Spectral Embedded Clustering*

Feiping Nie[1,2], Dong Xu[2], Ivor W. Tsang[2] and Changshui Zhang[1]
[1]State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology(TNList)
Department of Automation, Tsinghua University, Beijing 100084, China
[2]School of Computer Engineering, Nanyang Technological University, Singapore
nfp03@mails.tsinghua.edu.cn; DongXu@ntu.edu.sg; IvorTsang@ntu.edu.sg; zcs@mail.tsinghua.edu.cn

Abstract

In this paper, we propose a new spectral clustering method, referred to as Spectral Embedded Clustering (SEC), to minimize the normalized cut criterion in spectral clustering as well as control the mismatch between the cluster assignment matrix and the low dimensional embedded representation of the data. SEC is based on the observation that the cluster assignment matrix of high dimensional data can be represented by a low dimensional linear mapping of data. We also discover the connection between SEC and other clustering methods, such as spectral clustering, Clustering with local and global regularization, K-means and Discriminative K-means. The experiments on many real-world data sets show that SEC significantly outperforms the existing spectral clustering methods as well as K-means clustering related methods.

1 Introduction

Clustering is a fundamental task of many machine learning, data mining and pattern recognition problems. Clustering aims at grouping the similar patterns into the same cluster, and discovering the meaningful structure of the data [Jain and Dubes, 1988]. In the past decades, many clustering algorithms have been developed such as K-means clustering, mixture models [McLachlan and Peel, 2000], spectral clustering [Ng *et al.*, 2001; Shi and Malik, 2000; Yu and Shi, 2003], support vector clustering [Ben-Hur *et al.*, 2001], and maximum margin clustering [Xu *et al.*, 2005; Zhang *et al.*, 2007; Li *et al.*, 2009].

It is a challenging task to partition the high dimensional data into different clusters. In practice, many high dimensional data may exhibit dense grouping in a low dimensional space. Hence, the researchers usually first project the high dimensional data onto the low dimensional subspace via some dimension reduction techniques such as Principle Component Analysis (PCA). To achieve better clustering performance, several works have been proposed to perform K-means clustering and dimension reduction iteratively for high dimensional data [la Torre and Kanade, 2006; Ding and Li, 2007; Ye *et al.*, 2007]. Recently, [Ye *et al.*, 2008] proposed Discriminative K-means (DisKmeans) to unify the iterative procedure of dimension reduction and K-means clustering into a unified trace maximization problem. The improved clustering performance was also demonstrated, when compared with the standard K-means. However, DisKmeans did not consider the geometry structure (a.k.a. manifold) of the data.

The use of manifold information in Spectral Clustering (SC) has shown the state-of-the-art clustering performance in many computer vision applications, such as segmentation [Shi and Malik, 2000; Yu and Shi, 2003]. But, the existing SC methods did not map the data into the low dimensional space for clustering. In this paper, we first show that the cluster assignment matrix of data can be represented by a low dimensional linear mapping of data, when the dimensionality of data is high enough. Thereafter, we explicitly incorporate this prior knowledge into spectral clustering. More specifically, we minimize the normalized cut criterion in SC as well as control the mismatch between the cluster assignment matrix and the low dimensional embedded representation of the data. The proposed clustering method is then referred to as Spectral Embedded Clustering (SEC).

The rest of this paper is organized as follows. Section 2 first revisit the Spectral Clustering and the cluster assignment methods. Our proposed method is presented in Section 3. Connections to other clustering methods are discussed in Section 4. Experimental results on real-world data sets are reported in Section 5 and the conclusion remarks are given in Section 6.

2 Brief Review of Spectral Clustering

Given a data set $\mathcal{X} = \{x_i\}_{i=1}^n$, clustering is to partition $\mathcal{X}$ into c clusters. Denote the cluster assignment matrix by $Y = [y_1, y_2, ..., y_n]^T \in \mathbb{B}^{n\times c}$, where $y_i \in \mathbb{B}^{c\times 1}(1 \le i \le n)$ is the cluster assignment vector for the pattern x_i. The j-th element of y_i is 1, if the pattern x_i is assigned to the j-th cluster; 0, otherwise. The main task of a clustering algorithm is to learn the cluster assignment matrix Y. Clustering is a non-trivial problem because Y is constrained as integer solution. In this Section, we first revisit spectral clustering method and the techniques to obtain the discrete cluster assignment matrix.

