Nonnegative Spectral Clustering with Discriminative Regularization

Yi Yang[1], Heng Tao Shen[1], Feiping Nie[2], Rongrong Ji[3], Xiaofang Zhou[1]

[1]School of Information Technology & Electrical Engineering, The University of Queensland.
[2] Department of Computer Science and Engineering, University of Texas at Arlington.
[3]Department of Electronic Engineering, Columbia University.
yangyi_zju@yahoo.com.cn, shenht@itee.uq.edu.au, feipingnie@gmail.com, rj2349@columbia.edu, zxf@itee.uq.edu.au.

Abstract

Clustering is a fundamental research topic in the field of data mining. Optimizing the objective functions of clustering algorithms, e.g. normalized cut and k-means, is an NP-hard optimization problem. Existing algorithms usually relax the elements of cluster indicator matrix from discrete values to continuous ones. Eigenvalue decomposition is then performed to obtain a relaxed continuous solution, which must be discretized. The main problem is that the signs of the relaxed continuous solution are mixed. Such results may deviate severely from the true solution, making it a nontrivial task to get the cluster labels. To address the problem, we impose an explicit nonnegative constraint for a more accurate solution during the relaxation. Besides, we additionally introduce a discriminative regularization into the objective to avoid overfitting. A new iterative approach is proposed to optimize the objective. We show that the algorithm is a general one which naturally leads to other extensions. Experiments demonstrate the effectiveness of our algorithm.

Introduction

Clustering plays an important role in many areas, such as pattern recognition and data indexing, navigation, organization and summarization. Recent research efforts have shown that spectral clustering and its variants are usually more capable to partition the data sampled from complicated structures into different clusters, mainly due to the utilization of data local structures (Shi and Malik 2000; Yu and Shi 2003; Wu and Schölkopf 2006; Nie et al. 2009; Yang et al. 2010b). However, there are two main limitations. Firstly, it is NP-hard to optimize the objective functions of spectral clustering algorithms. The traditional solution is to relax the elements of the cluster indicator matrix from discrete values to continuous ones. In that way, eigenvalue decomposition can be employed to compute the continuous valued cluster indicator matrix. Because the cluster indicator matrix obtained by eigenvalue decomposition is mixed signed and usually not sparse, it might severely deviate from the true solution (Ding, Li, and Jordan 2008). Additionally, given the mixed signed cluster indicator matrix, there is no straightforward method to discretize it. Frequently used algorithms to obtain the discretized cluster indicator matrix in previous works include EM-like algorithm, i.e. standard k-means clustering, and spectral rotation (Yu and Shi 2003). Secondly, most spectral clustering algorithms only focus on local structures of the input data distribution. It may, under certain circumstances, incur overfitting and therefore degrades the clustering performance.

In this paper, we propose a new spectral clustering algorithm, namely Nonnegative Spectral clustering with Discriminative Regularization (NSDR). Different from most of the previous spectral clustering algorithms (Shi and Malik 2000; Yu and Shi 2003; Wu and Schölkopf 2006; Yang et al. 2010b), nonnegative constraint is imposed to the cluster indicator matrix when it is relaxed to the continuous domain (Nie et al. 2010a). Such a constraint insures that the solution is much closer to the ideal cluster indicator matrix, and it can be readily used to assign cluster labels to each input datum. In this way, *no discretization is required*. Although nonnegative constraints have been previously imposed in matrix factorization for data representation (Cai et al. 2009; Lee and Seung 1999; Liu and Wu 2010), matrix factorization is not involved in our algorithm. The motivation of imposing nonnegative constraint is to make the relaxed cluster indicator matrix more accurate, making it intrinsically different from (Cai et al. 2009; Lee and Seung 1999; Liu and Wu 2010; Liu et al. 2010). We not only focus on local structures of the data distribution, but also incorporate the global discriminative information to avoid overfitting and make the results more robust. Experiments on different datasets show that our algorithm outperforms the state-of-the-art clustering algorithms.

The rest of this paper is organized as follows. After brief review of spectral clustering in section 2, we detail our NSDR algorithm in section 3. After that, we show that our algorithm leads to several extensions. Experiment are given in section 5 and section 6 concludes this paper.

