A Convex Formulation for Spectral Shrunk Clustering

Xiaojun Chang[1], Feiping Nie[2,3], Zhigang Ma[4], Yi Yang[1] and Xiaofang Zhou[5]

[1]Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Australia.
[2]Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Shaanxi, China.
[3]Department of Computer Science and Engineering, University of Texas at Arlington, USA.
[4]School of Computer Science, Carnegie Mellon University, USA.
[5]School of Information Technology & Electrical Engineering, The University of Queensland, Australia.
cxj273@gmail.com,feipingnie@gmail.com,kevinma@cs.cmu.edu,yiyang@cs.cmu.edu,zxf@itee.uq.edu.au.

Abstract

Spectral clustering is a fundamental technique in the field of data mining and information processing. Most existing spectral clustering algorithms integrate dimensionality reduction into the clustering process assisted by manifold learning in the original space. However, the manifold in reduced-dimensional subspace is likely to exhibit altered properties in contrast with the original space. Thus, applying manifold information obtained from the original space to the clustering process in a low-dimensional subspace is prone to inferior performance. Aiming to address this issue, we propose a novel convex algorithm that mines the manifold structure in the low-dimensional subspace. In addition, our unified learning process makes the manifold learning particularly tailored for the clustering. Compared with other related methods, the proposed algorithm results in more structured clustering result. To validate the efficacy of the proposed algorithm, we perform extensive experiments on several benchmark datasets in comparison with some state-of-the-art clustering approaches. The experimental results demonstrate that the proposed algorithm has quite promising clustering performance.

Introduction

Clustering has been widely used in many real-world applications (Jain and Dubes 1988; Wang, Nie, and Huang 2014). The objective of clustering is to cluster the original data points into various clusters, so that data points within the same cluster are dense while those in different clusters are far away from each other (Filippone et al. 2008). Researchers have proposed a variety of clustering algorithms, such as K-means clustering and mixture models (Wang et al. 2014; Nie, Wang, and Huang 2014; Nie et al. 2011b), *etc*.

The existing clustering algorithms, however, mostly work well when the samples' dimensionality is low. When partitioning high-dimensional data, the performance of these algorithms is not guaranteed. For example, K-means clustering iteratively assigns each data point to the cluster

with the closest center based on specific distance/similarity measurement and updates the center of each cluster. But the distance/similarity measurements may be inaccurate on high-dimensional data, which tends to limit the clustering performance. As suggested by some researchers, many high-dimensional data may exhibit dense grouping in a low-dimensional subspace (Nie et al. 2009). Hence, researchers have proposed to first project the original data into a low-dimensional subspace via some dimensionality reduction techniques and then cluster the computed low-dimensional embedding for high-dimensional data clustering. For instance, a popular approach is to use Principle component analysis (PCA) to reduce the dimensionality of the original data followed by Kmeans for clustering (PcaKm) (Xu and Wunsch 2005). Ding et al. present a clustering algorithm based on Linear discriminant analysis (LDA) method (Ding and Li 2007). Ye et al propose discriminative K-means (DisKmeans) clustering which unifies the iterative procedure of dimensionality reduction and K-means clustering into a trace maximization problem (Ye, Zhao, and Wu 2007).

Another genre of clustering, *i.e.*, spectral clustering (Shi and Malik 2000) integrates dimensionality reduction into its clustering process. The basic idea of spectral clustering is to find a clustering assignment of the data points by adopting the spectrum of similarity matrix that leverages the nonlinear manifold structure of original data. Spectral clustering has been shown to be easy to implement and oftentimes it outperforms traditional clustering methods because of its capacity of mining intrinsic geometric structures, which facilitates partitioning data with more complicated structures. The benefit of utilizing manifold information has been demonstrated in many applications, such as image segmentation and web mining. Due to the advantage of spectral clustering, different variants of spectral clustering algorithms have been proposed these years (Li et al. 2015). For example, local learning-based clustering (LLC) (Wu and Schlkopf 2006) utilizes a kernel regression model for label prediction based on the assumption that the class label of a data point can be determined by its neighbors. Self-tuning SC (Zelnik-Manor and Perona 2004) is able to tune parameters automatically in an unsupervised scenario. Normalized cuts is capable of balancing the volume of clusters for the usage of data density information

(Shi and Malik 2000).

Spectral clustering is essentially a two-stage approach, *i.e.*, manifold learning based in the original high-dimensional space and dimensionality reduction. To achieve proper clustering, spectral clustering assumes that two nearby data points in the high density region of the reduced-dimensional space have the same cluster label. However, this assumption does not always hold. More possibly, these nearest neighbors may be far away from each other in the original high-dimensional space due to the curse of dimensionality. That being said, the distance measurement of the original data could not precisely reflect the low-dimensional manifold structure, thus leading to suboptimal clustering performance.

