. Different from [28] and [29], the data matrix itself is considered in [30]. Under proper conditions, Lu et al. demonstrate that a relaxed Rcut spectral clustering algorithm is equivalent to nonnegative factorization of the data matrix into the product a nonnegative matrix and another nonnegative matrix with orthogonal columns. Similarly, Ncut spectral clustering is also proven to be equivalent to nonnegative factorization of the normalized data matrix [30].

Under this equivalence, four algorithms: NSC-Ncut, NSC-Rcut, NSSC-Ncut and NSSC-Rcut are proposed in [30]. These four algorithms all consider the global manifold structure of a dataset, but they fail to consider the discriminative structure which reveals the intrinsic structure of the data distribution. We know that both manifold information and discriminant information are of great importance for clustering. We expect to preserve the discriminant information of a dataset in the learning process.

In order to capture the global discriminative information of the dataset, an intuitive approach is taking the class labels as prior knowledge in the learning process. However, in unsupervised clustering, it is infeasible to get the class labels in advance. Fortunately, in recent years, we have witnessed some progresses in employing discriminative structural information under the unsupervised learning paradigm [31–38].

Discriminative cluster analysis (DCA) [31] uses discriminative features for clustering rather than generative ones. Thus, clustering in the low dimensional discriminative space is more effective and computationally efficient than clustering in principal components space. In [32], the proposed discriminative k-means algorithm performs linear discriminant analysis (LDA) subspace selection and clustering simultaneously. In [33], both the local manifold structure and the global discriminant information are preserved simultaneously through manifold discriminant learning. In [34], the proposed local discriminative and global integration clustering algorithm (LDMGI) combines the local discriminative models and manifold structure for clustering. In [35], the discriminative information and geometrical information are characterized in a weighted feature space, which can well estimate the clustering structure of the data. In [36], a new Laplacian matrix was integrated into a spectral embedded clustering framework to capture local and global discriminative information for clustering. In [37], the global discriminative regularization term in is introduced, which provides more discriminative information to enhance clustering performance. In [38], an effective feature extraction method used discriminant analysis, which facilites the learning power of the method.

These algorithms use the global discriminative information, and make their performance to be improved. However, the general global discriminative model is used in linear cases, so these algorithms cannot effectively deal with the nonlinear data. Fortunately, this problem can be solved with the development of kernel tricks [39–44]. Kernel trick has been applied to many learning algorithms, such as the kernel principal component analysis (KPCA) [39], the kernel trick for support vector machines (SVMs) [40] and the kernelized LDA [41–44]. In [41], a nonlinear method based on Fisher's discriminant was proposed, which called kernel fisher discriminant (KFD). Fisher discriminant can be computed efficiently in feature space by using the kernel trick. So KFD can be used to handle nonlinear data, and also maintains the advantages of Fisher's discriminant analysis. The results show that KFD is competitive to other state-of-the-art methods. In [42], a method to deal with nonlinear discriminant analysis using kernel function operator was proposed. It is effective for both simulated data and alternate kernels. In [43], Liang et al. proposed a method to solve kernel Fisher discriminant analysis. This method is effective and feasible in dealing with handwritten numeral characters. In [44], the method of KFD was analyzed and a more transparent KFD algorithm was proposed, in which KPCA was first performed and then LDA was used for a second feature extraction. Simulation results on CENPARMI handwritten numeral database showed the effectiveness of this algorithm. Therefore, the kernelized global discriminative model can be used for nonlinear data effectively.

Inspired by these ideas, we integrate the global geometrical structure and the global discrimination structure in a joint unsupervised framework. We propose two novel spectral clustering algorithms named global discriminative-based nonnegative spectral clustering (GDBNSC-Ncut and GDBNSC-Rcut). The proposed approaches are expected to keep the connection between spectral clustering and NMF, and learn a compact data representation. This compact data representation can preserve not only the global geometric information but also has the global discriminant ability, both of which are crucial for effective clustering. Different from previous work [18,19,29,30], the proposed algorithms preserve both discriminative information and the geometrical information of the dataset, while still keeping the connection between NMF and spectral clustering.

We know that some former algorithms [28–30] just perform nonnegative matrix factorization of matrices to keep connection between spectral clustering and NMF. We go a step further by integrating discriminative information in the objective function to detect the intrinsic structure of the dataset.

It is worthwhile to highlight the main contributions of the proposed algorithms here:

1. The proposed methods do not only connect spectral clustering algorithm with NMF, but also characterize both the underlying global geometrical information and the global discriminative information of the dataset, and the proposed algorithms have good ability to handle out-of-sample data.
2. For the proposed algorithms, we give the objective functions, develop iterative multiplicative updating schemes, and analyze the convergence.
3. The remainder of this paper is organized as follows. In Section 2, we introduce some related work. In Section 3, we present the proposed algorithms, deduce iterative multiplicative updating rules, and then provide the convergence proof of the optimization scheme. Experimental part is presented in Section 4. Finally, some concluding remarks and several issues of future's work are given in Section 5.

2. Related works

In this section, we briefly review some recent work closely related to our algorithms.

2.1. Rcut spectral clustering

Let $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_N]\in\mathbb{R}^{M\times N}$ denote the data matrix, $\boldsymbol{x}_i\in\mathbb{R}^M$denotes the i-th data point, M is the dimensionality of original data, and N is the number of samples. The dataset is expected to group into K classes. We construct an undirected similarity graph $G=(V,E)$, where each node corresponds to a data

point in set $V=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_N\}$, and E denotes the edge set. We can compute a similarity weight matrix $\boldsymbol{W}\in\mathbb{R}^{N\times N}$ from the data points. In this paper, we assume the pairwise similarity being nonnegative, i.e., $W_{ij}\geq 0$, and $\boldsymbol{W}$ is symmetric.

The unnormalized graph Laplacian matrix is defined as follows:

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W} \tag{1}$$

where $\boldsymbol{D}$ is a diagonal matrix whose entries are the column (or row) sums of $\boldsymbol{W}$, $\boldsymbol{D}_{ii}=\sum_j \boldsymbol{W}_{ij}$. $\boldsymbol{L}$ is a symmetric positive semi-definite matrix.

Let the class indicator vector for the lth class C_l be $\boldsymbol{h}_l\in\mathbb{R}^N$, defined as

$$\mathbf{h}_l(i)=\begin{cases}1, & x\in C_l\\ 0, & otherwise\end{cases} \tag{2}$$

We define cluster indicator matrix $\mathbf{H}\in\mathbb{R}^{N\times K}$ as

$$\boldsymbol{H}=\left(\frac{\boldsymbol{h}_1}{\|\boldsymbol{h}_1\|},\frac{\boldsymbol{h}_2}{\|\boldsymbol{h}_2\|},...\frac{\boldsymbol{h}_K}{\|\boldsymbol{h}_K\|}\right) \tag{3}$$

Obviously, $\boldsymbol{H}^T\boldsymbol{H}=\boldsymbol{I}$.

Rcut spectral clustering solves the following problem:

$$\min_{\boldsymbol{H},\mathbf{H}^T\mathbf{H}=\mathbf{I}} tr\left(\boldsymbol{H}^T\boldsymbol{L}\boldsymbol{H}\right) \tag{4}$$

where $tr(\cdot)$ denotes the trace of a matrix.

2.2. Ncut spectral clustering

For Ncut spectral clustering, we define the cluster indicator vector as

$$\mathbf{z}_l=\boldsymbol{D}^{1/2}\boldsymbol{h}_l/\|\boldsymbol{D}^{1/2}\boldsymbol{h}_l\| \tag{5}$$

where, $\boldsymbol{Z}=(\mathbf{z}_1,\mathbf{z}_2,...,\mathbf{z}_K)$ is the cluster indicator matrix. It is obviously that $\mathbf{Z}^T\mathbf{Z}=\mathbf{I}$.

The symmetric normalized graph Laplacian matrix is defined as

$$\boldsymbol{L}_s=\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} \tag{6}$$

Ncut spectral clustering solves the following problem:

$$\min_{\boldsymbol{Z},\boldsymbol{Z}^T\boldsymbol{Z}=\mathbf{I}} tr\left(\boldsymbol{Z}^T\boldsymbol{L}_s\boldsymbol{Z}\right) \tag{7}$$

2.3. Nonnegative Laplacian Embedding (NLE)

In [29], it has been proven that nonnegative Laplacian embedding is equivalent to the following symmetric NMF:

$$\boldsymbol{W}-\boldsymbol{D}+\sigma\boldsymbol{I}\approx\boldsymbol{Q}\boldsymbol{Q}^T\ s.t.\ \boldsymbol{Q}^TQ=I,\ \ \boldsymbol{Q}\geq 0 \tag{8}$$

Consider this equivalence, nonnegative Laplacian embedding (NLE) solves the following problem:

$$\min_{\mathbf{Q}} tr\left(\mathbf{Q}^T(\boldsymbol{W}-\boldsymbol{D}+\sigma\boldsymbol{I})\mathbf{Q}\right)$$
$$s.t.\ \mathbf{Q}^TQ=I,\ \mathbf{Q}\geq 0 \tag{9}$$

where σ is the largest eigenvalue of $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$.