Spectral Clustering

Before getting started, we first summarize the notations which will be frequently used in this paper. Denote $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ as the input data set to be clustered, where $x_i \in \mathbb{R}^d (1 \leq i \leq n)$ is the i-th datum and n is the total number of input data. The task of clustering is to partition $\mathcal{X}$

into c clusters $\{C_j\}_{j=1}^c$. $Y = [y_1, y_2, ..., y_n]^T \in \{0,1\}^{n \times c}$, where $y_i \in \{0,1\}^{c \times 1} (1 \leq i \leq n)$ is the cluster indicator vector for the datum x_i. The j-th element of y_i is 1 if $x_i \in C_j$, and 0 otherwise. Following (Ye, Zhao, and Wu 2008), we define the scaled cluster indicator matrix F as $F = [F_1, ..., F_n]^T = Y(Y^TY)^{-1/2}$, where F_i is the scaled cluster indicator of x_i. It follows that the j-th column of F is given by (Ye, Zhao, and Wu 2008)

$$f_j = [\underbrace{0, ..., 0}_{\sum_{i=1}^{j-1} n_i}, \underbrace{1, ...1}_{n_j}, \underbrace{0, ..., 0}_{\sum_{i=j+1}^{c} n_i}]^T / \sqrt{n_j}, \tag{1}$$

where n_j is the number of data in j-th cluster. The objective function of spectral clustering algorithms can be unified as

$$\min_F Tr(F^TLF) \qquad s.t. \quad F = Y(Y^TY)^{-1/2}, \tag{2}$$

where $Tr(\bullet)$ is the trace operator and L is a Laplacian matrix computed according to local data structure using different strategies (Yang et al. 2010b). Let us define

$$A_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right) & x_i \text{ and } x_j \text{ are } k \text{ nearest neighbors;} \\ 0 & \text{otherwise,} \end{cases}$$

where σ is the bandwidth parameter. Denote I as the identity matrix. The normalized Laplacian matrix L_n is defined as

$$L_n = I - D^{-1/2}AD^{-1/2}, \tag{3}$$

where $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix with its element defined as $D_{ii} = \sum_{j=1}^n A_{ij}$. If we replace L in (2) by L_n, (2) turns to the objective function of the well-known spectral clustering algorithm normalized cut (Shi and Malik 2000; Yu and Shi 2003). Similarly, if we replace L in (2) by L_l which is the Laplacian matrix obtained by local learning (Wu and Schölkopf 2006), (2) is then the objective function of Local Learning Clustering (LLC).

The Proposed NSDR Algorithm

The Objective Function

Because the exploration of local data structures is a good choice to deal with the data sampled from non-linear manifolds (Nie et al. 2010b; Roweis and Saul 2000; Yang et al. 2010a), most of the existing spectral clustering algorithms only utilize local structures for clustering. Although the term "global" occurs in some spectral clustering algorithms such as (Yang et al. 2010b), the Laplacian matrix is purely based on k nearest neighbors of the input data. Sometimes, emphasizing local structures only may induce overfitting and thus degrades the clustering performance. Therefore, we additionally incorporate a regularization term $\Omega(F)$ in the objective function and propose to minimize the following for clustering

$$\begin{aligned} &\min_F Tr(F^TLF) + \lambda\Omega(F) \\ &s.t. \quad F = Y(Y^TY)^{-1/2}, \end{aligned} \tag{4}$$

where $\lambda \geq 0$ is a regularization parameter. $H = I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is the centering matrix and $\mathbf{1}_n \in \mathbb{R}^n$ is a vector of 1s. $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{d \times n}$ represents the data matrix. The between-cluster scatter and total scatter are defined as follows (Fukunaga 1990)

$$S_b = \tilde{X}FF^T\tilde{X}^T, \tag{5}$$

$$S_t = \tilde{X}\tilde{X}^T, \tag{6}$$

where $\tilde{X} = XH$ is the centered data matrix. We assume that the distance between data from different clusters should be as large as possible and the distance between data from the same cluster should be as small as possible. Under such a criterion, it is reasonable to maximize the following

$$\max_F Tr[(S_t + \mu I)^{-1}S_b], \tag{7}$$

where $\mu > 0$ is a parameter and μI is added to make $(S_t + \mu I)$ invertible. Note that

$$\begin{aligned} Tr(F^THF) &= Tr(F^T(I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)F) \\ &= c - \frac{1}{n}Tr[\mathbf{1}_n^TY(Y^TY)^{-1}Y^T\mathbf{1}_n] = c - 1. \end{aligned}$$