Intuitively, if the manifold structure in the low-dimensional space is precisely captured, the clustering performance could be enhanced when applied to high-dimensional data clustering. Aiming to achieve this goal, we propose a novel clustering algorithm that is able to mine the inherent manifold structure of the low-dimensional space for clustering. Moreover, compared to traditional spectral clustering algorithms, the shrunk pattern learned by the proposed algorithm does not have an orthogonal constraint, giving it more flexibility to fit the manifold structure. It is worthwhile to highlight the following merits of our work:

- The proposed algorithm is more capable of uncovering the manifold structure. Particularly, the shrunk pattern does not have the orthogonal constraint, making it more flexible to fit the manifold structure.

- The integration of manifold learning and clustering makes the former particularly tailored for the latter. This is intrinsically different from most state-of-the-art clustering algorithms.

- The proposed algorithm is convex and converges to global optimum, which indicates that the proposed algorithm does not rely on the initialization.

The rest of this paper is organized as follows. After reviewing related work on spectral clustering in section 2, we detail the proposed algorithm in section 3. Extensive experimental results are given in section 4 and section 5 concludes this paper.

Related Work

Our work is inspired by spectral clustering. Therefore, we review the related work on spectral clustering in this section.

Basics of Spectral Clustering

To facilitate the presentation, we first summarize the notations that will be frequently used in this paper. Given a dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^d (1 \leq i \leq n)$ is the i-th datum and n is the total number of data points. The objective of clustering is to partition χ into c clusters. Denote the cluster assignment matrix by $Y = \{y_1, \ldots, y_n\} \in \mathbb{R}^{n \times c}$, 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 is clustered to the j-th cluster, and 0 otherwise.

Existing spectral clustering algorithms adopt a weighted graph to partition the data. Let us denote $\mathcal{G} = \{\mathcal{X}, A\}$ as a weighted graph with a vertex set $\mathcal{X}$ and an affinity matrix $A \in \mathbb{R}^{n \times n}$. A_{ij} is the affinity of a pair of vertexes of the weighted graph. A_{ij} is commonly defined as:

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

where δ is the parameter to control the spread of neighbors. The Laplacian matrix L is computed according to $L = D - A$, where D is a diagonal matrix with the diagonal elements as $D_{ii} = \sum_j A_{ij}, \forall i$. Following the work in (Ye, Zhao, and Wu 2007), we denote the scaled cluster indicator matrix F as follows:

$$F = [F_1, F_2, \ldots, F_n]^T = Y(Y^T Y)^{-\frac{1}{2}}, \quad (1)$$

where F_i is the scaled cluster indicator of x_i. The j-th column of F is defined as follows by (Ye, Zhao, and Wu 2007):

$$f_j = \left[\underbrace{0, \ldots, 0}_{\sum_{i=1}^{j-1} n_i}, \underbrace{\frac{1}{\sqrt{n_j}}, \ldots, \frac{1}{\sqrt{n_j}}}_{n_j}, \underbrace{0, \ldots, 0}_{\sum_{i=j+1}^{c} n_k} \right], \quad (2)$$

which indicates which data points are partitioned into the j-th cluster C_j. n_j is the number of data points in cluster C_j.

The objective function of spectral clustering algorithm is generally formulated as follows:

$$\min_F Tr(F^T L F) \\ s.t. F = Y(Y^T Y)^{-\frac{1}{2}} \quad (3)$$

where $Tr(\cdot)$ denotes the trace operator. By denoting I as an identity matrix, we can define the normalized Laplacian matrix L_n as:

$$L_n = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}. \quad (4)$$

By replacing L in Eq. (3) with the normalized Laplacian matrix, the objective function becomes the well-known SC algorithm normalized cut (Shi and Malik 2000). In the same manner, if we replace L in Eq. (3) by the Laplacian matrix obtained by local learning (Yang et al. 2010)(Wu and Schlkopf 2006), the objective function converts to Local Learning Clustering (LLC).

Progress on Spectral Clustering

Being easy to implement and promising for many applications, spectral clustering has been widely studied for different problems. Chen et al. propose a Landmark-based Spectral Clustering (LSC) for large scale clustering problems (Chen and Cai 2011). Specifically, a few representative data points are first selected as the landmarks and the original data points are then represented as the linear combinations of these landmarks. The spectral clustering is performed on the landmark-based representation.

Yang et al. propose to utilize a nonnegative constraint to relax the elements of cluster indicator matrix for spectral clustering (Yang et al. 2011). Liu et al. propose to compress the original graph used for spectral clustering into a sparse bipartite graph. The clustering is then performed on the bipartite graph instead, which improved the efficiency for large-scale data (Liu et al. 2013). Xia et al. propose a multi-view spectral clustering method based on low-rank and sparse decomposition (Xia et al. 2014). Yang et al. propose to use Laplacian Regularized L1-Graph for clustering (Yang et al. 2014). Tian et al. recently propose to adopt deep learning in spectral clustering (Tian et al. 2014).