2.4. Nonnegative and sparse spectral clustering

Given the dataset $X\geq 0$, the similarity matrix is measured by inner product, i.e., the similarity matrix $\boldsymbol{W}=\boldsymbol{X}^T\boldsymbol{X}$. In [30], the intrinsic connection between spectral clustering and NMF has been revealed. They also prove that spectral clustering can be regarded as NMF of data matrix or scaled data matrix with orthogonal constraints.

2.4.1. Nonnegative spectral clustering for Ncut (NSC-Ncut)

We first introduce a theorem demonstrated in [30].

Theorem 1. *If the data matrix $\boldsymbol{X}\geq 0$ and the similarity matrix $\boldsymbol{W}=\boldsymbol{X}^T\boldsymbol{X}$, Ncut spectral clustering (7) is equivalent to the nonnegative matrix factorization of the scaled data matrix $\boldsymbol{D}^{-1/2}\boldsymbol{X}^T\approx\boldsymbol{Z}\boldsymbol{Y}$ subject to $\boldsymbol{Z}^TZ=I$.*

Based on Theorem 1, nonnegative spectral clustering for Ncut (NSC-Ncut) solves the following problem:

$$\min_{Z,Y}\|\boldsymbol{D}^{-1/2}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{Y}\|_F^2$$
$$s.t.\ \boldsymbol{Z}^TZ=I \tag{10}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\boldsymbol{Z}\in\mathbb{R}^{N\times K}$ and $\boldsymbol{Y}\in\mathbb{R}^{K\times M}$are two nonnegative matrices, the rows of $\boldsymbol{Z}$ serve as a clustering indicator vector for each data point, and the columns of $\boldsymbol{Z}$ are clustering indicator vector of each cluster.

2.4.2. Nonnegative spectral clustering for Rcut (NSC-Rcut)

Similarly, as for Rcut spectral clustering, the Rcut spectral clustering is relaxed and then it can be casted into an NMF problem. There is also a theorem.

Theorem 2. *If the data matrix $\boldsymbol{X}\geq 0$ and the similarity matrix $\boldsymbol{W}=\boldsymbol{X}^T\boldsymbol{X}$, Rcut spectral clustering (4) can be relaxed such that it is equivalent to the nonnegative matrix factorization of the data matrix $\boldsymbol{X}^T\approx\boldsymbol{H}\boldsymbol{Y}$ subject to $\boldsymbol{H}^TH=I$.*

Based on Theorem 2, nonnegative spectral clustering for Rcut (NSC-Rcut) solves the following problem:

$$\min_{H,Y}\|\boldsymbol{X}^T-\mathrm{HY}\|_F^2$$
$$s.t.\ \boldsymbol{H}^TH=I \tag{11}$$

Where, $\boldsymbol{H}\in\mathbb{R}^{N\times K}$and $\boldsymbol{Y}\in\mathbb{R}^{K\times M}$ are two nonnegative matrices, the rows of $\boldsymbol{H}$ serve as a clustering indicator vector for each data point, and the columns of $\boldsymbol{H}$ are clustering indicator vector of each cluster.

2.4.3. Nonnegative and sparse spectral clustering for Ncut (NSSC-Ncut)

Sparse constraints are added to the cluster indicator matrices in the nonnegative matrix factorization framework to increase robustness of spectral clustering. l_1-*norm* is used to measure sparseness, the two corresponding nonnegative and sparse spectral clustering (NSSC) algorithms are named NSSC-Ncut and NSSC-Rcut.

The objective function of NSSC-Ncut is as follows:

$$\min_{Z,Y}\frac{1}{2}\|\boldsymbol{D}^{-1/2}\boldsymbol{X}^T-\boldsymbol{Z}\boldsymbol{Y}\|_F^2+\lambda\|\boldsymbol{Z}\|_1$$
$$s.t.\ \boldsymbol{Z}^TZ=I \tag{12}$$

where $\lambda>0$ is the trade-off parameter that balances the reconstruction item and sparse item, $\|\cdot\|_1$ denotes l_1-norm, and for a matrix $\boldsymbol{B}\in\mathbb{R}^{n\times d}$, the l_1-norm of $\boldsymbol{B}$ is

$$\|B\|_1=\sum_{i=1}^{n}\sum_{j=1}^{d}|B_{ij}| \tag{13}$$

2.4.4. Nonnegative and sparse spectral clustering for Rcut (NSSC-Rcut)

Similarly, NSSC-Rcut solves the following problem:

$$\min_{H,Y}\frac{1}{2}\|\boldsymbol{X}^T-\boldsymbol{H}\boldsymbol{Y}\|_F^2+\lambda\|\boldsymbol{H}\|_1$$
$$s.t.\ \boldsymbol{H}^TH=I \tag{14}$$

where $\lambda>0$ is the trade-off parameter that balances the reconstruction item and sparse item.

3. Global discriminative-based nonnegative spectral clustering algorithm (GDBNSC)

We know that NSC-Ncut, NSC-Rcut, NSSC-Ncut and NSSC-Rcut factorize the data matrix or the normalized data matrix into the product of a nonnegative matrix and another nonnegative matrix with orthogonal columns. They just keep the equivalence between spectral clustering and NMF. The global discriminative structure of the dataset is not considered, and their clustering performance needs to be improved. To compensate this drawback, we add global discriminative regularization term into the nonnegative matrix factorization framework. Both the global geometric information and global discriminative information are preserved for clustering. We propose two novel nonnegative spectral clustering: GDBNSC-Ncut and GDBNSC-Rcut. Next, we will present their objective functions, and deduce iterative updating rules. The convergence proof of the algorithm is also given.

3.1. The global discriminative model

In order to obtain good clustering results, the discriminative information should be considered. Here discriminative analysis model is introduced. We first define a centering matrix $\boldsymbol{H}_N = \boldsymbol{I}_N - \frac{1}{N}\boldsymbol{i}_N\boldsymbol{i}_N^T$, $\boldsymbol{i}_{\mathbf{N}} \in \mathbb{R}^N$is an N-dimensional vector with all-one, $\mathbf{I}_N$is an identity matrix. Let $\boldsymbol{X}^{\sim} = \boldsymbol{X}\boldsymbol{H}_N$be the centered matrix.

We bring in the between-cluster matrix ($\mathbf{S}_B$) and the within-cluster matrix ($\mathbf{S}_W$) [32] as follows:

$$\boldsymbol{S}_B = \widetilde{\boldsymbol{X}}\boldsymbol{H}\boldsymbol{H}^T\widetilde{\boldsymbol{X}}^T \tag{15}$$

$$\boldsymbol{S}_W = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T - \widetilde{\boldsymbol{X}}\boldsymbol{H}\boldsymbol{H}^T\widetilde{\boldsymbol{X}}^T \tag{16}$$

$$\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T \tag{17}$$

Where, matrix $\mathbf{S}_T$ denotes the total scatter.

As a modern nonlinear cluster method, spectral clustering can capture the nonlinear manifold structure. However, the above global discriminative model is a linear version of fisher discriminative analysis (LFDA). Therefore, we need to kernelize the above discriminative model.

Kernel trick has been widely applied to effective analysis of nonlinear data. Using the idea of the kernel trick, the input data are mapped into an implicit feature space by a nonlinear mapping and these data are handled in the feature space [40–44].

The input space $\boldsymbol{x}_i \in \mathbb{R}^M$can be mapped into the feature space $\boldsymbol{F}$ by a nonlinear mapping function $\boldsymbol{\phi}$:

$$\boldsymbol{\phi} : \mathbb{R}^M \to \boldsymbol{F},\ \boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x}) \tag{18}$$

Then we get the following kernel discriminative model:

$$\hat{\mathbf{S}}_B = \boldsymbol{\phi}(\boldsymbol{X})\mathbf{H}_N\mathbf{H}\mathbf{H}^T\mathbf{H}_N{}^T\boldsymbol{\phi}(\boldsymbol{X})^T \tag{19}$$

$$\hat{\boldsymbol{S}}_T = \boldsymbol{\phi}(\boldsymbol{X})\boldsymbol{H}_N\boldsymbol{H}_N{}^T\boldsymbol{\phi}(\boldsymbol{X})^T \tag{20}$$

Note that $\boldsymbol{H}_N\boldsymbol{H}_N{}^T = \boldsymbol{H}_N$, so (20) can be rewritten as follows:

$$\hat{\boldsymbol{S}}_T = \boldsymbol{\phi}(\boldsymbol{X})\boldsymbol{H}_N\boldsymbol{\phi}(\boldsymbol{X})^T \tag{21}$$

The goal of clustering is to maximize the between-cluster scatter matrix and minimize the within-cluster scatter matrix simultaneously, i.e., to solve the following optimization problem.

$$\max_{\mathbf{H}} tr\left(\left(\hat{\boldsymbol{S}}_T + \mu\boldsymbol{I}_N\right)^{-1}\hat{\boldsymbol{S}}_B\right) \tag{22}$$

where, $\mu > 0$, $\mu\mathbf{I}_N$ is added to make the matrix $\left(\hat{\mathbf{S}}_T + \mu\boldsymbol{I}_N\right)$ invertible. Here we fix $\mu = 10^{-12}$.

By maximizing the above equation, we can characterize the discriminating power and get compact data representation.