*This material is based upon work funded by Singapore National Research Foundation Interactive Digital Media R&D Program (Grant No. NRF2008IDM-IDM-004-018) and NSFC (Grant No. 60835002).

2.1 Spectral Clustering

Since last decade, Spectral Clustering (SC) has attracted much attention. Several algorithms have been proposed in the literature [Ng *et al.*, 2001; Shi and Malik, 2000; Yu and Shi, 2003]. Here, we focus on the spectral clustering algorithm with k-way normalized cut [Yu and Shi, 2003].

Let us denote $\mathcal{G} = \{\mathcal{X}, A\}$ as an undirected weighted graph with a vertex set $\mathcal{X}$ and an affinity matrix $A \in \mathbb{R}^{n\times n}$, in which each entry A_{ij} of the symmetric matrix A represents the affinity of a pair of vertices. The common choice of A_{ij} is defined by

$$A_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right) & x_i \text{ and } x_j \text{ are neighbors;} \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where σ is the parameter to control the spread of neighbors. The graph Laplacian matrix L is then defined by $L = D - A$, where D is a diagonal matrix with the diagonal elements as $D_{ii} = \sum_j A_{ij}, \forall\, i$. Let us denote $tr(A)$ as the trace operator of a matrix A. The minimization of the normalized cut criterion can be transformed to the following maximization problem [Yu and Shi, 2003]:

$$\max_{Z^T D Z = I} tr(Z^T A Z), \quad (2)$$

where $Z = Y(Y^T D Y)^{-1/2}$.

Let us define a scaled cluster assignment matrix F by

$$F = D^{1/2} Z = D^{1/2} Y (Y^T D Y)^{-1/2} = f(Y).$$

Then the objective function (2) can be rewritten as:

$$\max_{F^T F = I} tr(F^T D^{-1/2} A D^{-1/2} F). \quad (3)$$

where $F = D^{1/2} Y (Y^T D Y)^{-1/2}$. Note that the elements of F are constrained to be discrete values, which makes the problem (3) hard to solve.A well-known solution to this problem is to relax the matrix F from the discrete values to the continuous ones. Then the problem becomes:

$$\max_{F^T F = I} tr(F^T K F), \quad (4)$$

where $K = D^{-1/2} A D^{-1/2}$.

The optimal solution of problem (4) can be obtained[1] by the eigenvalue decomposition of the matrix K. Based on the relaxed continuous solution, the final discrete solution is then obtained by K-means or spectral rotation.

2.2 Cluster Assignment Methods

With the relaxed continuous solution $F \in \mathbb{R}^{n\times c}$ from spectral decomposition, K-means or spectral rotation can be used to calculate the discrete solution $Y \in \mathbb{B}^{n\times c}$.

K-Means

The input to K-means clustering is n points, in which the i-th data point is the i-th row of F. The standard K-means algorithm is performed to obtain the discrete-valued cluster assignment for each pattern. [Ng *et al.*, 2001] used this technique for assigning cluster labels.

[1]A trivial eigenvector $D^{1/2}\mathbf{1}$ corresponding to the largest eigenvalue of K is removed in spectral clustering.

Spectral Rotation

Note that the global optimal F of the optimization problem (4) is not unique. Let $F^* \in \mathbb{R}^{n\times c}$ be the matrix whose columns consist of top c eigenvectors of K and $R \in \mathbb{R}^{c\times c}$ be an orthogonal matrix. Then F can be F^*R for any R. To obtain the final clustering result, we need to find a discrete-valued cluster assignment matrix which is close to F^*R. The work in [Yu and Shi, 2003] also defined a mapping to obtain the corresponding Y^*:

$$Y^* = f^{-1}(F^*) = Diag(F^* F^{*T})^{-1/2} F^*,$$

where $Diag(M)$ denotes a diagonal matrix with the same size and the same diagonal elements as the square matrix M. It can be easily verified that $f^{-1}(F^*R) = Y^*R$.