Because $Tr(F^THF)$ is a constant, (7) is equivalent to

$$\min_F Tr[F^THF - (S_t + \mu I)^{-1}S_b]. \tag{8}$$

Therefore, the regularization term $\Omega(F)$ in (4) is given by

$$\begin{aligned} \Omega(F) &= Tr[F^THF - (S_t + \mu I)^{-1}S_b] \\ &= Tr[F^THF - F^T\tilde{X}^T(\tilde{X}\tilde{X}^T + \mu I)^{-1}\tilde{X}F]. \end{aligned} \tag{9}$$

We can see that $\Omega(F)$ reflects the global discriminative information. Substituting $\Omega(F)$ in (4) by (9), we arrive at

$$\begin{aligned} &\min_F Tr\{F^T[L + \lambda(H - \tilde{X}^T(\tilde{X}\tilde{X}^T + \mu I)^{-1}\tilde{X})]F\} \\ &\qquad\qquad s.t. \quad F = Y(Y^TY)^{-1/2}. \end{aligned} \tag{10}$$

Recall that F is the scaled cluster indicator matrix. According to definition, in each row of F, only one element is positive and all the others are 0. Such a constraint makes the optimization of (10) an NP-hard problem, which should be relaxed to make the problem solvable.[1] Note that

$$F^TF = (Y^TY)^{-1/2}Y^TY(Y^TY)^{-1/2} = I. \tag{11}$$

Traditional clustering algorithms usually relax the aforementioned constraint and keep the orthogonality intact. Although no element of F can be negative by definition, it is ignored by most of the existing clustering algorithms. If we simply follow the convention of spectral clustering, as those in (Shi and Malik 2000; Yu and Shi 2003; Wu and Schölkopf 2006; Yang et al. 2010b), (10) is relaxed to

$$\min_F Tr[F^T(L + \lambda R)F] \quad s.t. \quad F^TF = I, \tag{12}$$

where $R = H - \tilde{X}^T(\tilde{X}\tilde{X}^T + \mu I)^{-1}\tilde{X}$. The solution of (12) for the relaxed F can be obtained by generalized eigenvalue

[1] For the same reason, (2) is an NP-hard problem as well.

decomposition. Although it is easier to be solved with such a relaxation, the eigenvector solution has mixed signs and it may severely deviate from the ideal solution. Moreover, the mixed signs make it difficult to get the cluster labels. Spectral rotation or k-means is used in previous works to get the cluster labels. Differently, we propose a more accurate relaxation by the nonnegative constraint. The objective function of NSDR is given by

$$\min_F Tr[F^T(L+\lambda R)F] \quad s.t. \quad F^TF=I, F\geq 0. \qquad (13)$$

While other sophisticated Laplacian matrices can be used here, we employ L_n in NSDR for simplicity.

Optimization

In this subsection, we give an algorithm to solve the optimization problem shown in (13). First, we rewrite the objective function of NSDR as follows

$$\min_F Tr[F^T(L+\lambda R)F]+\xi\|F^TF-I\|^2 \qquad s.t. F\geq 0, \qquad (14)$$

where $\xi>0$ is a parameter to control the orthogonality condition. Usually, ξ should be large enough to insure the orthogonality satisfied and we fix it as 10^6 in our experiments. Following (Lee and Seung 2001), we introduce an iterative procedure to optimize (14). Let us define

$$h_{ij}(F_{ij})=h(F)=TrF^TLF+\xi Tr(F^TF-I)^T(F^TF-I).$$

Then we have

$$h'_{ij}(F_{ij})=\frac{\partial h(F)}{\partial F_{ij}}=(2LF+4\xi FF^TF-4\xi F)_{ij}, \qquad (15)$$

which leads to the following updating rules.

- $F_{ij}\leftarrow F_{ij}\frac{(2\xi F)_{ij}}{(LF+2\xi FF^TF)_{ij}}$;
- Normalize F such that $(F^TF)_{ii}=1$ for $i=1,...,n$.

Next, we show that the above updating rules converges.