Note that

$$\begin{aligned} tr(\boldsymbol{H}^T\boldsymbol{H}_N\boldsymbol{H}) &= tr\left(\boldsymbol{H}^T\left(\boldsymbol{I}_N - \frac{1}{N}\boldsymbol{i}_N\boldsymbol{i}_N{}^T\right)\boldsymbol{H}\right) \\ &= K - 1 \end{aligned} \tag{23}$$

It indicates that $tr(\boldsymbol{H}^T\boldsymbol{H}_N\boldsymbol{H})$ is a constant [45], then the above optimization problem can be rewritten as follow:

$$\begin{aligned} &\min_{\mathbf{H}} tr(\boldsymbol{H}^T\boldsymbol{H}_N\boldsymbol{H} - (\hat{\boldsymbol{S}}_T + \mu\boldsymbol{I}_N)^{-1}\hat{\boldsymbol{S}}_B) \\ &= \min_{\mathbf{H}} tr(\boldsymbol{H}^T\boldsymbol{H}_N\boldsymbol{H} - (\boldsymbol{\phi}(\boldsymbol{X})\mathbf{H}_N\boldsymbol{\phi}(\boldsymbol{X})^T + \mu\boldsymbol{I}_N)^{-1} \\ &\quad \times\boldsymbol{\phi}(\boldsymbol{X})\mathbf{H}_N\mathbf{H}\mathbf{H}^T\mathbf{H}_N{}^T\boldsymbol{\phi}(\boldsymbol{X})^T) \\ &= \min_{\mathbf{H}} tr(\boldsymbol{H}^T\boldsymbol{H}_N\mathbf{H} - \mathbf{H}^T\mathbf{H}_N{}^T\boldsymbol{\phi}(\boldsymbol{X})^T(\boldsymbol{\phi}(\boldsymbol{X})\mathbf{H}_N\boldsymbol{\phi}(\boldsymbol{X})^T \\ &\quad + \mu\boldsymbol{I}_N)^{-1}\boldsymbol{\phi}(\boldsymbol{X})\mathbf{H}_N\mathbf{H}) \\ &= \min_{\mathbf{H}} tr(\boldsymbol{H}^T(\boldsymbol{H}_N - \mathbf{H}_N{}^T\boldsymbol{\phi}(\boldsymbol{X})^T(\boldsymbol{\phi}(\boldsymbol{X})\boldsymbol{H}_N\boldsymbol{\phi}(\boldsymbol{X})^T + \mu\boldsymbol{I}_N)^{-1} \\ &\quad \times\boldsymbol{\phi}(\boldsymbol{X})\mathbf{H}_N)\mathbf{H}) \\ &= \min_{\mathbf{H}} tr\left(\boldsymbol{H}^T(\boldsymbol{H}_N - \mathbf{H}_N{}^T(\boldsymbol{H}_N + \mu(\boldsymbol{\phi}(\boldsymbol{X})^T\boldsymbol{\phi}(\boldsymbol{X}))^{-1})^{-1}\mathbf{H}_N)\mathbf{H}\right) \end{aligned} \tag{24}$$

Let $\boldsymbol{K}_1 = \boldsymbol{\phi}(\boldsymbol{X})^T\boldsymbol{\phi}(\boldsymbol{X})$ be the kernel matrix, then (24) can be rewritten as follows

$$\min_{\mathbf{H}} tr(\boldsymbol{H}^T(\boldsymbol{H}_N - \boldsymbol{H}_N{}^T(\boldsymbol{H}_N + \mu\boldsymbol{K}_1{}^{-1})^{-1}\boldsymbol{H}_N)\boldsymbol{H}) \tag{25}$$

There are various kernel functions, such as Linear kernel, Gaussian kernel, Polynomial kernel, Cosine kernel and Hyperbolic kernel [46]. We use the Gaussian kernel in this paper, and it is defined as $\boldsymbol{K}_1(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right)$.

We define $\boldsymbol{Q} = \boldsymbol{H}_N - \boldsymbol{H}_N{}^T(\boldsymbol{H}_N + \mu\boldsymbol{K}_1{}^{-1})^{-1}\boldsymbol{H}_N$, so (25) can be rewritten as follows

$$\min_{\mathbf{H}} tr(\boldsymbol{H}^T Q H) \tag{26}$$

Now, we integrate this global discriminative item and the global geometrical structure in a nonnegative matrix factorization framework. We get two novel spectral clustering algorithms: GDBNSC-Ncut and GDBNSC-Rcut. Next, we will present their objective functions respectively.

3.1.1. The objective function of GDBNSC-Rcut

The objective function of GDBNSC-Rcut is

$$\begin{aligned} &\min_{\mathbf{H}^T\mathbf{H} = \mathbf{I}} \frac{1}{2}\|\boldsymbol{X}^T - \boldsymbol{H}\boldsymbol{Y}\|_F^2 + \alpha tr(\boldsymbol{H}^T\boldsymbol{Q}\boldsymbol{H}), \\ &s.t.\quad \boldsymbol{H} \geq 0, \boldsymbol{Y} \geq 0, \boldsymbol{H}^T\boldsymbol{H} = \boldsymbol{I}, \end{aligned} \tag{27}$$

where, $\alpha \geq 0$ is the regularization parameter that balances the reconstruction error in the first term and the discriminative regularization in the second term. When letting $\alpha = 0$, GDBNSC-Rcut degenerates to NSC-Rcut.

3.1.2. The objective function of GDBNSC-Ncut

GDBNSC-Ncut solves the following problem:

$$\min_{\mathbf{Z},\mathbf{Y},\mathbf{Z}^T\mathbf{Z} = \mathbf{I}} \left\{\frac{1}{2}\|\boldsymbol{D}^{-1/2}\boldsymbol{X}^T - \boldsymbol{Z}\boldsymbol{Y}\|_F^2 + \alpha tr(\boldsymbol{Z}^T\boldsymbol{Q}\boldsymbol{Z})\right\} \tag{28}$$

where the trade-off parameter $\alpha \geq 0$ balances the error reconstruction regularization in the first term and the global discriminative regularization the second term item. Letting $\alpha = 0$, GDBNSC-Ncut degenerates to NSC-Ncut.

In traditional spectral clustering approaches, the eigenvectors of the graph Laplacian matrices may involve negative components. The existence of negative components may cause deviation of the results from actual cluster labels [47]. Fortunately, cluster indicator matrices $\boldsymbol{H}$ and $\boldsymbol{Z}$ are nonnegative in our NMF based spectral

clustering frameworks. Therefore, the results of our proposed algorithms are much closer to the real results. Additionally, the proposed algorithms are equipped with discriminating power by the utilization of the global discriminant information.

3.2. Iterative updating rules

Taken GDBNSC-Rcut as an example, we show the process of developing the iterative updating rules for the algorithm. Let us write the objective function:

$$L=\frac{1}{2}\|\boldsymbol{X}^T-\boldsymbol{HY}\|_F^2+\alpha tr(\boldsymbol{H}^T\boldsymbol{QH}) \tag{29}$$

The updating rule for $\boldsymbol{Y}$ is as follows:

$$\boldsymbol{Y}_{ij}{}^T=(\boldsymbol{XH})_{ij} \tag{30}$$

We have noticed that, GDBNSC-Rcut and NSSC-Rcut have the same updating rule for $\boldsymbol{Y}$. Next, we will concentrate on the updating rules for $\boldsymbol{H}$.

The objective function can be rewritten as follow:

$$\begin{aligned}L(\boldsymbol{H},\boldsymbol{Y})&=\frac{1}{2}\|\boldsymbol{X}^T-\boldsymbol{HY}\|_F^2+\alpha tr(\boldsymbol{H}^T\boldsymbol{QH})\\&=\frac{1}{2}tr(\boldsymbol{XX}^T-2\boldsymbol{Y}^T\boldsymbol{H}^T\boldsymbol{X}^T+\boldsymbol{Y}^T\boldsymbol{H}^T\boldsymbol{HY})+\alpha tr(\boldsymbol{H}^T\boldsymbol{QH})\end{aligned} \tag{31}$$

The partial derivative is

$$\frac{\partial L}{\partial \boldsymbol{H}^T}=-\boldsymbol{YX}+\boldsymbol{YY}^T\boldsymbol{H}^T+2\alpha\boldsymbol{H}^T\boldsymbol{Q} \tag{32}$$

Let $\boldsymbol{Q}=\boldsymbol{Q}^+-\boldsymbol{Q}^-$, $\boldsymbol{Q}_{ij}{}^+=(|\boldsymbol{Q}_{ij}|+\boldsymbol{Q}_{ij})/2$, $\boldsymbol{Q}_{ij}{}^-=(|\boldsymbol{Q}_{ij}|-\boldsymbol{Q}_{ij})/2$, with application of the gradient descent method, we have the following updating rule for $\boldsymbol{H}^T$:

$$\boldsymbol{H}_{ij}{}^T\leftarrow\boldsymbol{H}_{ij}{}^T+\eta_{ij}[\boldsymbol{YX}+2\alpha\boldsymbol{H}^T\boldsymbol{Q}^--(\boldsymbol{YY}^T\boldsymbol{H}^T+2\alpha\boldsymbol{H}^T\boldsymbol{Q}^+)] \tag{33}$$

We set the step η_{ij} similar with that in [24] as follows:

$$\eta_{ij}=\frac{\boldsymbol{H}_{ij}{}^T}{\boldsymbol{YY}^T\boldsymbol{H}^T+2\alpha\boldsymbol{H}^T\boldsymbol{Q}^+} \tag{34}$$

With η_{ij} substituted into the updating formula (33), we get

$$\boldsymbol{H}_{ij}{}^T=\boldsymbol{H}_{ij}{}^T\frac{\boldsymbol{YX}+2\alpha\boldsymbol{H}^T\boldsymbol{Q}^-}{\boldsymbol{YY}^T\boldsymbol{H}^T+2\alpha\boldsymbol{H}^T\boldsymbol{Q}^+} \tag{35}$$

For GDBNSC-Ncut algorithm, we have a similar iterative updating rule

$$\boldsymbol{Z}_{ij}^T=\boldsymbol{Z}_{ij}^T\frac{\boldsymbol{YXD}^{-1/2}+2\alpha\boldsymbol{Z}^T\boldsymbol{Q}^-}{\boldsymbol{YY}^T\boldsymbol{Z}^T+2\alpha\boldsymbol{Z}^T\boldsymbol{Q}^+} \tag{36}$$

The process of the proposed algorithm is shown as follows in Table 1.