In spite of the encouraging progress, few of the existing spectral clustering methods have considered learn the manifold in the low-dimensional subspace more precisely, not to mention integrating such manifold learning and clustering into a unified framework. This issue shall be addressed in this paper for boosted clustering performance.

The Proposed Algorithm

In this section, we present the details of the proposed algorithm. A fast iterative method is also proposed to solve the objective function.

Problem Formulation

Our algorithms is built atop the aim of uncovering the utmost manifold structure in the low-dimensional subspace of original data. Inspired by (Hou et al. 2013), we adopt the pattern shrinking during the manifold learning and the shrunk patterns are exploited for clustering simultaneously.

To begin with, we have the following notations. Denote the shrunk patterns of n data samples as $\{g_1, \cdots, g_n\}$, where $g_i \in \mathbb{R}^c$. We first obtain spectral embedding F of the original samples by minimizing the traditional spectral clustering algorithm $\min Tr(F^T L_n F)$, where L_n is a normalized Laplacian matrix.

Next, the shrunk patterns are computed by satisfying the following requirements. (1) The shrunk patterns should keep consistency with the spectral embedding. To be more specific, the shrunk patterns should not be far away from the spectral clustering. (2) Note that nearby points are more likely to belong to the same cluster. We thus design a similarity matrix to measure pair similarity of any two spectral embedding, which the shrunk patters should follow.

To characterize the manifold structure of the spectral embedding $\{f_1, \cdots, f_n\}$, a k-nearest neighbor graph is constructed by connecting each point to its k nearest neighbors. The similarity matrix, W, is computed by $W_{ij} = exp(-\frac{\|f_i - f_j\|^2}{\delta^2})$.

From this similarity matrix, we can observe that if two spectral embeddings are nearby, they should belong to the same cluster and the corresponding weight should be large, which satisfies the first requirement (Nie et al. 2011a).

To keep the local similarity of spectral embedding, we propose to optimize the following objective function.

$$\min_G \sum_{ij} W_{ij} \|g_i - g_j\|_2 \qquad (5)$$

We also aim to keep the consistency between spectral embedding and shrunk patterns. Hence, we propose to minimize the following loss function directly.

$$\min_G \|G - F\|_2^2 \qquad (6)$$

To this end, we formulate the objection function as follows:

$$\min_G \|G - F\|_2^2 + \gamma \sum_{i,j} W_{ij} \|g_i - g_j\|_2 \qquad (7)$$

where γ is a balance parameter.

It can be easily proved that our formulation is convex. Due to the space limit, we omit the proof here. Since our method exploits shrunk patterns as the input for clustering, we name it Spectral Shrunk Clustering (SSC).

As indicated in (Ma et al. 2012; Kong, Ding, and Huang 2011), the least square loss function is not robust to outliers. To make our method even more effective, we follow (Ma et al. 2012; Nie et al. 2010) and employ $l_{2,1}$-norm to handle the outliers. The objective function is rewritten as follows:

$$\min_G \|G - F\|_{2,1} + \gamma \sum_{i,j} W_{ij} \|g_i - g_j\|_2 \qquad (8)$$

Optimization

The proposed function involves the $l_{2,1}$-norm, which is difficult to solve in a closed form. We propose to solve this problem in the following steps. Denote $H = G - F$ and $H = [h^1, \cdots, h^d]$, where d is the dimension of spectral embedding. The objective function can be rewritten as follows:

$$\min_G Tr((G - F)^T S(G - F)) + \gamma \sum_{ij} w_{ij} \|g_i - g_j\|_2 \qquad (9)$$

where

$$S = \begin{bmatrix} \frac{1}{2\|h^1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|h^d\|_2} \end{bmatrix}. \qquad (10)$$

Denote a Laplacian matrix $\widetilde{L} = \widetilde{D} - \widetilde{W}$, where $\widetilde{W}$ is a re-weighted weight matrix defined by

$$\widetilde{W}_{ij} = \frac{W_{ij}}{2\|g_i - g_j\|_2} \qquad (11)$$

$\widetilde{D}$ is a diagonal matrix with the i-th diagonal element as $\sum_j \widetilde{W}_{ij}$.

By simple mathematical deduction, the objective function arrives at:

$$\min_G Tr((G - F)^T S(G - F)) + \gamma Tr(G^T \widetilde{L} G). \qquad (12)$$

By setting the derivative of Eq. (12) to G to 0, we have:

$$G = (S + \gamma \widetilde{L})^{-1} SF. \qquad (13)$$

Based on the above mathematical deduction, we propose an iterative algorithm to optimize the objective function in Eq. (3), which is summarized in Algorithm 1. Once the shrunk patterns G are obtained, we perform K-means clustering on it to get the final clustering result.