Definition 1. *$G(f,f')$ is an auxiliary function of $h(f)$ provided that $G(f,f')\geq h(f)$ and $G(f,f)=h(f)$ are satisfied.*

Lemma 1. *If G is an auxiliary function of h, then h is nonincreasing under the following update rule.*

$$f^{(t+1)}=\arg\min_f G(f,f^t). \qquad (16)$$

Detailed proof of Lemma 1 can be found in (Lee and Seung 2001).

Theorem 1. *The objective value of (14) is nonincreasing using the proposed updating rules.*

Proof: Let us define

$$G(F_{ij},F^t_{ij})=h_{ij}(F^t_{ij})+h'_{ij}(F^t_{ij})(F_{ij}-F^t_{ij}) +\frac{(LF^t+2\xi F^tF^{tT}F^t)_{ij}}{F^t_{ij}}(F_{ij}-F^t_{ij})^2. \qquad (17)$$

$G(F_{ij},F^t_{ij})$ is usually an auxiliary function of $h_{ij}(F_{ij})$. By setting $\frac{\partial G(F_{ij},F^t_{ij})}{\partial F_{ij}}=0$, we have

$$h'_{ij}(F^t_{ij})+2\frac{(LF^t+2\lambda F^tF^{tT}F^t)_{ij}}{F^t_{ij}}(F_{ij}-F^t_{ij})=0$$

Then we have

$$\begin{aligned}F_{ij}&=F^t_{ij}-\frac{(LF^t+2\lambda F^tF^{tT}F^t)_{ij}}{F^t_{ij}}h'_{ij}(F^t_{ij})\\&=F^t_{ij}-\frac{(LF^t+2\lambda F^tF^{tT}F^t)_{ij}}{F^t_{ij}}(LF^t+2\lambda F^tF^{tT}F^t-2\lambda F^t)_{ij}\end{aligned}$$

Let us define $F^{t+1}_{ij}=F_{ij}=F^t_{ij}\frac{(2\lambda F^t)_{ij}}{(LF^t+2\lambda F^tF^{tT}F^t)_{ij}}$. It is obvious that

$$h_{ij}(F^{t+1}_{ij})\leq G(F^{t+1}_{ij},F^t_{ij})\leq G(F^t_{ij},F^t_{ij})=h_{ij}(F^t_{ij}).$$

According to Lemma 1, we can see that the objective function value is nonincreasing under the updating rule $F_{ij}\leftarrow F_{ij}\frac{(2\lambda F)_{ij}}{(LF+2\lambda FF^TF)_{ij}}$. □

Discussions

Optimization for the objective function of spectral clustering shown in (2) is an NP-hard problem. Although the elements of F should be nonnegative by definition, most of the existing spectral clustering algorithms simply ignore this constraint when they relax the problem to make it solvable. The mixed signs of F increases the difficulty in getting the cluster labels. Typically, EM-like algorithm or spectral rotation is performed to get the cluster labels in previous spectral clustering algorithms.

Besides the global discriminative regularizer added into the objective function, the major difference between NSDR and existing spectral clustering algorithms is that the nonnegative constraint is explicitly imposed, making the results much closer to the ideal solution. Provided that both orthogonal and nonnegative constraints are satisfied, in each row of F, only one element is of positive value and all of the others are 0. Therefore, F can be directly used to assign the cluster labels for the input data. Taking the UMIST dataset[2] as an example, we plot the absolute values of the first 20 rows of the optimal F corresponding to (12) and (13) respectively in Fig.1. Fig.1(a) is the cluster indicator matrix obtained from the traditional relaxation and Fig.1(b) is the cluster indicator matrix obtained from nonnegative relaxation. Note that the first 20 images in UMIST dataset are from an identical cluster. In Fig.1(b) these 20 images are directly grouped into one cluster (the 4-th cluster). However, it remains unclear how to assign the cluster labels to the input data according to Fig.1(a) directly.

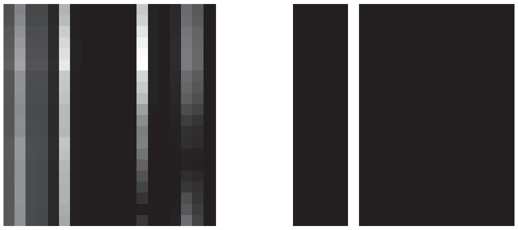
The algorithm proposed in this paper is a general one, which naturally leads to other nonnegative clustering algorithms. Next, we show some examples by three propositions.