In Table 1, as for handling the out-of-sample data, the cluster indicator matrix $\boldsymbol{Y}$ is K*M. When a new data point $\boldsymbol{x}$ (M*1) comes, its cluster indicator vector $\boldsymbol{Yx}$ is computed of size K*1. And its cluster membership can be obtained by k-means clustering.

Table 1
The global discriminative-based nonnegative spectral clustering algorithms.

Input: dataset $\boldsymbol{X}$, kernel matrix K_1, the number of clusters K, the maximum number of iterations t, and regularization parameter α.
Output: clustering labels

1. Random initialize of the two nonnegative matrixes $\boldsymbol{H}$ ($\boldsymbol{Z}$ for GDBNSC-Ncut) and $\boldsymbol{Y}$.
2. Updating $\boldsymbol{Y}$ and $\boldsymbol{H}$ ($\boldsymbol{Z}$ for GDBNSC-Ncut) iteratively using corresponding rules.
3. After reaching the number of iterations, output $\boldsymbol{Y}$ and $\boldsymbol{H}$.
4. Clustering $\boldsymbol{H}$ ($\boldsymbol{Z}$ for GDBNSC-Ncut) into k classes using *k-means* algorithm.
5. Once a new data comes, its cluster indicator vector can be calculated by $\boldsymbol{Yx}$, and its cluster membership can be calculated by *k-means* clustering algorithm.

3.3. Convergence analysis of the proposed algorithms

In this section, we analyze the convergence of the iterative updating rules in case of GDBNSC-Rcut. We will prove that the iterative updating schemes state in Eqs. (30) and (35) lead to local minima of the objective function in Eq. (29).

We have the following theorem:

Theorem 3. *For given matrices $\boldsymbol{X}\in\mathbb{R}^{M\times N}$, $\boldsymbol{H}\in\mathbb{R}^{N\times K}$, $\boldsymbol{Y}\in\mathbb{R}^{K\times M}\geq 0$, the objective function in formula (29) is non-increasing under the alternative iterative updating rules in (30) and (35).*

Next, we give a detailed proof of the theorem. Our proof follows the ideas in the proof of NMF [24] and [48].

Lemma 1. *When the following conditions are satisfied:*

$$G(u,u')\geq F(u) \tag{37}$$

And

$$G(u,u)=F(u) \tag{38}$$

$G(u,u')$ is an auxiliary function of $F(u)$. So under the updating formula

$$u^{(t+1)}=\arg\min_u G(u,u^{(t)}) \tag{39}$$

the function F is non-increasing.

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

Note that only when $u^{(t)}$ is the local minimum for $F(u,u^{(t)})$, $\boldsymbol{H}(u^{(t+1)})=\boldsymbol{H}(u^{(t)})$ holds [24].

Since the updating rule for $\boldsymbol{H}$ contains global manifold discrimination information, here we demonstrate a proof of convergence of only the updating rule of $\boldsymbol{H}$, the convergence proof of the updating rule of $\boldsymbol{Y}$ is a similar case.

For notation convenience, we let $\boldsymbol{V}=\boldsymbol{H}^T$.

Lemma 2. *Let F' be the first partial derivatives of L with respect to $\boldsymbol{H}^T(\boldsymbol{V})$, the function*

$$G(v,v_{ab}^t)=F_{ab}(v_{ab}^t)+F'_{ab}(v-v_{ab}^t)+\frac{(\boldsymbol{YY}^T\boldsymbol{H}^T+2\alpha\boldsymbol{H}^T\boldsymbol{Q}^++\lambda)_{ab}}{v_{ab}^t}(v-v_{ab}^t)^2 \tag{40}$$

is an auxiliary function of F_{ab}.

Proof By. the above equation, we have $G(v,v)=F_{ab}(v)$, so we only need to prove that $G(v,v_{ab}^t)\geq F_{ab}(v)$. To this end, we compare the Taylor expansion of $G(v,v_{ab}^t)$ with $F_{ab}(v)$

$$F_{ab}^t(v)=F_{ab}(v_{ab}^t)+F'_{ab}(v-v_{ab}^t)+\frac{1}{2}F''_{ab}(v-v_{ab}^t)^2 \tag{41}$$

where, F_{ab}'' is the second order partial derivative of L with respect to $\boldsymbol{H}^T$

$$F''_{ab}=\boldsymbol{YY}^T+2\alpha\boldsymbol{Q} \tag{42}$$

$$(\boldsymbol{YY}^T\boldsymbol{H}^T)_{ab}=\sum_{l=1}^{K}(\boldsymbol{YY}^T)_{al}v_{lb}^t\geq(\boldsymbol{YY}^T)_{aa}v_{ab}^t \tag{43}$$

$$(\boldsymbol{H}^T\boldsymbol{Q}^+)_{ab}=\sum_{l=1}^{N}v_{al}^t\boldsymbol{Q}_{lb}^+\geq v_{ab}^t\boldsymbol{Q}_{bb}^+\geq v_{ab}^t(\boldsymbol{Q}^+-\boldsymbol{Q}^-)_{bb} \tag{44}$$

In summary, we have the following inequality

$$\frac{(\boldsymbol{YY}^T\boldsymbol{H}^T+2\alpha\boldsymbol{H}^T\boldsymbol{Q}^++\lambda)_{ab}}{v_{ab}^t}\geq\frac{1}{2}F''_{ab} \tag{45}$$

Then the inequality $G(v,v_{ab}^t)\geq F_{ab}(v)$ is satisfied, and the lemma is proven.

From Lemma 2, we know that $G(v, v_{ab}^t)$ is an auxiliary function of $F_{ab}(v_{ab})$. Then from Lemma 1, we solve the problem

$$v^{(t+1)} = \arg\min_u F(v, v^{(t)}) \tag{46}$$

We get

$$v_{ab}^{(t+1)} = v_{ab}^{(t)} \frac{(\boldsymbol{YX} + 2\alpha \boldsymbol{VQ}^-)_{ab}}{(\boldsymbol{YY}^T\boldsymbol{V} + 2\alpha \boldsymbol{VQ}^+)_{ab}} \tag{47}$$

So the updating rule for $\mathbf{H}^T$ is as follows

$$\boldsymbol{H}_{ij}^T = \boldsymbol{H}_{ij}^T \frac{\boldsymbol{YX} + 2\alpha \boldsymbol{H}^T\boldsymbol{Q}^-}{\boldsymbol{YY}^T\boldsymbol{H}^T + 2\alpha \boldsymbol{H}^T\boldsymbol{Q}^+} \tag{48}$$

Similarly, we can get the updating rule for $\boldsymbol{Y}^T$:

$$\boldsymbol{Y}_{ij}{}^T = (\boldsymbol{XH})_{ij} \tag{49}$$

Table 2
Features of UCI and AT&T datasets.

Datasets	#Sample	#Dimension	#Class
Dermatology	366	33	6
Glass	214	9	6
Soybean	47	35	4
Vehicle	846	18	4
Zoo	101	16	7
AT&T	400	10,304	40

4. Experimental results and analysis

In this section, we carry out extensive experiments on some datasets in comparison with some other spectral clustering methods.

4.1. Data sets

We will make comparisons of the clustering quality in term of clustering accuracy with some spectral clustering algorithms. The compared algorithms include traditional Ncut [18] and Rcut [19] spectral clustering algorithms, NLE [29], NSC-Ncut, NSC-Rcut, NSSC-Ncut and NSSC-Rcut [30]. Six datasets include five UCI datasets and an AT&T face dataset [30].

The datasets we use are the same as in [30] and the features are shown as follows in Table 2.

4.2. Evaluation metric

In the experiments, the number of cluster is set as the true number of classes. We use a common clustering evaluation index, namely clustering accuracy (ACC) [49–51] to evaluate the effectiveness of the above clustering algorithms.