As F^*R is the optimal solution to the relaxed problem (4) for arbitrary orthogonal matrix R, a suitable R should be selected such that Y^*R is closest to a discrete cluster assignment matrix Y. The optimal R and Y are then obtained by solving the following optimization problem [Yu and Shi, 2003]:

$$\min_{Y\in\mathbb{B}^{n\times c}, R\in\mathbb{R}^{c\times c}} \|Y - Y^*R\|^2$$
$$\text{subject to} \quad Y\mathbf{1}_c = \mathbf{1}_n,\ R^T R = I,$$

where $\mathbf{1}_c$ and $\mathbf{1}_n$ denote the $c\times 1$ and $n\times 1$ vectors of all 1's respectively. [Yu and Shi, 2003] used this technique to obtain the cluster assignment matrix by iteratively solving Y and R.

3 Spectral Embedded Clustering

Denote the data matrix by $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$. For simplicity, we assume the data is centered, i.e. $X\mathbf{1}_n = \mathbf{0}$. Let us define the total scatter matrix S_t, the between-cluster scatter matrix S_b and the within-cluster scatter matrix S_w as:

$$S_t = XX^T, \quad (5)$$
$$S_b = XGG^TX^T, \quad (6)$$
$$S_w = XX^T - XGG^TX^T, \quad (7)$$

where $G = Y(Y^TY)^{-1/2}$, and Y is defined as in Section 2. It is easy to verify that $G^TG = I$.

In next subsections, we will introduce our proposed clustering method, referred to as Spectral Embedded Clustering (SEC).

3.1 Low Dimensional Embedding for Cluster Assignment Matrix

Traditional SC methods partition data based only on the manifold structure of data. However, when the manifold is not well-defined, the SC method may not perform well. To improve the clustering performance, we will apply the following theorem in the design of SEC [2]

Theorem 1. *If $rank(S_b) = c - 1$ and $rank(S_t) = rank(S_w) + rank(S_b)$, then the true cluster assignment matrix can be represented by a low dimensional linear mapping of the data, that is, there exist $W \in \mathbb{R}^{d\times c}$ and $b \in \mathbb{R}^{c\times 1}$ such that $Y = X^TW + \mathbf{1}_n b^T$.*

[2]Due to the space limitation, we omit the proof of this theorem in the paper. The proof can be downloaded at: http://feipingnie.googlepages.com/ijcai09_clustering_proof.pdf.

As noted in [Ye, 2007], the conditions in Theorem 1 are usually satisfied for the high-dimensional and small-sample-size problem, which is usually the case in many real-world applications. According to Theorem 1, the true cluster assignment matrix can be always embedded into a low dimensional linear mapping of the data. To utilize such constraints, we explicitly add a new regularizer into the objective function in SEC.

3.2 Proposed Formulation

In spectral clustering, the optimization problem (4) is equivalent to the following problem:

$$\min_{F^TF=I} tr(F^T\tilde{L}F), \tag{8}$$

where $\tilde{L} = D^{-\frac{1}{2}}LD^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ is the normalized Laplacian matrix.

In addition, we expect that the learned F is close to a linear space spanned by the data X. To this end, we propose to solve the following optimization problem:

$$\min_{F^TF=I,W,b} tr(F^T\tilde{L}F)+\mu(trW^TW+\gamma\|X^TW+\mathbf{1}_nb^T-F\|^2), \tag{9}$$

where μ and γ are two tradeoff parameters to balance three terms. In (9), the first term reflects the smoothness of data manifold; while the third term characterizes the mismatch between the relaxed cluster assignment matrix F and the low dimensional representation of the data.

3.3 Detailed Algorithm

To obtain the optimal solution to (9), we set the derivatives of the objective function with respect to b and W to zeros. Note that the data are centered, i.e, $X\mathbf{1}_n = \mathbf{0}$. Then we have:

$$b = \frac{1}{n}F^T\mathbf{1}_n \quad \text{and} \quad W = \gamma(\gamma XX^T + I)^{-1}XF. \tag{10}$$

Replacing W and b in (9) by (10), the optimization problem (9) becomes:

$$\min_{F^TF=I} F^T(\tilde{L}+\mu\gamma H_c-\mu\gamma^2X^T(\gamma XX^T+I)^{-1}X)F, \tag{11}$$

where $H_c = I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is the centering matrix. The global optimal solution F^* to (11) can be obtained by eigenvalue decomposition. The columns of F^* are from the bottom c eigenvectors of the matrix $\tilde{L} + \mu\gamma H_c - \mu\gamma^2X^T(\gamma XX^T + I)^{-1}X$. Based on F^*, the discrete-valued cluster assignment matrix can be obtained by K-means or spectral rotation. The details of the proposed SEC are outlined in Algorithm 1.