Proposition 1. NSDR leads to nonnegative spectral clustering when $\lambda=0$.

Proof: This proposition naturally holds because if $\lambda=0$ (4) reduces to (2). □

According to Proposition1, our optimization approach is readily to extend any spectral clustering to its nonnegative version. For example, if we set $\lambda=0$ and replace L in (13) by the Laplacian matrix L_l, which is proposed in Local Learning Clustering (LLC) (Wu and Schölkopf 2006), it leads to nonnegative LLC. Similarly, under our framework

[2] http://images.ee.umist.ac.uk/danny/database.html



(a) Mixed-signs (b) Nonnegative

Figure 1: First 20 cluster indicator vectors for UMIST. Each row is a cluster indicator vector for an input datum. The results are normalized for a clearer illustration.

any spectral clustering algorithm can be extended to its nonnegative version.

Proposition 2. NSDR leads to nonnegative k-means clustering when $\mu \to \infty$ and $\lambda \to \infty$.

Proof: It has been shown in (Zha et al. 2001) that the objective function of k-means clustering is

$$\max_{F^TF=I} Tr(F^T\tilde{X}^T\tilde{X}F). \tag{18}$$

If we keep the nonnegative constraint for k-means clustering during relaxation, we have

$$\max_{F} Tr(F^T\tilde{X}^T\tilde{X}F) \quad s.t. F^TF = I, F \geq 0. \tag{19}$$

Because $Tr(F^THF) = c - 1$ is a constant, if $\lambda \to \infty$, (13) is equivalent to

$$\max_{F} Tr(F^T\tilde{X}^T(\tilde{X}\tilde{X}^T + \mu I)^{-1}\tilde{X}F) \qquad s.t. F^TF = I, F \geq 0. \tag{20}$$

If $\mu \to \infty$, it is easy to see that (20) is equivalent to (19). □

Recently, some researchers suggested to incorporate dimension reduction and clustering into a joint framework for high dimensional data. Let W be the projection matrix for dimension reduction, this family of algorithms, referred to as discriminative k-means, tries to optimize the following objective function

$$\max_{W,F} Tr\left[\left(W^T(\tilde{X}\tilde{X}^T + \gamma I)W\right)^{-1} W^T\tilde{X}FF^T\tilde{X}^TW\right]. \tag{21}$$

In (21), there are two variables, i.e. W and F, to be optimized. Initial work usually iteratively optimizes W and F (Torre and Kanade 2006). More recently, (Ye, Zhao, and Wu 2008) proved that the optimization problem can be solved by optimize F only and they simplify (21) as follows:

$$\max_{F^TF=I} Tr\left\{F^T\left[I - (I + \frac{1}{\gamma}\tilde{X}^T\tilde{X})^{-1}\right]F\right\}. \tag{22}$$

In discriminative k-means proposed by (Torre and Kanade 2006; Ye, Zhao, and Wu 2008), the nonnegative constraint is also ignored. Discriminative k-means can be extended to its nonnegative version under our framework by the following proposition.

Proposition 3. NSDR leads to nonnegative discriminative k-means clustering (Ye, Zhao, and Wu 2008) when $\lambda \to \infty$.

Table 1: Database Descriptions.

Dataset	Size	Dimension	# of Classes
Ecoil	336	343	8
Yeast	1484	1470	10
UMIST	575	644	20
MSRA	1799	256	12
USPS	9298	256	10
WebKB	814	4029	7

Proof: As shown previously, when $\lambda \to \infty$, (13) is equivalent to (20). Note that

$$\begin{aligned}
&\max_{F^TF=I} Tr\left\{F^T\left[I - (I + \frac{1}{\gamma}\tilde{X}^T\tilde{X})^{-1}\right]F\right\} \\
\Leftrightarrow\ &\max_{F^TF=I} Tr\left\{F^T\left[(I + \frac{1}{\gamma}\tilde{X}^T\tilde{X})^{-1}(\frac{1}{\gamma}\tilde{X}^T\tilde{X} + I - I)\right]F\right\} \\
\Leftrightarrow\ &\max_{F^TF=I} Tr\left[F^T\tilde{X}^T(\frac{1}{\gamma}\tilde{X}\tilde{X}^T + I)^{-1}\tilde{X}F\right].
\end{aligned}$$