Algorithm 1: Optimization Algorithm for SSC

Data: Data $X \in \mathbb{R}^{d \times n}$, Parameter γ and the number of clusters c
Result:
  The discrete cluster assignment $Y \in \mathbb{R}^{n \times c}$
1   Compute the normalized Laplacian matrix L_n ;
2   Obtain the spectral embedding F by using the traditional spectral clustering ;
3   Compute the similarity matrix W using the spectral embedding F ;
4   Obtain the Laplacian matrix with the reweighted weight matrix according to Eq. (11) ;
5   Set $t = 0$;
6   Initialize $G_0 \in \mathbb{R}^{n \times c}$;
7   **repeat**
8   Compute $H_t = G_t - F$;
9   Compute the diagonal matrix S_t according to (10) ;
10   Compute G_{t+1} according to
  $G_{t+1} = (S_t + \gamma \widetilde{W})^{-1} S_t X$;
11   t = t + 1 ;
12   **until** *Convergence*;
13   Based on G^*, compute the discrete cluster assignment matrix Y by using K-means clustering;
14   Return the discrete cluster assignment matrix Y.

Convergence Analysis

To prove the convergence of the Algorithm 1, we need the following lemma (Nie et al. 2010).

Lemma 1. *For any nonzero vectors* $g, g_t \in \mathbb{R}^c$*, the following inequality holds:*

$$\|g\|_2 - \|g\|_2^2/2\|g_t\|_2 \leq \|g_t\|_2 - \|g_t\|_2^2/2\|g_t\|_2 \quad (14)$$

The following theorem guarantees that the problem in Eq. (8) converges to the global optimum by Algorithm 1 .

Theorem 1. *The Algorithm 1 monotonically decreases the objective function value of the problem in Eq.* (8) *in each iteration, thus making it converge to the global optimum.*

Proof. Define $f(G) = Tr((G - F)^T S(G - F))$. According to Algorithm 1, we know that

$$G_{t+1} = \arg\min_G f(G) + \gamma \sum_{i,j}(\widetilde{W})_{ij}\|g_i - g_j\|_2^2 \quad (15)$$

Note that $(\widetilde{W}_t)_{ij} = \frac{W_{ij}}{2\|g_i^t - g_j^t\|_2}$, so we have

$$f(G_{t+1}) + \gamma \sum_{ij} \frac{W_{ij}\|g_i^{t+1} - g_j^{t+1}\|_2^2}{2\|g_i^t - g_j^t\|_2}$$
$$\leq f(G_t) + \gamma \sum_{ij} \frac{W_{ij}\|g_i^t - g_j^t\|_2^2}{2\|g_i^t - g_j^t\|_2} \quad (16)$$

According to Lemma 1, we have

$$\sum_{ij} W_{ij}(\|g_i^{t+1} - g_j^{t+1}\|_2 - \frac{\|g_i^{t+1} - g_j^{t+1}\|_2^2}{2\|g_i^t - g_j^t\|_2})$$
$$\leq \sum_{ij} W_{ij}(\|g_i^t - g_j^t\|_2 - \frac{\|g_i^t - g_j^t\|_2^2}{2\|g_i^t - g_j^t\|_2}) \quad (17)$$

By summing Eq. (16) and Eq. (17), we arrive at:

$$f(G_{t+1}) + \gamma \sum_{ij} W_{ij}\|g_i^{t+1} - g_j^{t+1}\|_2$$
$$\leq f(G_t) + \gamma \sum_{ij} W_{ij}\|g_i^t - g_j^t\|_2 \quad (18)$$

Thus, Algorithm 1 monotonically decreases the objective function value of the problem in Eq. (8) in each iteration t. When converged, G_t and $\widetilde{L}_t$ satisfy Eq. (13). As the problem in Eq. (8) is convex, satisfying Eq. (13) indicates that G_t is the global optimum solution of the problem in Eq. (8). Therefore, using Algorithm 1 makes the problem in Eq. (8) converge to the global optimum.   □

Experiment

In this section, we perform extensive experiments on a variety of applications to test the performance of our method SSC. We compare SSC to several clustering algorithms including the classical K-means, the classical spectral clustering (SC), PCA_Kmeans (Xu and Wunsch 2005), PCA spectral clustering (PCA_SC), LDA_Kmeans (Ding and Li 2007), LDA spectral clustering (LDA_SC), Local Learning Clustering (LLC) (Wu and Schlkopf 2006) and SPLS (Hou et al. 2013).