Algorithm 1 : The algorithm of SEC

Given a sample set $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$ and the number of clusters c.

1: Compute the normalized Laplacian matrix $\tilde{L}$.
2: Solve (11) with eigenvalue decomposition and obtain the optimal F^*.
3: Based on F^*, compute the discrete cluster assignment matrix Y by using K-means or spectral rotation.

4 Connections to Prior Work

In this Section, we discuss the connection between SEC and Spectral Clustering, Clustering with Local and Global Regularization, K-means and Discriminative K-means.

4.1 Connection between SEC and Spectral Clustering

SEC reduces to spectral clustering, if μ is set as zero. Therefore spectral clustering is a special case of SEC.

4.2 Connection between SEC and Clustering with Local and Global Regularization

Recently, [Wang *et al.*, 2007] proposed Clustering with Local and Global Regularization (CLGR), which solves the following problem:

$$\min_{F^TF=I} tr(F^T(L+\mu L_l)F), \tag{12}$$

where L_l is another Laplacian matrix constructed using local learning regularization [Wu and Schölkopf, 2007].

Let us denote the cluster assignment matrix $F = [f_1, ..., f_n]^T \in \mathbb{R}^{n\times c}$. We also define the k neighbors of x_i as $\mathcal{N}(x_i) = \{x_{i_1}, ..., x_{i_k}\}$, $X_i = [x_{i_1}, ..., x_{i_k}] \in \mathbb{R}^{d\times k}$ and $F_i = [f_{i_1}, ..., f_{i_k}]^T \in \mathbb{R}^{k\times c}$. In local learning regularization, for each x_i, a locally linear projection $W_i \in \mathbb{R}^{d\times c}$ is learned by minimizing the following structural risk functional [Wang *et al.*, 2007]:

$$\min_{W_i} \sum_{x_j\in\mathcal{N}(x_i)} \|W_i^Tx_j - f_j\|^2 + \gamma tr(W_i^TW_i).$$

One can obtain the closed form solution for W_i:

$$W_i = (X_iX_i^T + \gamma I)^{-1}X_iF_i. \tag{13}$$

After all the locally linear projections are learnt, the cluster assignment matrix F can be found by minimizing the following criterion:

$$\mathcal{J}(F) = \sum\nolimits_{i=1}^{n} \|x_i^TW_i - f_i^T\|^2. \tag{14}$$

Substituting (13) back to (14), we have

$$\mathcal{J}(F) = tr(F^T(N-I)^T(N-I)F) = tr(F^TL_lF),$$

where $L_l = (N-I)^T(N-I)$ and $N \in \mathbb{R}^{n\times n}$ with its (i,j)-th entry as:

$$N_{ij} = \begin{cases} a_h^i, & if\ x_j \in \mathcal{N}(x_i)\ and\ j = i_h (h = 1, ..., k); \\ 0, & otherwise; \end{cases}$$

in which a_h^i denotes the h-th entry of $a^i = x_i^T(X_iX_i^T + \gamma I)^{-1}X_i$.

One can observe that $L + \mu L_l$ in (12) is also a Laplacian matrix, and so CLGR is just one variant of SC, which combines the objectives of spectral clustering and the clustering using local learning regularization in (14). Therefore, CLGR is also a spectral case of SEC when $L + \mu L_l$ is used in (8).

It is worthwhile to mention that our SEC is fundamentally different from CLGR in the following two aspects: 1) CLGR uses two-step approach to learn the linear regularized models

and the cluster assignment matrix. First, it calculates a series of *local projection matrices* $W_i (i = 1, ..., n)$ and then obtains the cluster assignment matrix F using (12). In contrast, SEC solves the *global projection matrix* W and the cluster assignment matrix F simultaneously. 2) It is unclear how to use CLGR to cope with the new-coming data. In contrast, the global projection matrix W in SEC can be used for clustering new-coming data.

4.3 Connection between SEC and K-means

K-means is a simple and frequently used clustering algorithm. As shown in [Zha *et al.*, 2001], the objective of K-means is to minimize the following criterion:

$$\min_{G^TG=I} tr(S_w) = \min_{G^TG=I} tr(XX^T - XGG^TX^T) \quad (15)$$

where G is defined as in (6). The problem (15) is simplified as the following problem:

$$\max_{G^TG=I} tr(G^TX^TXG). \quad (16)$$

Traditional K-means uses an EM-like iterative method to solve the above problem. The spectral relaxation can also be used to solve the K-means problem [Zha *et al.*, 2001].