Therefore, if we add nonnegative constraint to (22) and when $\lambda \to \infty$, it is equivalent to (13). □

Experiments

Experiment Setup

We compare our NSDR with k-means (KM), discriminative k-means (DKM) (Ye, Zhao, and Wu 2008) and two representative spectral clustering algorithms, i.e., Nonnegative Normalized Cut (NNcut) (Ding, Li, and Jordan 2008), and Local Learning Clustering (LLC) (Wu and Schölkopf 2006). To show the effectiveness of imposing the nonnegative constraint during the relaxation, we additionally report the results corresponding to (12), where the nonnegative constraint is removed in NSDR and we denote it as SDR.

We set k, which specifies the size of neighborhood, to 5 for all the spectral clustering algorithms. We perform the self-tuning algorithm (Zelnik-Manor and Perona 2004) to determine σ in (3) for NNcut, SDR and NSDR. For the parameters in DKM, SDR, NSDR and LLC, we tune them from $\{10^{-6}, 10^{-3}, 10^{0}, 10^{3}, 10^{6}\}$ and report their best results. The results of all the clustering algorithms depend on initialization. To reduce statistical variation, each clustering algorithms is independently repeated 20 times with random initialization and we report the results corresponding to best objective function values. For LLC and SDR, spectral rotation is used to discretize the relaxed cluster indicator matrix in order to obtain the cluster labels of the input data. For NSDR, we use the results from SDR as initialization.

In our experiment, we have collected 6 public datasets, including two UCI datasets Ecoil and Yeast[3], two face image datasets UMIST and MSRA (He et al. 2004), one hand written digital image dataset USPS[4] and one text database WebKB collected by the University of Texas (Craven et al. 1998). Detailed information of the six datasets is summarized in Table 1.

[3] http://archive.ics.uci.edu/ml/

[4] http://www-i6.informatik.rwth-aachen.de/ keysers/usps.html

Table 2: Performance Comparison (ACC %) of KM, DKM, NNCut, LLC, SDR and NSDR

	KM	DKM	NNcut	LLC	SDR	NSDR
Ecoil	51.8	60.8	53.1	48.5	60.1	**61.3**
Yeast	32.8	35.0	29.8	26.3	38.1	**41.2**
UMIST	43.6	46.3	61.1	61.0	61.3	**64.5**
MSRA	52.7	67.7	55.6	56.3	85.3	**86.3**
USPS	67.3	68.8	72.6	70.8	80.9	**82.5**
WebKB	56.5	57.8	58.4	42.6	58.9	**60.2**

Table 3: Performance Comparison (NMI %) of KM, DKM, NNCut, LLC, SDR and NSDR

	KM	DKM	NNcut	LLC	SDR	NSDR
Ecoil	48.9	52.6	50.2	47.1	50.6	**55.0**
Yeast	16.5	17.5	15.9	14.5	22.9	**23.8**
UMIST	66.1	66.7	80.6	78.2	80.6	**81.8**
MSRA	62.5	72.5	68.3	66.4	90.9	**92.8**
USPS	61.6	63.2	83.6	75.7	82.5	**86.2**
WebKB	14.6	13.3	15.6	8.3	17.0	**17.6**

Evaluation Metrics

Following the convention of clustering study, we use Accuracy (ACC) and Normalized Mutual Information (NMI) as evaluation metrics. Denote q_i as the clustering result from the clustering algorithm and p_i as the ground truth label of x_i. ACC is defined as:

$$ACC = \frac{\sum_{i=1}^{n} \delta(p_i, map(q_i))}{n} \tag{23}$$

where $\delta(x, y) = 1$ if $x = y$; $\delta(x, y) = 0$ otherwise, and $map(q_i)$ is the best mapping function that permutes clustering labels to match the ground truth labels using the Kuhn-Munkres algorithm. A larger ACC indicates better performance.

