A Unified Framework for Discrete Spectral Clustering

Yang Yang[†], Fumin Shen[†], Zi Huang[§], Heng Tao Shen[†§]
[†]University of Electronic Science and Technology of China, Chengdu, China
[§]The University of Queensland, Brisbane, Australia
{dlyyang, fumin.shen}@gmail.com, huang@itee.uq.edu.au, shenhengtao@hotmail.com

Abstract

Spectral clustering has been playing a vital role in various research areas. Most traditional spectral clustering algorithms comprise two independent stages (i.e., first learning continuous labels and then rounding the learned labels into discrete ones), which may lead to severe information loss and performance degradation. In this work, we study how to achieve discrete clustering as well as reliably generalize to unseen data. We propose a unified spectral clustering scheme which jointly learns discrete clustering labels and robust out-of-sample prediction functions. Specifically, we explicitly enforce a discrete transformation on the intermediate continuous labels, which leads to a tractable optimization problem with a discrete solution. Moreover, to further compensate the unreliability of the learned labels, we integrate an adaptive robust module with $\ell_{2,p}$ loss to learn prediction function for unseen data. Extensive experiments conducted on various data sets have demonstrated the superiority of our proposal as compared to existing clustering approaches.

1 Introduction

Clustering [Jain *et al.*, 1999; Rodriguez and Laio, 2014; Bühler and Hein, 2009; Belkin *et al.*, 2016] has long been serving as a critical technique in modern research fields, such as image segmentation [Shi and Malik, 2000; Felzenszwalb and Huttenlocher, 2004], gene expression analysis [Jiang *et al.*, 2004], face analysis [Elhamifar and Vidal, 2013], content-based image retrieval [Gordon *et al.*, 2003], image annotation [Wang *et al.*, 2008], heterogeneous data analysis [Liu *et al.*, 2015] and image hashing [Shen *et al.*, 2015].

As one of the most classic clustering approaches, k-means has been extensively applied in practice. The typical procedure of k-means iteratively assigns data points to their closest clusters and updates clustering centers. Nonetheless, k-means suffers from the problem of "curse of dimensionality" [Ding and Li, 2007]. Recent research endeavors focus on finding a low-dimensional projection using dimensionality reduction techniques (e.g., PCA) and then performing k-means. Furthermore, several works have tried to employ discriminative analysis [De la Torre and Kanade, 2006; Ye *et al.*, 2007a; Yang *et al.*, 2011] to generate better partition of data. In [De la Torre and Kanade, 2006; Ding and Li, 2007], k-means and LDA were integrated together to discover discriminative subspace. In [Ye *et al.*, 2007b], Ye *et al.* proposed a discriminative k-means (DKM) algorithm to formalize clustering as a trace maximization problem for learning better clustering labels.

As an alternative promising direction, spectral clustering [Filippone *et al.*, 2008] has demonstrated its strong capability in group objects by analyzing complex data structural information. Spectral clustering has been extensively used in real-world applications, such as image/video segmentation [Galasso *et al.*, 2014]. The advantage of spectral clustering family lies in the exploration of the intrinsic data structures [Xia *et al.*, 2014; Yang *et al.*, 2013; 2015], which are fully employed for predicting clustering labels by exploiting the different similarity graphs of data points. For instance, in [Wu and Scholkopf, 2007], an effective algorithm, termed as local learning based clustering (LLC), was developed according to the assumption that the cluster label of a data point can be determined by its neighbors.

It is well-known that optimizing the spectral clustering models will lead to an NP-hard problem due to the discrete constraint on the clustering labels. To achieve a feasible approximate solution, most spectral clustering algorithms follow a common practical paradigm. It first relaxes the discrete constraint to allow the clustering label matrix to be continuous-valued and performs eigenvalue decomposition on the specific Laplacian matrix to generate an approximate indicator with continuous values. Then, we can discretize the clustering label matrix by employing certain independent technique, such as k-means. Furthermore, to enable clustering new unseen data, one may learn an additional prediction function in an independent stage (module). In [Yang *et al.*, 2013; Nie *et al.*, 2011], the out-of-sample problem is addressed by introducing a regression learning module, and discriminative information is injected into the construction of the similarity matrix to improve clustering performance. Although existing spectral clustering has been applied in practice widely, they may easily achieve poor performance due to the following drawbacks:

- High risk of severe deviation of approximate solution

from the genuine discrete clustering labels;

- Information loss among separate independent stages, i.e., continuous label generation, label discretization and prediction function learning;
- Unreliability of the predicted cluster labels leading to poor prediction functions.

To cope with the aforementioned problems, we propose a spectral clustering scheme which directly learns discrete clustering labels and robust out-of-sample prediction functions in a unified manner. Specifically, in order to alleviate the influence caused by the information loss during the relaxation of traditional spectral clustering, we deliberately recover the abandoned discrete constraint with a smooth transformation (e.g., rotation) from the relaxed continuous clustering labels to a discrete solution. In this sense, the continuous clustering label just serves as an intermediate product. We integrate a discrete rotation functionality, which guarantees a tractable optimization problem with a discrete solution. Moreover, to further compensate the unreliability of the learned labels, we integrate an adaptive robust module to learn prediction function for unseen data. In particular, we devise a novel noise modelling approach by utilizing an effective $\ell_{2,p}$ loss term over the prediction error residual to capture unreliability of clustering labels in a more adaptive way. The $\ell_{2,p}$ loss is capable of inducing sample-wise sparsity, which naturally identifies unreliable predicted labels. Besides, different choices of p enables sufficient control on unpredictable conditions of label noise.