We will prove that the objective function of the proposed SEC reduces to that of K-means, when $\gamma \to 0$ and $\mu\gamma \to \infty$ in SEC. The objective function of SEC in (11) is equivalent to the following optimization problem:

$$\max_{F^TF=I} F^T(K + \frac{\mu\gamma}{n}\mathbf{1}_n\mathbf{1}_n^T + \mu\gamma^2 X^T(\gamma XX^T + I)^{-1}X)F, \quad (17)$$

where K is the same matrix as in (4).

When $\mu\gamma \to \infty$, (17) reduces to:

$$\max_{F^TF=I} F^T(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T + \gamma X^T(\gamma XX^T + I)^{-1}X)F.$$

This problem has a trivial solution $\mathbf{1}_n$ corresponding to the largest eigenvalue of the matrix $\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T + \gamma X^T(\gamma XX^T + I)^{-1}X$. Therefore, we add a new constraint $F^T\mathbf{1}_n = \mathbf{0}$:

$$\begin{aligned}
&\max_{F^TF=I, F^T\mathbf{1}_n=\mathbf{0}} F^T(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T + \gamma X^T(\gamma XX^T + I)^{-1}X)F \\
\Leftrightarrow &\max_{F^TF=I, F^T\mathbf{1}_n=\mathbf{0}} F^T(X^T(\gamma XX^T + I)^{-1}X)F \\
\Leftrightarrow &\max_{F^TF=I} F^T(X^T(\gamma XX^T + I)^{-1}X)F. \quad (18)
\end{aligned}$$

When $\gamma \to 0$, the optimization problem in (18) reduces to the optimization problem in (16). Therefore, the objective function of SEC reduces to that of K-means algorithm, if $\gamma \to 0$ and $\mu\gamma \to \infty$.

4.4 Connection between SEC and Discriminative K-means

Subspace clustering methods were proposed to learn the low-dimensional subspace and data cluster simultaneously [Ding *et al.*, 2002; Li *et al.*, 2004], possibly because high dimensional data may exhibit dense grouping in a low dimensional space. For instances, Discriminative Clustering methods solve the following optimization problem:

$$\max_{W,G} tr(W^T(\gamma S_t + I)W)^{-1}W^TS_bW, \quad (19)$$

where S_t and S_b are defined in (5) and (6), respectively.

There are two sets of variables, the projection matrix W and the scaled cluster assignment matrix G, in (19). Most of the existing works optimize W and G iteratively [la Torre and Kanade, 2006; Ding and Li, 2007; Ye *et al.*, 2007]. However, a recent work Discriminative K-means [Ye *et al.*, 2008] simplified (19) by optimizing G only, which is based on the following observation [Ye, 2005]:

$$tr(W^T(\gamma S_t + I)W)^{-1}W^TS_bW \leq tr(\gamma S_t + I)^{-1}S_b, \quad (20)$$

where the equality holds when $W = VM$, and V is composed of the eigenvectors of $(\gamma S_t + I)^{-1}S_b$ corresponding to all the nonzero eigenvalues, M is an arbitrary nonsingular matrix.

Based on (20), the optimization problem (19) can be simplified as:

$$\max_G tr(\gamma S_t + I)^{-1}S_b. \quad (21)$$

Replacing (5) and (6) into (21) and adding the constraint $G^TG = I$ in (21), we arrive at:

$$\max_{G^TG=I} trG^T(X^T(\gamma XX^T + I)^{-1}X)G. \quad (22)$$

Recall that (17) reduces to (18) in SEC, when γ is a nonzero constant and $\mu \to \infty$. We also observe that the optimization problem (18) in SEC and (22) in Discriminative K-means [Ye *et al.*, 2008] are exactly the same. Therefore, when $\mu \to \infty$, SEC reduces to Discriminative K-means algorithm, if the spectral relaxation is used to solve the cluster assignment matrix in Discriminative K-means algorithm.