Datasets

A variety of datasets are used in our experiments which are described as follows. The AR dataset (Martinez and Benavente 1998) contains 840 faces of 120 different people. We utilize the pixel value as the feature representations. The JAFFE dataset (Lyons et al. 1997) consists of 213 images of different facial expressions from 10 different Japanese female models. The images are resized to 26×26 and represented by pixel values. The ORL dataset (Samaria and Harter 1994) consists of 40 different subjects with 10 images each. We also resize each image to 32×32 and use pixel values to represent the images. The UMIST face dataset (Graham and M 1998) consists of 564 images of 20 individuals with mixed race, gender and appearance. Each individual is shown in a range of poses from profile

to frontal views. The pixel value is used as the feature representation. The BinAlpha dataset contains 26 binary hand-written alphabets and we randomly select 30 images for every alphabet. The MSRA50 dataset contains 1799 images from 12 different classes. We resize each image to 32×32 and use the pixel values as the features. The YaleB dataset (Georghiades, Belhumeur, and Kriegman 2001) contains 2414 near frontal images from 38 persons under different illuminations. Each image is resized to 32×32 and the pixel value is used as feature representation. We additionally use the USPS dataset to validate the performance on handwritten digit recognition. The dataset consists of 9298 gray-scale handwritten digit images. We resize the images to 16×16 and use pixel values as the features.

Setup

The size of neighborhood, k is set to 5 for all the spectral clustering algorithms. For parameters in all the comparison algorithms, we tune them in the range of $\{10^{-6}, 10^{-3}, 10^0, 10^3, 10^6\}$ and report the best results. Note that the results of all the clustering algorithms vary on different initialization. To reduce the influence of statistical variation, we repeat each clustering 50 times with random initialization and report the results corresponding to the best objective function values. For all the dimensionality reduction based K-means and Spectral clustering, we project the original data into a low dimensional subspace of 10 to 150 and report the best results.

Evaluation Metrics

Following most work on clustering, we use clustering accuracy (ACC) and normalized mutual information (NMI) as our evaluation metrics in our experiments.

Let q_i represent the clustering label result from a clustering algorithm and p_i represent the corresponding ground truth label of an arbitrary data point x_i. Then ACC is defined as follows:

$$ACC = \frac{\sum_{i=1}^n \delta(p_i, map(q_i))}{n}, \quad (19)$$

where $\delta(x, y) = 1$ if $x = y$ and $\delta(x, y) = 0$ otherwise. $map(q_i)$ 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 clustering performance.

For any two arbitrary variables P and Q, NMI is defined as follows (Strehl and Ghosh 2003):

$$NMI = \frac{I(P, Q)}{\sqrt{H(P)H(Q)}}, \quad (20)$$

where $I(P, Q)$ computes the mutual information between P and Q, and $H(P)$ and $H(Q)$ are the entropies of P and Q. Let t_l represent the number of data in the cluster $\mathcal{C}_l (1 \leq l \leq c)$ generated by a clustering algorithm and $\widetilde{t_h}$ represent the number of data points from the h-th ground truth class. NMI metric is then computed as follows (Strehl and Ghosh 2003):

$$NMI = \frac{\sum_{l=1}^c \sum_{h=1}^c t_{l,h} log(\frac{n \times t_{l,h}}{t_l \widetilde{t_h}})}{\sqrt{(\sum_{l=1}^c t_l \log \frac{t_l}{n})(\sum_{h=1}^c \widetilde{t_h} \log \frac{\widetilde{t_h}}{n})}}, \quad (21)$$

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

Experimental Results

The experimental results on listed in Table 1 and Table 2. We can see from the two tables that our method is consistently the best algorithm using both evaluation metrics. We also observe that:

1. The spectral clustering algorithm and its variants achieve better performance than the classical k-means and its variants. This observation suggests that it is beneficial to utilize the pairwise similarities between all data points from a weighted graph adjacency matrix that contains helpful information for clustering.

2. PCA_Kmeans and LDA_Kmeans are better than K-means whereas PCA_SC and LDA_SC are better than SC. This demonstrates that dimensionality reduction is helpful for improving the cluster performance.

3. LDA_Kmeans outperforms PCA_Kmeans while LDA_SC outperforms PCA_SC. This indicates that LDA is more capable of keeping the structural information than PCA when doing dimensionality reduction.

4. Among various spectral clustering variants, LLC is the most robust algorithm. This means using a more sophisticated graph Laplacian is beneficial for better exploitation of manifold structure.

5. SPLS is the second best clustering algorithm. This is because it incorporates both the linear and nonlinear structures of original data.

6. Our proposed Spectral Shrunk Clustering (SSC) consistently outperforms the other K-means based and spectral clustering based algorithms. This advantage is attributed to the optimal manifold learning in the low-dimensional subspace and it being tightly coupled with the clustering optimization.

Parameter Sensitivity

In this section, we study the sensitivity of our algorithm *w.r.t.* the parameter γ in Eq. (3). Fig 1 shows the accuracy (y-axis) of SSC for different γ values (x-axis) on all the experimental datasets. It can be seen from the figure that the performance varies when different values of γ are used. However, except on MSRA50 and USPS datasets, our method attains the best/respectable performance when $\gamma = 1$. This indicates that our method has a consistent preference on parameter setting, which makes it uncomplicated to get optimal parameter value in practice.

Table 1: Performance comparison (ACC%±Standard Deviation) between K-means, Spectral Clustering, PCA_Kmeans, LDA_Kmeans, PCA_SC, LDA_SC, LLC, SPLS and SSC.