For two arbitrary variables *P* and *Q*, NMI is defined as follows (Strehl and Ghosh 2002):

$$NMI(P, Q) = \frac{I(P, Q)}{\sqrt{H(P)H(Q)}}, \tag{24}$$

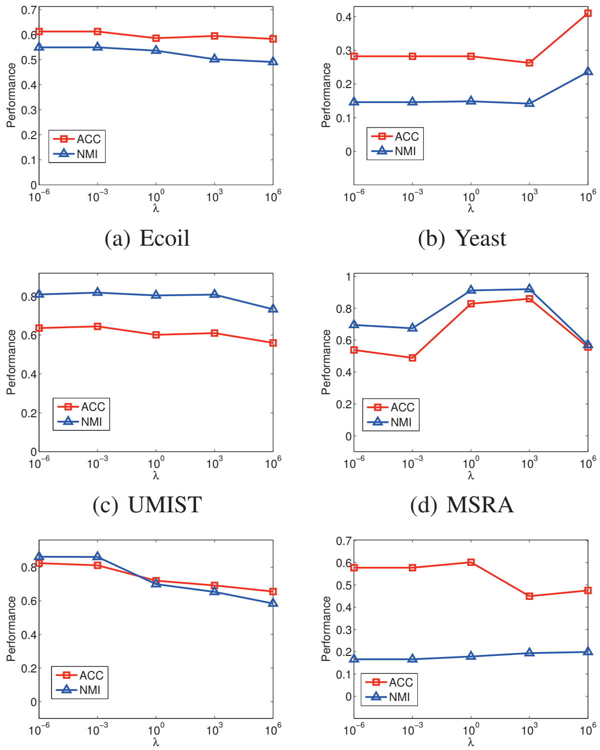
where $I(P, Q)$ is the mutual information between P and Q, and $H(P)$ and $H(Q)$ are the entropies of P and Q. t_l is the number of data in the cluster $\mathcal{C}_l$ $(1 \leq l \leq c)$ obtained from clustering algorithms and $\tilde{t}_h$ be the number of data in the h-th ground truth class $(1 \leq h \leq c)$. NMI is defined as (Strehl and Ghosh 2002):

$$NMI = \frac{\sum_{l=1}^{c} \sum_{h=1}^{c} t_{l,h} \log(\frac{n \cdot t_{l,h}}{t_l \tilde{t}_h})}{\sqrt{\left(\sum_{l=1}^{c} t_l \log \frac{t_l}{n}\right) \left(\sum_{h=1}^{c} \tilde{t}_h \log \frac{\tilde{t}_h}{n}\right)}}, \tag{25}$$

where $t_{l,h}$ is the number of samples that are in the intersection between the cluster $\mathcal{C}_l$ and the h-th ground truth class. Similarly, a larger NMI indicates better clustering results.

Experiment Results

Table 2 and Table 3 are the clustering results of different algorithms over the 6 datasets. Fig.2 shows the performance variation *w.r.t.* the regularization parameter λ. Fig.3 shows convergence curves of NSDR over all the six datasets. From the tables and figures, we have the following observations.



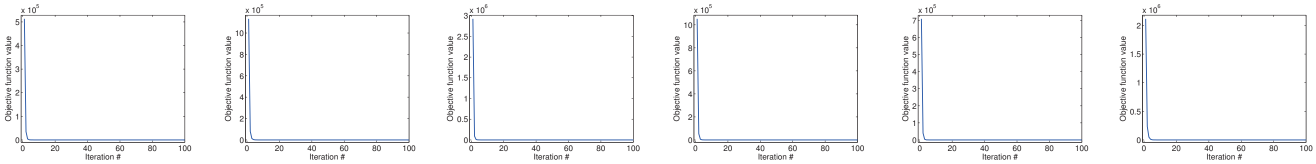
(a) Ecoil (b) Yeast

(c) UMIST (d) MSRA

(e) USPS (f) WebKB

Figure 2: Performance variation of NSDR *w.r.t.* the regularization parameter λ.

- DKM outperforms KM because it incorporates discriminative dimension reduction and clustering into a joint framework (Ye, Zhao, and Wu 2008). This observation also indicates that it is helpful to utilize discriminative information for clustering.
- Although the Laplacian matrix of LLC is more sophisticated than that of NNcut, NNcut generally shows better performance. A possible reason is that the nonnegative constraint is imposed in NNCut during spectral relaxation while LLC ignores it.
- SDR outperforms KM, DKM, NNcut and LLC, demonstrating that simultaneously utilizing local data structures and global discriminative information is beneficial for data clustering.
- NSDR achieves the best performance for all of the six datasets because
  - NSDR simultaneously exploits local data structure information and discriminative information for data clustering;
  - the nonnegative constraint is imposed during spectral relaxation, making the results more faithful.