In addition, we observe that K-means and Discriminative K-means will lead to the same results, if the spectral relaxation is used to solve the cluster assignment matrices. Note that $X^T(\gamma XX^T + I)^{-1}X = \frac{1}{\gamma}I - \frac{1}{\gamma}(\gamma X^TX + I)^{-1}$. Thus $X^T(\gamma XX^T + I)^{-1}X$ in the optimization problem (22) and X^TX in the optimization problem (16) have the same top c eigenvectors. The results from K-means and Discriminative K-means are reported to be different because EM-like method is used to solve the cluster assignment matrices of the optimization problem in (16) and (22) for K-means and Discriminative K-means respectively.

5 Experiments

In this Section, we compare the proposed Spectral Embedded Clustering (SEC) with Spectral Clustering (SC) [Yu and Shi, 2003], CLGR [Wang *et al.*, 2007], K-means (KM) and Discriminative K-means(DKM) [Ye *et al.*, 2008]. We employ the spectral relaxation + spectral rotation to compute the assignment matrix for SEC, SC and CLGR. For KM and DKM, we still use the EM-like method to assign cluster labels as in [Ye *et al.*, 2008]. We also implement K-means and Discriminative K-means by using the spectral relaxation + spectral rotation for cluster assignment. As K-means and Discriminative K-means turn to the same when the spectral relaxation is used, we denote the results as KM-r in this work.

5.1 Experimental Setup

Eight data sets are used in the experiments, including two UCI data sets, Iris and Vote[3], one object data set, COIL-20,

[3]http://www.ics.uci.edu/ mlearn/MLRepository.html

Table 1: Dataset Description.

Dataset	Size	Dimensions	Classes
Iris	150	4	3
Vote	435	16	2
COIL-20	1440	1024	20
UMIST	575	644	20
AT&T	400	644	40
AR	840	768	120
YALE-B	2414	1024	38
CMU PIE	3329	1024	68

and five face data sets, UMIST, AT&T, AR, YALE-B and CMU PIE. Some data sets are resized, and Table 1 summarizes the details of the datasets used in the experiments.

SC and SEC need to determine the parameter σ in (1). In this work, we use the self-tune spectral clustering [Zelnik-Manor and Perona, 2004] method to determine the parameter σ. We also need to set the regularization parameters for SEC, CLGR and DKM beforehand. For fair comparison, we set the parameter γ in SEC and CLGR as 1, and set the parameter μ in SEC and CLGR, and the parameter γ in DKM as $\{10^{-10}, 10^{-7}, 10^{-4}, 10^{-1}, 10^{2}, 10^{5}, 10^{8}\}$. We report the best clustering result from the best parameter for SEC, CLGR and DKM.

The results of all clustering algorithms depend on the initialization (either EM-like or the spectral rotation). To reduce statistical variety, we independently repeat all clustering algorithms for 50 times with random initialization, and then we report the results corresponding to the best objective values.

5.2 Evaluation Metrics

We use the following two popular evaluation metrics to evaluate the performance for all the clustering algorithms.

Clustering Accuracy (ACC) is defined as:

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

where l_i is the true class label and c_i is the obtained cluster label of x_i, $\delta(x, y)$ is the delta function, and $map(\cdot)$ is the best mapping function. Note $\delta(x, y) = 1$, if $x = y$; $\delta(x, y) = 0$, otherwise. The mapping function $map(\cdot)$ matches the true class label and the obtained cluster label and the best mapping is solved by Kuhn-Munkres algorithm. A larger ACC indicates a better performance.

Normalized Mutual Information (NMI) is calculated by:

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

where C is a set of clusters obtained from the true labels and C' is a set of clusters obtained from the clustering algorithm. $MI(C, C')$ is the mutual information metric, and $H(C)$ and $H(C')$ are the entropies of C and C' respectively. See [Cai *et al.*, 2005] for more information. NMI is between 0 and 1. Again, a larger NMI value indicates a better performance.

5.3 Experimental Results

The clustering results from various algorithms are reported in Table 2 and Table 3. Moreover, the results of SEC with different μ and DKM with different γ are also shown in Figure 1. We have the following observations:

1) When the traditional EM-like technique is used in KM and DKM to assign cluster labels, DKM and KM lead to different results. In some data sets, DKM significantly outperforms KM. But DKM is slightly worse than KM in other data sets.

2) When EM-like and spectral relaxation + spectral rotation methods are used to solve the cluster assignment matrix for the same clustering algorithm (KM or DKM), there is no consistent winner on all the databases.