Given a data point $\boldsymbol{x}_i$, $\boldsymbol{c}_i$ and $\boldsymbol{g}_i$ are the labels of the clustering algorithm and the ground truth label respectively. The accuracy (ACC) of a clustering method is defined as

$$ACC = \frac{\sum_{i=1}^{n} \delta(g_i, map(c_i))}{n} \tag{50}$$

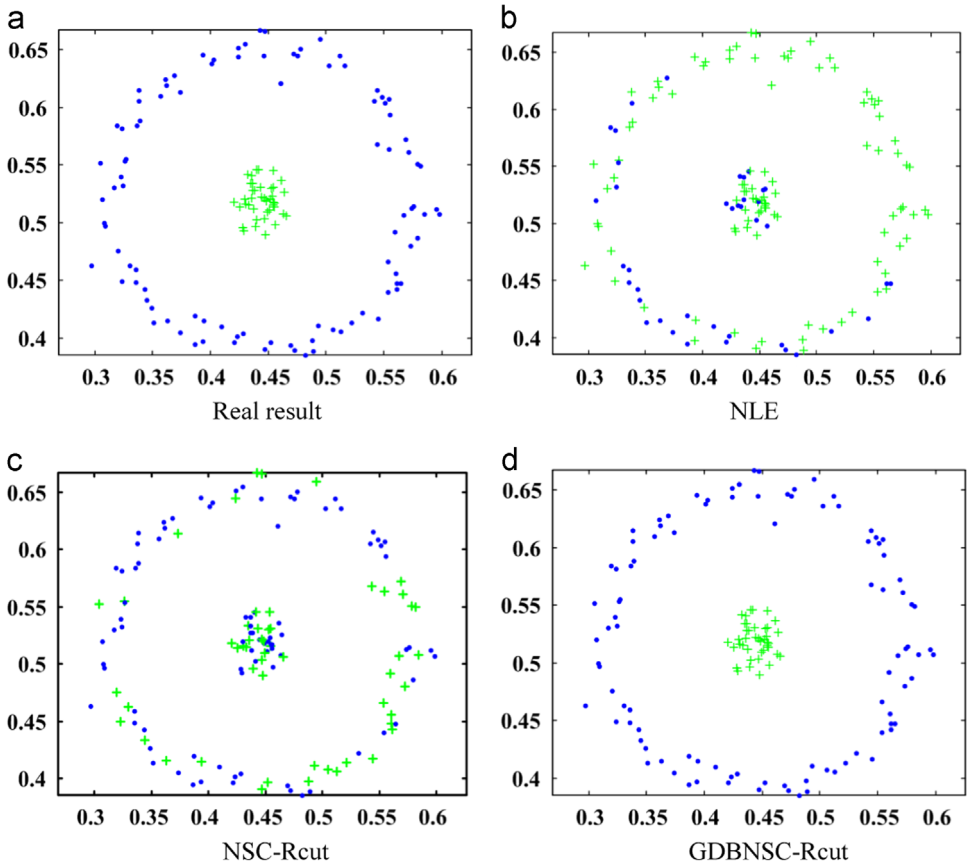

Fig. 1. Real clustering result and clustering results of NLE, NSC-Rcut and GDBNSC-Rcut on a two-dimensional synthetic dataset. (a) Real result (b) NLE (c) NSC-Rcut (d) GDBNSC-Rcut.

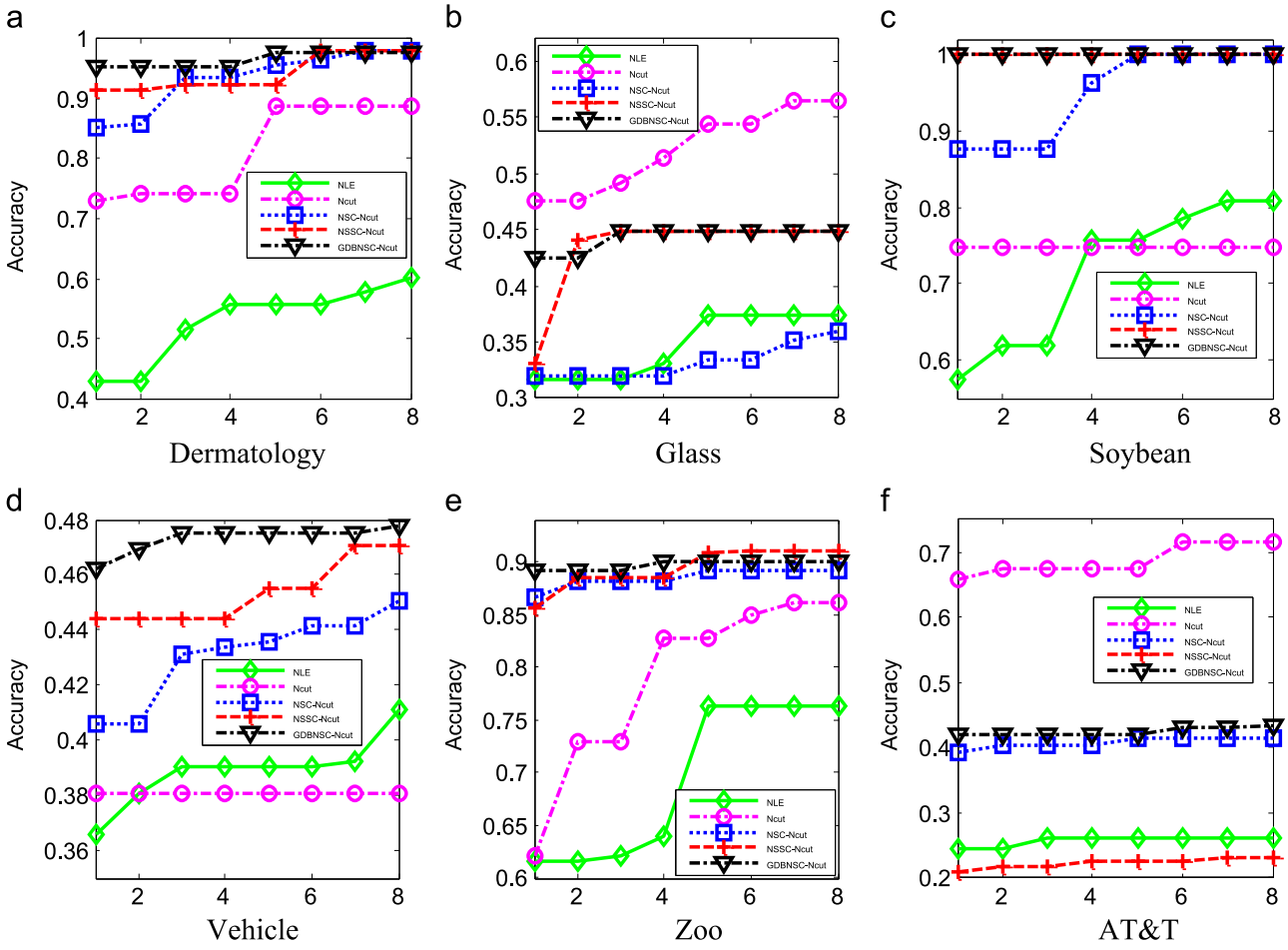

Fig. 2. The clustering accuracy of NLE, Ncut, NSC-Ncut, NSSC-Ncut and GDBNSC-Ncut on five UCI datasets and AT&T dataset. (a) Dermatology, (b) Glass, (c) Soybean, (d) Vehicle, (e) Zoo and (f) AT&T. (a) Dermatology (b) Glass (c) Soybean (d) Vehicle (e) Zoo (f) AT&T.

where, n is the total number of data samples, $\delta(.,.)$ is the *delta* function as follows:

$$\delta(x,y)=\begin{cases}1, & x=y\\ 0, & otherwise\end{cases} \tag{51}$$

map(.) is the optimal mapping function, using the *Hungarian* algorithm [52] to permute the clustering labels and the ground truth labels. The higher the accuracy is, the better the performance is.

4.3. Toy experiment

We apply NLE, NSC-Rcut and GDBNSC-Rcut to a two-dimensional synthetic dataset, shown in Fig. 1(a). The clustering results of NLE, NSC-Rcut and GDBNSC-Rcut on this dataset are given in Fig. 1(b), (c), and (d) respectively. In Fig. 1, different shapes indicate different classes.

From Fig. 1(b), (c), and (d), we can see that NLE and NSC-Rcut are much inferior to the proposed GDBNSC-Rcut. It can be seen from Fig. 1(d) that GDBNSC-Rcut can differentiate the two classes successfully, which shows the effectiveness of GDBNSC-Rcut. This is due to that the proposed GDBNSC-Rcut uses the kernelized global discriminative model which makes it effective for nonlinear case.

4.4. Experimental setup

With the same experimental setting in [30], for each clustering algorithm, we independently run 256 times. For each run, we first randomly initialize matrices, and then iterate 300 times to achieve convergence and obtain the cluster indicator matrix $\boldsymbol{H}$ (or $\boldsymbol{Z}$). k-means algorithm is applied to the rows of $\boldsymbol{H}$ (or $\boldsymbol{Z}$, $\boldsymbol{Q}$) to obtain the clustering results finally. We fix the parameter μ in matrix $\boldsymbol{Q}=\boldsymbol{H}_N-\boldsymbol{H}_N^T(\boldsymbol{H}_N+\mu\boldsymbol{K}_1^{-1})^{-1}\boldsymbol{H}_N$ as $\boldsymbol{\mu}=10^{-12}$. We tune the regularization parameter α from $\{10^{-7},10^{-5},10^{-4},10^{-3},10^{-2},10^{-1},1,10^{1},10^{2},10^{3}\}$. And for the sake of simplify, we fix the parameter of the Gaussian kernel $\sigma=1$. For each dataset, we choose the value of α so that the algorithm has the highest accuracy.