	AR	JAFFE	ORL	UMIST	binalpha	MSRA50	YaleB	USPS
K-means	36.3 ± 1.4	75.6 ± 1.8	60.5 ± 1.8	41.3 ± 1.6	41.7 ± 1.1	46.2 ± 1.7	14.4 ± 1.5	65.4 ± 1.7
SC	41.6 ± 2.1	76.1 ± 1.6	72.7 ± 2.3	52.2 ± 1.4	43.6 ± 1.5	52.3 ± 1.8	34.8 ± 1.4	64.3 ± 1.4
PCA_Kmeans	39.8 ± 1.8	75.8 ± 1.5	64.5 ± 2.3	48.8 ± 1.7	42.4 ± 1.5	56.0 ± 1.9	24.9 ± 1.8	69.4 ± 1.8
PCA_SC	43.2 ± 1.7	76.9 ± 1.8	67.8 ± 2.1	54.1 ± 1.9	44.3 ± 1.8	54.9 ± 1.6	36.8 ± 2.0	69.1 ± 1.5
LDA_Kmeans	40.4 ± 1.5	76.5 ± 1.7	65.6 ± 2.6	49.7 ± 1.8	42.9 ± 1.7	56.4 ± 1.8	26.1 ± 1.8	70.1 ± 1.3
LDA_SC	44.5 ± 1.3	77.4 ± 1.9	68.3 ± 2.4	54.7 ± 1.5	45.1 ± 1.4	55.1 ± 1.7	38.2 ± 1.6	70.4 ± 1.5
LLC	48.7 ± 1.6	78.6 ± 1.5	71.5 ± 2.2	63.3 ± 1.8	40.7 ± 1.8	48.1 ± 1.4	38.2 ± 1.5	63.9 ± 1.7
SPLS	49.2 ± 1.4	79.5 ± 2.1	74.2 ± 1.8	70.4 ± 1.6	48.5 ± 1.9	60.3 ± 1.3	47.3 ± 1.7	71.4 ± 1.6
SSC	$\mathbf{51.3 \pm 1.5}$	$\mathbf{81.2 \pm 1.6}$	$\mathbf{76.0 \pm 1.6}$	$\mathbf{71.1 \pm 1.8}$	$\mathbf{49.4 \pm 1.3}$	$\mathbf{63.2 \pm 1.1}$	$\mathbf{49.8 \pm 1.6}$	$\mathbf{75.5 \pm 1.9}$

Table 2: Performance Comparison (NMI%±Standard Deviation) between K-means, Spectral Clustering, PCA_Kmeans, LDA_Kmeans, PCA_SC, LDA_SC, LLC, SPLS and SSC.

	AR	JAFFE	ORL	UMIST	binalpha	MSRA50	YaleB	USPS
K-means	68.7 ± 3.0	79.4 ± 0.8	80.3 ± 1.8	64.4 ± 1.5	58.6 ± 1.4	56.7 ± 1.8	17.3 ± 1.5	67.3 ± 1.8
SC	71.3 ± 2.6	80.2 ± 0.9	85.8 ± 1.9	72.1 ± 1.7	59.7 ± 1.6	70.0 ± 1.6	55.6 ± 1.6	69.5 ± 1.6
PCA_Kmeans	69.4 ± 2.8	79.8 ± 0.8	80.6 ± 1.6	68.2 ± 1.8	59.1 ± 1.8	60.3 ± 1.5	26.7 ± 1.8	73.1 ± 1.9
PCA_SC	70.3 ± 2.4	81.5 ± 1.3	86.3 ± 1.4	72.9 ± 1.5	60.6 ± 1.9	72.4 ± 1.8	38.6 ± 1.5	74.2 ± 1.8
LDA_Kmeans	69.9 ± 1.9	82.1 ± 1.4	81.1 ± 2.1	68.8 ± 1.5	59.8 ± 1.6	61.1 ± 1.9	29.4 ± 1.6	75.1 ± 1.6
LDA_SC	70.8 ± 1.5	81.9 ± 0.9	86.8 ± 1.7	74.1 ± 2.0	61.3 ± 1.7	73.2 ± 1.6	39.9 ± 1.4	75.4 ± 1.7
LLC	71.2 ± 2.4	82.5 ± 1.7	84.9 ± 1.5	77.3 ± 1.8	61.4 ± 1.9	66.2 ± 1.6	34.1 ± 1.3	67.5 ± 1.5
SPLS	72.4 ± 1.7	83.1 ± 2.1	87.2 ± 1.8	82.2 ± 1.6	63.6 ± 1.8	69.6 ± 1.8	41.4 ± 1.6	76.5 ± 1.8
SSC	$\mathbf{73.2 \pm 1.4}$	$\mathbf{84.3 \pm 1.6}$	$\mathbf{88.6 \pm 1.5}$	$\mathbf{84.1 \pm 1.5}$	$\mathbf{64.1 \pm 1.4}$	$\mathbf{72.2 \pm 1.4}$	$\mathbf{46.8 \pm 1.3}$	$\mathbf{79.8 \pm 1.6}$