3) CLGR sightly outperforms SC in all the cases. SC and CLGR significantly outperform KM and DKM in some cases, but they are also significantly worse in other cases.

4) Our method SEC outperforms KM, DKM, KM-r, SC and CLGR in most cases. For the image datasets (such as AR and CMU PIE) with strong lighting variations, we observe significant improvement of SEC over SC and CLGR. Even for the dataset with clear manifold structure such as COIL-20 and UMIST, SEC is still better than SC and CLGR.

5) For low dimensional data sets (*e.g.*, Iris and Vote), SEC is slightly better than DKM with some range of parameters μ, and DKM slightly outperforms SEC with other range of parameters γ. However, for all high dimensional data sets, SEC outperforms DKM in most range of parameters μ in term of both ACC and NMI.

Table 2: Performance comparison of clustering accuracy from KM, DKM, KM-r, SC, CLGR and SEC on eight databases.

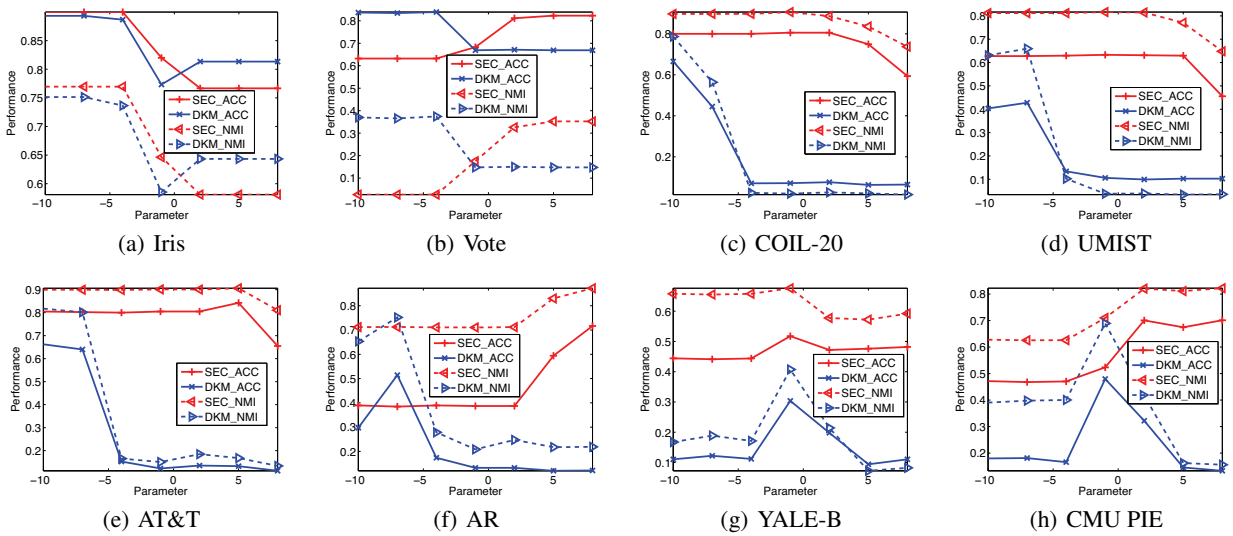
	KM	DKM	KM-r	SC	CLGR	SEC
Iris	89.3	89.3	76.0	74.6	78.0	**90.0**
Vote	83.6	**83.9**	78.8	66.9	68.3	82.3
COIL-20	69.5	66.6	58.2	72.5	79.8	**80.6**
UMIST	45.7	42.8	50.9	60.3	61.5	**63.3**
AT&T	60.8	66.2	68.7	74.7	77.5	**84.2**
AR	30.7	51.5	69.8	38.8	42.9	**71.6**
YALE-B	11.9	30.3	45.8	45.6	45.9	**51.8**
CMU PIE	17.5	47.9	65.7	46.2	51.9	**70.1**

Table 3: Performance comparison of normalized mutual information from KM, DKM, KM-r, SC, CLGR and SEC on eight databases.