4.5. Experimental results and analysis

For each clustering algorithm, we iterate 300 times, run 256 times independently and continuously. We choose the best results among the first 2^N, N=1, 2, …, 8 runs to plot figures. The horizontal axis is the logarithm of the number of runs, and the vertical axis is the clustering accuracy of the algorithm. The accuracy of NLE, Ncut, NSC-Ncut, NSSC-Ncut and GDBNSC-Ncut algorithms on six data sets is shown in Fig. 2.

From Fig. 2, for Ncut group spectral clustering algorithms, the proposed GDBNSC-Ncut can achieve good results within only a few runs. The proposed GDBNSC-Ncut always outperforms NLE, NSC-Ncut and NSSC-Ncut. However, NLE performs the worst. On "Glass" and "AT&T" datasets, Ncut achieves the best results. On other four datasets, GDBNSC-Ncut gets the best results. It is worth mentioning that, of clustering the 400 × 10304 "AT&T" dataset into 40 classes, GDBNSC-Ncut achieves an accuracy of about 0.4, while

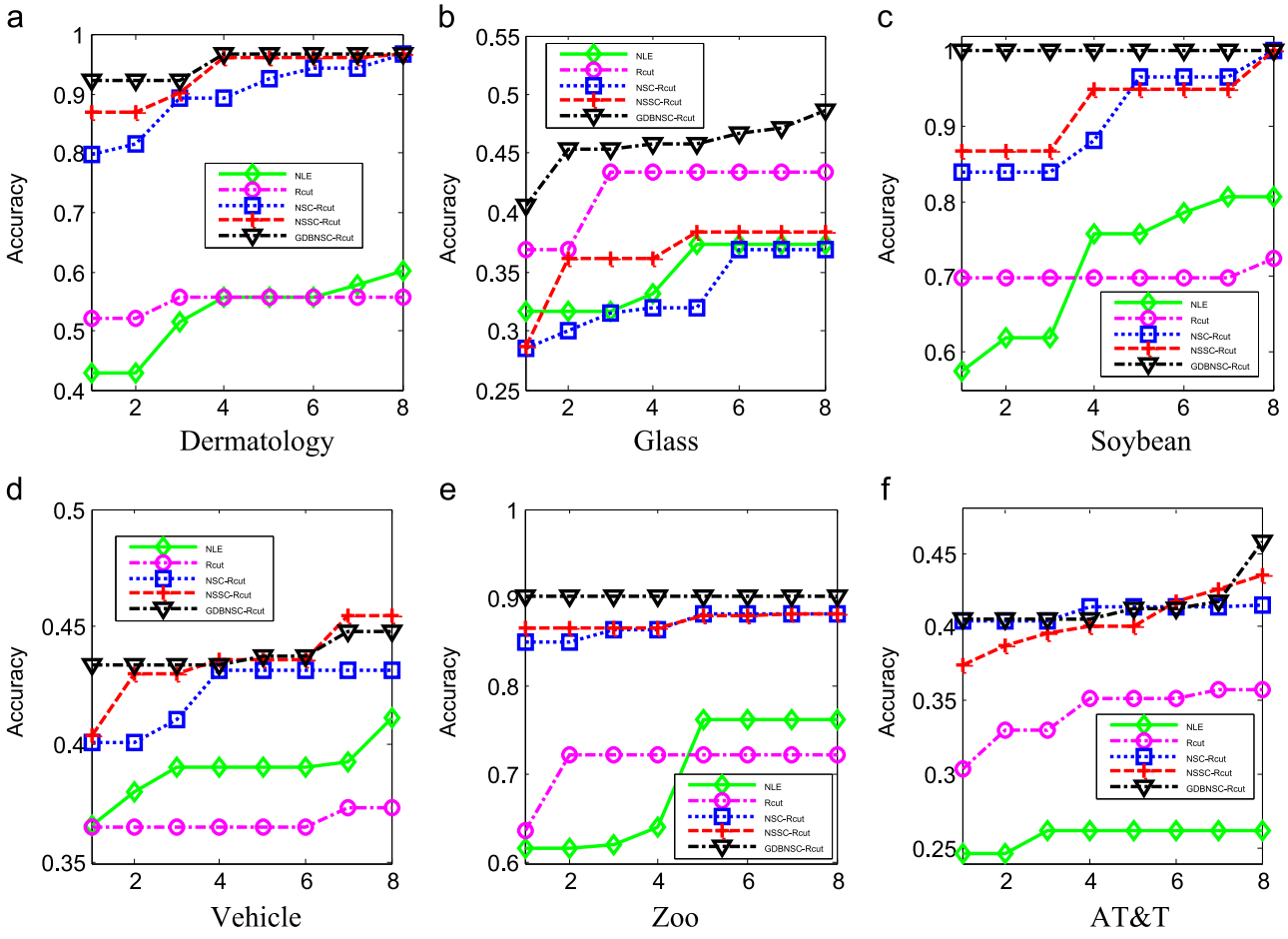

Fig. 3. The clustering accuracy of NLE, Rcut, NSC-Rcut, NSSC-Rcut and GDBNSC-Rcut on five UCI datasets and AT&T dataset. (a) Dermatology, (b) Glass, (c) Soybean, (d) Vehicle, (e) Zoo and (f) AT&T. (a) Dermatology (b) Glass (c) Soybean (d)Vehicle (e) Zoo (f) AT&T.

Table 3
The clustering results of NLE and Ncut group spectral clustering algorithms on 6 datasets.

Mthods		Dermatology	Glass	Soybean	Zoo	Vehicle	AT&T
NLE	Ave	0.3489	0.2505	0.4771	0.4932	0.2824	0.2072
	Best	0.6011	0.3738	0.8085	0.7624	0.4113	0.2625
Ncut	Ave	0.7547	**0.4653**	0.7046	0.6317	0.3776	0.6224
	Best	0.888	**0.5654**	0.7447	0.8614	0.3806	0.715
NSC-Ncut	Ave	0.7129	0.2522	0.7143	0.6159	0.3957	0.3518
	Best	**0.9809**	0.3598	**1**	0.8911	0.4504	0.415
NSSC-Ncut	Ave	0.7185	0.3477	0.7167	**0.661**	0.4196	0.025
	Best	**0.9809**	0.4486	**1**	**0.9109**	0.4704	0.2325
GDBNSC-Ncut	Ave	**0.8260**	0.4126	**0.7977**	0.6504	**0.4605**	0.3854
	Best	0.9781	0.4486	**1**	0.8812	**0.4775**	0.435

the result of NSSC-Ncut is about 0.22, and the improvement of GDBNSC-Ncut is obvious.

The accuracy results of NLE, Rcut, NSC-Rcut, NSSC-Rcut, GDBNSC-Rcut algorithms on six datasets are shown in Fig. 3.

From Fig. 3, for Rcut group spectral clustering algorithms, the proposed GDBNSC-Rcut can achieve good results within only a few runs. The proposed GDBNSC-Rcut algorithm outperforms all the other algorithms on the six datasets. Both NLE and Rcut are much inferior to the other methods. NSC-Rcut and NSSC-Rcut perform the second best on all datasets except for the "Glass" dataset. On the "Glass" dataset, NSC-Rcut has clustering result of about 0.32, a relatively poor result. The GDBNSC-Rcut achieves an accuracy of about 0.42 on the 'Glass' dataset, an apparent improvement.

For each clustering algorithm, we independently run 256 times, we record the best results from the optimal parameter and the average results are reported. All the results are shown in Tables 3 and 4. The best results for each dataset are highlighted in bold.

From the data in Tables 3 and 4, we can get the following conclusions:

1. Ncut group spectral clustering algorithms usually perform better than Rcut group spectral clustering algorithms. NLE performs the worst among all the nine algorithms.
2. GDBNSC-Ncut outperforms or has competitive clustering results compared with NSC-Ncut and NSSC-Ncut. GDBNSC-Ncut outperforms Ncut except for "Glass" and "AT&T" datasets. GDBNSC-Rcut outperforms or has competitive clustering results compared with NSC-Rcut and NSSC-Rcut. The results of GDBNSC-Rcut are much better than that of Rcut overall. This demonstrates that it is crucial to exploit the discriminative information in unsupervised clustering.
3. Although the results of GDBNSC-Ncut on "Glass" and "AT&T" datasets are worse than that of Ncut, On "Glass" and "AT&T" datasets, the results of GDBNSC-Ncut are much better than those of NSC-Ncut and NSSC-Ncut on "Glass" and "AT&T" datasets.
4. Almost all the best results are acquired by GDBNSC-Ncut and GDBNSC-Rcut which illustrates the effectiveness of the proposed algorithms.
5. The results have shown that the clustering performance can be significantly enhanced with the global geometrical structure exploited and the global discriminative structure considered.

Table 4
The clustering results of NLE and Rcut group spectral clustering algorithms on 6 datasets.