(a) Ecoil (b) Yeast (c) UMIST (d) MSRA (e) USPS (f) WebKB

Figure 3: Convergence curve of NSDR.

- NSDR is not very sensitive to the regularization parameter though the clustering performance is different when we set this parameter as different values. The optimal value for algorithmic parameter is data dependent. How to choose the optimal parameter automatically will be studied in our future work.
- The proposed iterative approach to optimize the objective function of NSDR always converges very fast, usually less than 50 iterations.

Conclusion

In this paper, we proposed a new spectral clustering algorithm NSDR. While most of the existing spectral clustering algorithms only utilize local data structures for clustering, we additionally take global discriminative information into account to make the results more robust. It is an NP-hard problem to optimize the objective function of any spectral clustering algorithm. Traditional relaxation usually neglects the nonnegative constraint. The relaxed cluster indicator matrix has mixed signs, making it difficult to get the cluster labels. Moreover, the mixed signed cluster indicator matrix may deviate severely from the ideal solution. We therefore impose the nonnegative constraint during the relaxation. We proposed a new and efficient iterative algorithm to solve the nonnegative optimization problem. We show that the proposed algorithm is a general one which naturally leads to many other extensions.

Acknowledgement

This work is supported by ARC DP1094678 and partially supported by the NSFC project No.61003127.

References

Cai, D.; He, X.; Bao, H.; and Han, J. 2009. Locality perserving nonnegative matrix factorization. In *IJCAI*.

Craven, M.; DiPasquo, D.; Freitag, D.; McCallum, A.; Mitchell, T. M.; Nigam, K.; and Slattery, S. 1998. Learning to extract symbolic knowledge from the world wide web. In *AAAI/IAAI*.

Ding, C.; Li, T.; and Jordan, M. I. 2008. Nonnegative matrix factorization for combinatorial optimization: Spectral clustering, graph matching, and clique finding. In *ICDM*.

Fukunaga, K. 1990. *Introduction to Statistical Pattern Recognition,2nd Edition.* Boston, MA: Academic Press.

He, X.; Yan, S.; Hu, Y.; Niyogi, P.; and Zhang, H.-J. 2004. Face recognition using laplacianfaces. *IEEE TPAMI* 27(3):328–340.

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

Lee, D., and Seung, H. 2001. Algorithms for nonnegative matrix factorization. In *NIPS*.

Liu, H., and Wu, Z. 2010. Non-negative matrix factorization with constraints. In *AAAI*.

Liu, Y.; Wu, F.; Zhang, Z.; Zhuang, Y.; and Yan, S. 2010. Sparse representation using nonnegative curds and whey. In *CVPR*, 3578–3585.

Nie, F.; Xu, D.; Tsang, I. W.; and Zhang, C. 2009. Spectral embeded clustering. In *IJCAI*.

Nie, F.; Ding, C. H. Q.; Luo, D.; and Huang, H. 2010a. Improved minmax cut graph clustering with nonnegative relaxation. In *ECML/PKDD*, 451–466.

Nie, F.; Xu, D.; Tsang, I. W.-H.; and Zhang, C. 2010b. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Transactions on Image Processing* 19(7):1921–1932.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE TPAMI* 22(8):888–905.

Strehl, A., and Ghosh, J. 2002. Cluster ensembles–a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research* 3:583–617.

Torre, F. D., and Kanade, T. 2006. Discriminative cluster analysis. In *ICML*.

Wu, M., and Schölkopf, B. 2006. A local learning approach for clustering. In *NIPS*.

Yang, Y.; Nie, F.; Xiang, S.; Zhuang, Y.; and Wang, W. 2010a. Local and global regressive mapping for manifold learning with out-of-sample extrapolation. In *AAAI*.

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

Ye, J.; Zhao, Z.; and Wu, M. 2008. Discriminative k-means for clustering. In *NIPS*.

Yu, S. X., and Shi, J. 2003. Multiclass spectral clustering. In *ICCV*.

Zelnik-Manor, L., and Perona, P. 2004. Self-tuning spectral clusering. In *NIPS*.

Zha, H.; He, X.; Ding, C.; Gu, M.; and Simon, H. 2001. Spectral relaxation for k-means clustering. In *NIPS*.