	KM	DKM	KM-r	SC	CLGR	SEC
Iris	75.1	75.1	58.0	53.3	54.6	**77.0**
Vote	37.0	**37.4**	29.1	14.8	18.3	35.3
COIL-20	78.5	78.6	73.6	87.3	89.2	**90.7**
UMIST	65.4	66.0	67.6	80.5	81.2	**81.6**
AT&T	80.7	81.8	82.9	87.1	89.6	**90.4**
AR	66.3	75.2	86.5	71.0	71.8	**87.3**
YALE-B	17.9	40.8	57.2	66.5	66.6	**67.6**
CMU PIE	39.7	68.9	80.6	62.8	68.1	**82.1**

6 Conclusions

Observing that the cluster assignment matrix can always be represented by a low dimensional linear mapping of the high-dimensional data, we propose Spectral Embedded Clustering



(a) Iris (b) Vote (c) COIL-20 (d) UMIST

(e) AT&T (f) AR (g) YALE-B (h) CMU PIE

Figure 1: Clustering Performance of SEC with $\gamma = 1$ and different μ and DKM with different γ. The horizontal axis is shown in log space.

(SEC) to minimize the objective function of spectral clustering as well as control the mismatch between the cluster assignment matrix and the low dimensional representation of data. We also prove that spectral clustering, CLGR, K-means and Discriminative K-means are all the special cases of SEC in terms of the objective functions. The exhaustive experiments on eight data sets show that SEC generally outperforms the existing spectral clustering methods, K-means and Discriminative K-means.

References

[Ben-Hur *et al.*, 2001] A. Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. Support vector clustering. 2:125–137, 2001.

[Cai *et al.*, 2005] Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. *IEEE Trans. Knowl. Data Eng.*, 17(12):1624–1637, 2005.

[Ding and Li, 2007] Chris H. Q. Ding and Tao Li. Adaptive dimension reduction using discriminant analysis and -means clustering. In *ICML*, pages 521–528, 2007.

[Ding *et al.*, 2002] Chris H. Q. Ding, Xiaofeng He, Hongyuan Zha, and Horst D. Simon. Adaptive dimension reduction for clustering high dimensional data. In *ICDM*, pages 147–154, 2002.

[Jain and Dubes, 1988] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[la Torre and Kanade, 2006] Fernando De la Torre and Takeo Kanade. Discriminative cluster analysis. In *ICML*, pages 241–248, 2006.

[Li *et al.*, 2004] Tao Li, Sheng Ma, and Mitsunori Ogihara. Document clustering via adaptive subspace iteration. In *SIGIR*, pages 218–225, 2004.

[Li *et al.*, 2009] Y. Li, I.W. Tsang, J. T. Kwok, and Z. Zhou. Tighter and convex maximum margin clustering. In *AISTATS*, 2009.

[McLachlan and Peel, 2000] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.

[Ng *et al.*, 2001] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2001.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[Wang *et al.*, 2007] Fei Wang, Changshui Zhang, and Tao Li. Clustering with local and global regularization. In *AAAI*, pages 657–662, 2007.

[Wu and Schölkopf, 2007] M. Wu and B. Schölkopf. Transductive classification via local learning regularization. In *AISTATS*, pages 628–635, 03 2007.

[Xu *et al.*, 2005] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. Cambridge, MA, 2005. MIT Press.

[Ye *et al.*, 2007] Jieping Ye, Zheng Zhao, and Huan Liu. Adaptive distance metric learning for clustering. In *CVPR*, 2007.

[Ye *et al.*, 2008] Jieping Ye, Zheng Zhao, and Mingrui Wu. Discriminative k-means for clustering. In *Advances in Neural Information Processing Systems 20*, pages 1649–1656. 2008.

[Ye, 2005] Jieping Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6:483–502, 2005.

[Ye, 2007] Jieping Ye. Least squares linear discriminant analysis. In *ICML*, pages 1087–1093, 2007.

[Yu and Shi, 2003] Stella X. Yu and Jianbo Shi. Multiclass spectral clustering. In *ICCV*, pages 313–319, 2003.

[Zelnik-Manor and Perona, 2004] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *NIPS*, 2004.

[Zha *et al.*, 2001] Hongyuan Zha, Xiaofeng He, Chris H. Q. Ding, Ming Gu, and Horst D. Simon. Spectral relaxation for k-means clustering. In *NIPS*, pages 1057–1064, 2001.

[Zhang *et al.*, 2007] K. Zhang, I.W. Tsang, and J.T. Kwok. Maximum margin clustering made practical. In *ICML*, Corvallis, Oregon, USA, June 2007.