Methods		Dermatology	Glass	Soybean	Zoo	Vehicle	AT&T
NLE	Ave	0.3489	0.2505	0.4771	0.4932	0.2824	0.2072
	Best	0.6011	0.3738	0.8085	0.7624	0.4113	0.2625
Ncut	Ave	0.4725	**0.4155**	0.6397	0.6096	0.3391	0.3136
	Best	0.5574	0.4346	0.7234	0.7228	0.3735	0.3575
NSC-Rcut	Ave	0.6602	0.2516	0.6986	0.6197	0.3863	0.3533
	Best	0.9672	0.3692	**1**	0.8812	0.4314	0.415
NSSC-Rcut	Ave	0.67	0.2616	0.6976	0.6185	0.3865	0.3513
	Best	**0.9699**	0.3832	**1**	0.8812	**0.4551**	0.435
GDBNSC-Rcut	Ave	**0.7324**	0.3640	**0.8004**	**0.6477**	**0.3886**	**0.3695**
	Best	**0.9699**	**0.486**	**1**	**0.901**	0.448	**0.4575**

Fig. 4. Thirty images of three persons from AT&T face database.

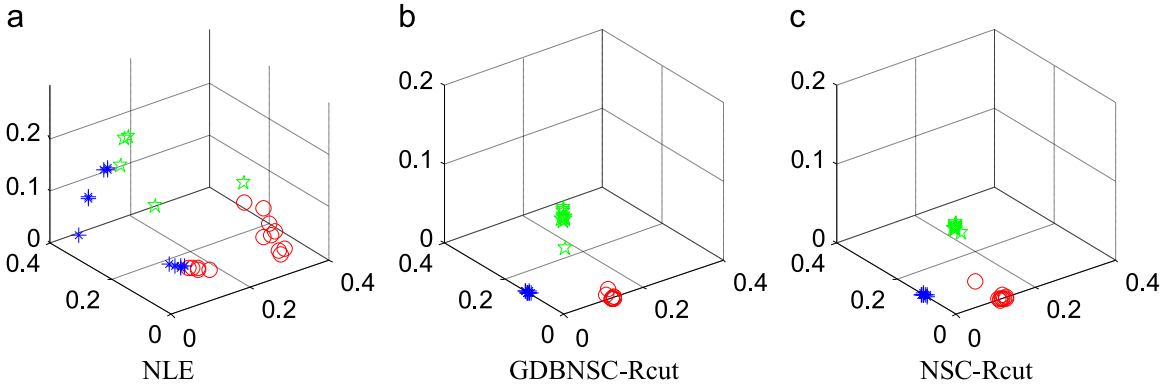

Fig. 5. Clustering effects of (a) NLE, (b) GDBNSC-Rcut and (c) NSC-Rcut. (a) NLE (b) GDBNSC-Rcut (c) NSC-Rcut.

4.6. An illustrative example

Fig. 4 shows the images of three persons from AT&T database, each person has 10 images with different expressions and poses. For each image, we reshape the image to a single vector to represent the image, and then the dataset for clustering is of the size $10{,}304 \times 30$. Experiments are also conducted on NSC-Rcut, NLE and GDBNSC-Rcut as comparison to illustrate the effectiveness of the proposed algorithms.

As the cluster indicator matrix $\boldsymbol{H}$ of GDBNSC-Rcut is 30×3, we can plot each row of $\boldsymbol{H}$ after 50 iterations as a point in a 3D plot shown in Fig. 5(b), where different clusters are represented by different colors and signs. Fig. 5(a) shows the image of $\boldsymbol{Q}$ matrix in NLE, and Fig. 5(c) shows the image of $\boldsymbol{H}$ matrix in NSC-Rcut [30].

From Fig. 5, we have the observation that the three clusters cannot be separated using NLE method (see Fig. 5(a)), and that the proposed method GDBNSC-Rcut and NSC-Rcut can separate the three clusters even in an iteration of 30. This experiment illustrates that the proposed algorithm GDBNSC-Rcut has better clustering discriminative ability than NLE.

4.7. Sensitivity to the selection of the parameter α

In this section, we will investigate the sensitivity with respect to the regularization parameter α on the 6 datasets. The iteration for the algorithms is 300. Fig. 6 shows how the average performance with 20 independent and continuous runs of GDBNSC-Ncut and GDBNSC-Rcut vary with α.

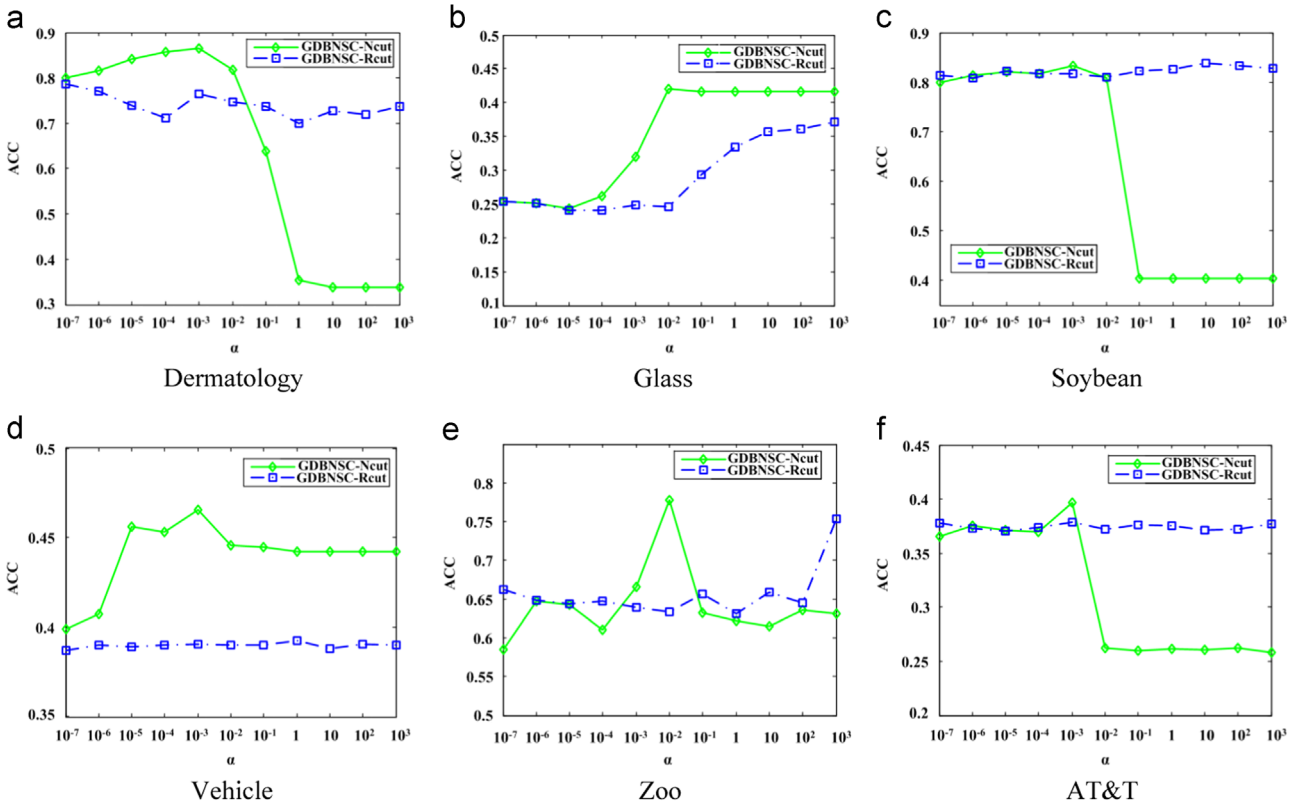

Fig. 6. The performances of GDBNSC-Ncut and GDBNSC-Rcut vary with the regularization parameterα (a) Dermatology (b) Glass (c) Soybean (d) Vehicle (e) Zoo (f) AT&T.

From the results shown in Fig. 6, we can observe that when the value of the regularization parameter α increases, the clustering results of GDBNSC-Ncut will first increase and then decrease but remain stable finally. The proposed GDBNSC-Rcut is robust to the value of the regularization parameter α.

5. Conclusions

In this paper, we proposed two novel global discriminative-based nonnegative spectral algorithms for clustering, named GDBNSC-Ncut and DBNSC-Rcut. Based on the equivalence between spectral clustering and NMF, we incorporate the global discriminative regularization terms into a nonnegative matrix factorization framework. The proposed algorithms preserve both the global geometrical and global discriminative structure of datasets. They have strong discriminative power and good clustering results. Experiments on real world datasets demonstrate the effectiveness of the proposed algorithms. However, not all algorithms including the proposed algorithms can achieve good enough results on some datasets such as “AT&T”, where the accuracy is below fifty percent. Therefore, the future work is to integrate feature selection into the framework for clustering, especially for high dimensional dataset such as “AT&T” to further improve the clustering accuracy.

Conflict of interest

We declare that we have no conflict of interest.

Acknowledgment

We would like to express our sincere appreciation to the editors and the anonymous reviewers for their valuable comments, which have greatly helped us in improving the quality of the paper. This work was partially supported by the National Basic Research Program (973 Program) of China under Grant 2013CB329402, the National Natural Science Foundation of China, under Grants 61371201, the Program for Cheung Kong Scholars and Innovative Research Team in University under Grant IRT1170.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/S0031-3203(16)00056-X.