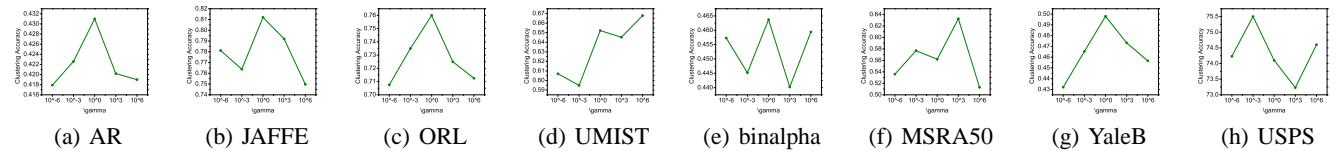(a) AR (b) JAFFE (c) ORL (d) UMIST (e) binalpha (f) MSRA50 (g) YaleB (h) USPS

Figure 1: The clustering performance (ACC) variation of our algorithm *w.r.t.* different parameter settings. From the experimental results, we observe that the proposed algorithm has a consistent preference on parameter setting, which makes it uncomplicated to get optimal parameter value in practice.

Convergence Study

As mentioned before, the proposed iterative approach in Algorithm 1 monotonically decreases the objective function value in Eq. (3). In this experiment, we show the convergence curves of the iterative approach on different datasets in Figure 2. The parameter γ is fixed at 1, which is the median value of the tuned range of the parameters.

It can be observed that the objective function value converges quickly. The convergence experiment demonstrates the efficiency of our algorithm.

Conclusion

In this paper, we have proposed a novel convex formulation of spectral shrunk clustering. The advantage of our method is three-fold. First, it is able to learn the manifold structure in the low-dimensional subspace rather than the original space. This feature contributes to more precise structural information for clustering based on the low-dimensional space. Second, our method is more capable of uncovering the manifold structure. Particularly, the shrunk pattern learned by the proposed algorithm does not have the orthogonal constraint, which makes it more flexible to fit the manifold structure. The learned manifold knowledge is particularly helpful for achieving better clustering result. Third, our algorithm is convex, which makes it easy to implement and very suitable for real-world applications. Extensive experiments on a variety of applications are given to show the effectiveness of the proposed algorithm. By comparing it to several state-of-the-art clustering approaches, we validate the advantage of our method.

References

[Chen and Cai 2011] Chen, X., and Cai, D. 2011. Large scale spectral clustering with landmark-based representation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*.

[Ding and Li 2007] Ding, C. H. Q., and Li, T. 2007. Adaptive dimension reduction using discriminant analysis and *K*-means clustering. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, 521–528.

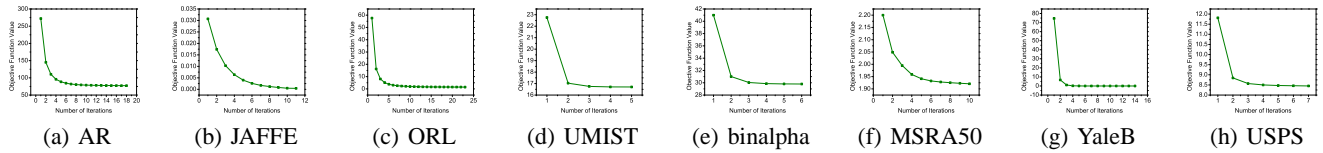| (a) AR | (b) JAFFE | (c) ORL | (d) UMIST | (e) binalpha | (f) MSRA50 | (g) YaleB | (h) USPS |

Figure 2: The convergence curves of our algorithm on different datasets. From the figures, we can observe that the objective function converges quickly, which demonstrates the efficiency of the proposed algorithm.

[Filippone et al. 2008] Filippone, M.; Camastra, F.; Masulli, F.; and Rovetta, S. 2008. A survey of kernel and spectral methods for clustering. *Pattern Recognition* 41(1):176–190.

[Georghiades, Belhumeur, and Kriegman 2001] Georghiades, A. S.; Belhumeur, P. N.; and Kriegman, D. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. PAMI* 23(6):643–660.

[Graham and M 1998] Graham, D. B., and M, A. N. 1998. Characterizing virtual eigensignatures for general purpose face recognition. *NATO ASI Series F, Computer and Systems Sciences* 446456.

[Hou et al. 2013] Hou, C.; Nie, F.; Jiao, Y.; Zhang, C.; and Wu, Y. 2013. Learning a subspace for clustering via pattern shrinking. *Inf. Process. Manage.* 49(4):871–883.