References

[1] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. ((CSUR)) 31 (1999) 264–323.
[2] F.H. Shang, L.C. Jiao, F. Wang, Graph dual regularization non-negative matrix factorization for co-clustering, Pattern Recognit. 45 (2012) 2237–2250.
[3] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data [M], Prentice-Hall, Englewood Cliffs, NJ, 1988.
[4] F. Tung, A. Wong, D.A. Clausi, Enabling scalable spectral clustering for image segmentation, Pattern Recognit. 43 (2010) 4069–4076.
[5] D. Jiang, C. Tang, A. Zhang, Cluster analysis for gene expression data: a survey, IEEE Trans. Knowl. Data Eng. 16 (2004) 1370–1386.
[6] K.M. Hammouda, M.S. Kamel, Efficient phrase-based document indexing for web document clustering, IEEE Trans. Knowl. Data Eng. 16 (2004) 1279–1296.
[7] S. Gordon, H. Greenspan, J. Goldberger, Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. ((CVPR)) (2003) 370–377.

[8] X.W. Kong, R. Wang, G. Li, Fuzzy clustering algorithms based on resolution and their application in image compression, Pattern Recognit. (2002) 2439–2444.
[9] M. Filippone, F. Camastra, F. Masulli, S. Rovetta, A survey of kernel and spectral methods for clustering, Pattern Recognit. 41 (2008) 176–190.
[10] U.V. Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (2007) 395–416.
[11] A. Ng, M. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, Adv. Neural Inform. Process. Syst. ((NIPS)) (2001) 849–856.
[12] F. Nie, X. Wang, H. Huang, Clustering and projected clustering with adaptive neighbors, in: Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2014, pp. 977–986.
[13] F. Nie, D. Xu, I.W. Tsang, C. Zhang, Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction, IEEE Trans. Image Process. 19 (7) (2010) 1921–1932.
[14] F. Nie, C. Ding, D. Luo, H. Huang, Improved minmax cut graph clustering with nonnegative relaxation, in: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2010, pp. 451–466.
[15] S.F. Ding, H.J. Jia, L.W. Zhang, F.X. Jin, Research of semisupervised spectral clustering algorithm based on pairwise constraints, Neural Comput. Appl. 24 (1) (2014) 211–219.
[16] H.J. Jia, S.F. Ding, H. Ma, W.Q. Xing, Spectral clustering with neighborhood attribute reduction based on information entropy, J. Comput. 9 (6) (2014) 1316–1324.
[17] H.J. Jia, S.F. Ding, H. Zhu, F.L. Wu, L.N. Bao, A feature weighted spectral clustering algorithm based on knowledge entropy, J. Softw. 8 (5) (2013) 1101–1108.
[18] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 888–905.
[19] P.K. Chan, M. Schlag, J.Y. Zien, Spectral k-way ratio cut partitioning and clustering, IEEE Trans. CAD-Integrated Circuits Syst. 13 (1994) 1088–1096.
[20] E.R. Barnes, An algorithm for partitioning the nodes of a graph, SIAM J. Algebraic Discret. Methods 3 (4) (1982) 541–550.
[21] C.H.Q. Ding, X. He, H. Zha, A min-max cut algorithm for graph partitioning and data clustering ICDM', Proc. IEEE Int. Conf. Data Min. (2001) 107–114.
[22] H.J. Jia, S.F. Ding, X.Z. Xu, R. Nie, The latest research progress on spectral clustering, Neural Comput. Appl. 24 (7 – 8) (2014) 1477–1486.
[23] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.
[24] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, Adv. Neural Inform. Process. Syst. ((NIPS)) (2001) 556–562.
[25] I. Biederman, Recognition-by-components: a theory of human image understanding, Psychol. Rev. 94 (1987) 115–147.
[26] D. Ross, R.S. Zemel, Learning parts-based representation of data, J. Mach. Learn. Res. ((JMLR)) 7 (2006) 2369–2397.
[27] C. Ding, T. Li, W. Peng, Nonnegative matrix factorization and probabilistic latent semantic indexing: equivalence, chi-square statistic, and a hybrid method (AAAI-06), Proc. Natl. Conf. Artif. Intell. (2006).
[28] C. Ding, X. He, H.D. Simon, On the equivalence of nonnegative matrix factorization and spectral clustering, Proc. SIAM Data Min. Conf. (2005).
[29] D. Luo, C. Ding, H. Huang, T. Li, Non-negative Laplacian embedding, Int. Conf. Data Min. ((ICDM)) (2009) 337–346.
[30] H. Lu, Z. Fu, X. Shu, Non-negative and sparse spectral clustering, Pattern Recognit. 47 (2014) 418–426.
[31] F. De la Torre, T. Kanade, Discriminative cluster analysis, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 241–248.
[32] J. Ye, Z. Zhao, M. Wu, Discriminative k-means for clustering, Adv. Neural Inform. Process. Syst. (2007) 1649–1656.
[33] P. Li, J. Bu, Y. Yang, R. Ji, C. Chen, D. Cai, Discriminative orthogonal nonnegative matrix factorization with flexibility for data representation, Expert Syst. Appl. 41 (2014) 1283–1293.
[34] Y. Yang, D. Xu, F. Nie, S. Yan, Y. Zhuang, Image clustering using local discriminative models and global integration, IEEE Trans. Image Process. 19 (2010) 2761–2773.
[35] L. Du, Z. Shen, X. Li, P. Zhou, Y. D. Shen, Local and global discriminative learning for unsupervised feature selection, in: Proceedings of 13th IEEE International Conference on Data Mining (ICDM) 2013, pp. 131–139.
[36] F. Nie, Z. Zeng, I.W. Tsang, D. Xu, C. Zhang, Spectral embedded clustering: a framework for in-sample and out-of-sample spectral clustering, IEEE Trans. Neural Netw. 22 (2011) 1796–1808.
[37] Y. Yang, H. Shen, Y. Zhang, X. Du, Discriminative nonnegative spectral clustering with out-of-sample extension, IEEE Trans. Knowl. Data Eng. 26 (2013) 1760–1770.
[38] F. Nie, S. Xiang, Y. Liu, C. Hou, C. Zhang, Orthogonal vs. uncorrelated least squares discriminant analysis for feature extraction, Pattern Recognit. Lett. 33 (5) (2012) 485–491.
[39] B. Schölkopf, A. Smola, K.R. Müller, Kernel principal component analysis, Artif. Neural Netw. 1327 (1997) 583–588.
[40] K.R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, An introduction to kernel-based learning algorithms, IEEE Trans. Neural Netw. 12 (02) (2001) 181–201.
[41] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.R. Müller, Fisher discriminant analysis with kernels, Neural Networks for Signal Processing IX, in: Proceedings of the IEEE Signal Processing Society Workshop, 1999, pp. 41–48.
[42] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Comput. 12 (10) (2000) 2385–2404.
[43] Z. Liang, P. Shi, An efficient and effective method to solve kernel Fisher discriminant analysis, Neurocomputing 61 (2004) 485–493.
[44] J. Yang, Z. Jin, J. Yang, et al., Essence of kernel Fisher discriminant: KPCA plus LDA, Pattern Recognit. 37 (10) (2004) 2097–2100.
[45] Y. Yang, H. Shen, F. Nie, R. Ji, X. Zhou, Nonnegative spectral clustering with discriminative regularization, in: Proceedings of the 25th AAAI Conference on Artifical Intelligence, 2011, pp. 555–560.
[46] J. Wang, H. Bensmail, X. Gao, Feature selection and multi-kernel learning for sparse representation on a manifold, Neural Netw. 51 (2014) 9–16.
[47] Z. Li, Y. Yang, J. Liu, X. Zhou, H. Lu, Unsupervised feature selection using nonnegative spectral analysis, in: Proceedings of the 26th AAAI Conference on Artifical Intelligence, 2012, pp. 1026–1032.
[48] C.J. Lin, On the convergence of multiplicative updating algorithms for non-negative matrix factorization, IEEE Trans. Neural Netw. 18 (2007) 1589–1596.
[49] D. Cai, X.F. He, J.W. Han, T.S. Huang, Graph regularized nonnegative matrix factorization for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 1548–1560.
[50] D. Cai, X. He, X. Wu, J. Han, Non-negative matrix factorization on manifold, in: Proceedings of the 8th IEEE International Conference on Data Mining, 2008, pp. 63–72.
[51] H. Liu, Z. Wu, X. Li, D. Cai, T.S. Huang, Constrained nonnegative matrix factorization for imagine representation, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 1299–1311.
[52] C.H. Papadimitriou, K. Steiglitz, Combinatorial Optimization: Algorithms and Complexity, Dover, New York, 1998.

Ronghua Shang is currently an associate Professor with Xidian University, Xi'an, China. She received the B.C. degree in Information and Computing Science, and the Ph.D. degree from Xidian University, in 2003 and 2008, respectively. Since 2008, she has been a lecturer of Xidian University. She was promoted to associate professor in 2010. Her research interests are broadly in the area of computational intelligence, with applications to optimization, learning, data mining and image understanding. She has published over thirty papers in journals and conferences, and holds three granted patents as the first inventor. She is leading or has completed nine projects as the PI, funded by the National Natural Science Foundation of China and others.