[Jain and Dubes 1988] Jain, A. K., and Dubes, R. C. 1988. Algorithms for clustering data. *Prentice-Hall, Inc.*

[Kong, Ding, and Huang 2011] Kong, D.; Ding, C.; and Huang, H. 2011. Robust nonnegative matrix factorization using l21-norm. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, 673–682.

[Li et al. 2015] Li, Y.; Nie, F.; Huang, H.; and Huang, J. 2015. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25 -29, 2015, Austin Texas, USA.*

[Liu et al. 2013] Liu, J.; Wang, C.; Danilevsky, M.; and Han, J. 2013. Large-scale spectral clustering on graphs. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013.*

[Lyons et al. 1997] Lyons, M. J.; Akamatsu, S.; Kamachi, M.; Gyoba, J.; and Budynek, J. 1997. The japanese female facial expression (jaffe) database. In *Database of digital images*.

[Ma et al. 2012] Ma, Z.; Nie, F.; Yang, Y.; Uijlings, J. R. R.; and Sebe, N. 2012. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Trans. Multimedia* 14(4):1021–1030.

[Martinez and Benavente 1998] Martinez, A. M., and Benavente, R. 1998. The ar face database. Technical report, Centre Vis. Comput., Univ. Autonoma Barcelona, Barcelona, Spain.

[Nie et al. 2009] Nie, F.; Xu, D.; Tsang, I. W.; and Zhang, C. 2009. Spectral embedded clustering. In *Proc. IJCAI*, 1181–1186.

[Nie et al. 2010] Nie, F.; Huang, H.; Cai, X.; and Ding, C. 2010. Efficient and robust feature selection via joint l21-norms minimization. In *Proc. NIPS*, 1813–1821.

[Nie et al. 2011a] Nie, F.; Wang, H.; Huang, H.; and Ding, C. H. Q. 2011a. Unsupervised and semi-supervised learning via l1-norm graph. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, 2268–2273.

[Nie et al. 2011b] Nie, F.; Zeng, Z.; Tsang, I. W.; Xu, D.; and Zhang, C. 2011b. Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. *IEEE Transactions on Neural Networks* 22(11):1796–1808.

[Nie, Wang, and Huang 2014] Nie, F.; Wang, X.; and Huang, H. 2014. Clustering and projected clustering with adaptive neighbors. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, 977–986.

[Samaria and Harter 1994] Samaria, F. S., and Harter, A. C. 1994. Parameterisation of a stochastic model for human face identification. In *Proc. Applications of Computer Vision*, 138–142.

[Shi and Malik 2000] Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence* 22(8):888–905.

[Strehl and Ghosh 2003] Strehl, A., and Ghosh, J. 2003. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Machine Learning Research* 3:583–617.

[Tian et al. 2014] Tian, F.; Gao, B.; Cui, Q.; Chen, E.; and Liu, T. 2014. Learning deep representations for graph clustering. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, 1293–1299.

[Wang et al. 2014] Wang, D.; Wang, Y.; Nie, F.; Yan, J.; Cai, T. W.; Saykin, A. J.; Shen, L.; and Huang, H. 2014. Human connectome module pattern detection using a new multi-graph minmax cut model. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2014 - 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part III*, 313–320.

[Wang, Nie, and Huang 2014] Wang, D.; Nie, F.; and Huang, H. 2014. Unsupervised feature selection via unified trace ratio formulation and k-means clustering (TRACK). In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part III*, 306–321.

[Wu and Schlkopf 2006] Wu, M., and Schlkopf, B. 2006. A local learning approach for clustering. In *Proc. NIPS*, 1529–1536.

[Xia et al. 2014] Xia, R.; Pan, Y.; Du, L.; and Yin, J. 2014. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, 2149–2155.

[Xu and Wunsch 2005] Xu, R., and Wunsch, D. 2005. Survey of clustering algorithms. *IEEE Trans. Neural Networks* 16(3):645–678.

[Yang et al. 2010] Yang, Y.; Xu, D.; Nie, F.; Yan, S.; and Zhuang, Y. 2010. Image clustering using local discriminant models and global integration. *IEEE Trans. Image Process.* 19(10):2761–2773.

[Yang et al. 2011] Yang, Y.; Shen, H. T.; Nie, F.; Ji, R.; and Zhou, X. 2011. Nonnegative spectral clustering with discriminative regularization. In *Proceedings of the Twenty-Fifth AAAI Conference on*

Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011.

[Yang et al. 2014] Yang, Y.; Wang, Z.; Yang, J.; Wang, J.; Chang, S.; and Huang, T. S. 2014. Data clustering by laplacian regularized l1-graph. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, 3148–3149.

[Ye, Zhao, and Wu 2007] Ye, J.; Zhao, Z.; and Wu, M. 2007. Discriminative k-means for clustering. In *Proc. NIPS*, 1649–1656.

[Zelnik-Manor and Perona 2004] Zelnik-Manor, L., and Perona, P. 2004. Self-tuning spectral clustering. In *Proc. NIPS*, 1601–1608.