The rest of this paper is organized as follows. Section 2 elaborates the details of the proposed model, including problem formulation, introduction of model, an efficient algorithm and the corresponding analysis. Experimental results are reported and analyzed in Section 3. In the last, we conclude our work in Section 4.

2 The Proposed Method

In this section, we elaborate the proposed discrete spectral clustering. We first present the formulation and then develop an efficient algorithm for optimization.

2.1 Problem Formulation

Given n data points of d dimensions, denoted as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, we aim to partition $\mathbf{X}$ into c groups $\{\mathbf{C}_j\}_{j=1}^{c}$ according to certain criteria. For instance, data in the same cluster should be similar to each other while those in different groups should have dissimilar representations. The family of spectral clustering algorithms is generally formulated as below:

$$\min_{\mathbf{Y}} Tr(\mathbf{Y}^T\mathbf{L}\mathbf{Y}) \quad \text{s.t. } \mathbf{Y} \in Idx, \tag{1}$$

where $Tr(\cdot)$ computes the trace of a matrix and $\mathbf{L} \in \mathbb{R}^{n\times n}$ is the Laplacian matrix of $\mathbf{X}$. $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]^T \in \mathbb{R}^{n\times c}$ is the clustering label matrix, and $\mathbf{Y} \in Idx$ means the clustering label vector of each sample $\mathbf{y}_i \in \{0,1\}^{c\times 1}$ contains one and only one element "1", which indicates the group membership of $\mathbf{x}_i$. It is recognized that the problem in Eq.(1) is NP-hard due to the discrete constraint on $\mathbf{Y}$. A commonly-used practical way is to relax $\mathbf{Y}$ to be continuous-valued and thus we arrive at

$$\min_{\mathbf{F}} Tr(\mathbf{F}^T\mathbf{L}\mathbf{F}) \quad \text{s.t. } \mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \tag{2}$$

where $\mathbf{F} \in \mathbb{R}^{n\times c}$ is the relaxed continuous clustering label matrix. $\mathbf{I}_c$ is an identity matrix of size $c \times c$. The orthogonal constraint $\mathbf{F}^T\mathbf{F} = \mathbf{I}_c$ is used to avoid trivial solution. After solving the above problem via eigenvalue decomposition, we further use traditional clustering method, e.g., k-means, to transform clustering labels into discrete ones.

Though the two-stage strategy provides a feasible solution, it may unpredictably deviate from the genuine discrete clustering labels. To avoid this situation, we intend to devise a unified spectral clustering model to directly generate the discrete clustering label matrix.

2.2 Model

To bypass the difficulty of the NP-hard problem in Eq.(1), we propose to re-introduce the "new" discrete clustering indicator $\mathbf{Y}$ into Eq.(2). We expect $\mathbf{Y}$ has exactly the same properties as that in Eq.(1). Meanwhile, $\mathbf{Y}$ should also be certain reasonable transformation from $\mathbf{F}$ in order to preserve the structural knowledge derived from data. Based on such analysis, we employ the following model to achieve these goals:

$$\begin{aligned} \min_{\mathbf{F},\mathbf{Q},\mathbf{Y}} \; & Tr(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \alpha\|\mathbf{Y} - \mathbf{F}\mathbf{Q}\|_F^2, \\ \text{s.t. } & \mathbf{F}^T\mathbf{F} = \mathbf{I}_c \wedge \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_c \wedge \mathbf{Y} \in Idx, \end{aligned} \tag{3}$$

where $\mathbf{Q} \in \mathbb{R}^{c\times c}$ is a rotation matrix which adjusts the continuous-valued clustering label matrix into the real discrete clustering label matrix $\mathbf{Y}$. α is a trade-off parameter.

Recall that existing clustering methods can hardly generalize to new unseen data. Possible solutions either learn new prediction functions independently from scratch [Bengio *et al.*, 2004] or incorporate learning model into clustering to generate prediction functions [Nie *et al.*, 2011; Yang *et al.*, 2013]. Denote prediction function learning component as $\mathcal{L}(\mathbf{W}; \mathbf{X}, \mathbf{Y})$. One may choose the following loss function:

$$\mathcal{L}(\mathbf{W}; \mathbf{X}, \mathbf{Y}) = \|\mathbf{Y} - \mathbf{X}^T\mathbf{W}\|_F^2 + \gamma\|\mathbf{W}\|_F^2, \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{d\times c}$ is the mapping variable.

However, integrating the above model into the discrete optimization framework as in Eq.(3) suffers from the following problem. The noise in $\mathbf{Y}$ can hardly be captured by Eq.(4). For example, if the clustering label of a training sample is incorrectly generated in previous process, then the error will be inevitably amplified due to the squared residual. The unreliability of the learned cluster labels probably in turn jeopardizes the subsequent learning of prediction functions. To handle the above problem, we propose to utilize a robust $\ell_{2,p}$ loss term [Yang *et al.*, 2014] to effectively control different levels of noise. Also, we also design a mutually reinforcing mechanism to explicitly make generation of clustering label codes and training of prediction functions interacted with each other, thereby exerting positive influence to the whole

learning process. The robust learning model is depicted as follows:

$$\mathcal{L}_{2,p}(\mathbf{W};\mathbf{X},\mathbf{Y}) = \|\mathbf{Y}-\mathbf{X}^T\mathbf{W}\|_{2,p} + \gamma\|\mathbf{W}\|_F^2, \quad (5)$$

where the loss function is changed to $\ell_{2,p}$ $(0 < p < 2)$ loss, which is capable of alleviating sample noise:

$$\|\mathbf{M}\|_{2,p} = \sum_{i=1}^{n} \|(\mathbf{M})_i\|_2^p, \quad (6)$$

where $(\mathbf{M})_i$ is the i-th row of matrix $\mathbf{M}$. The above $\ell_{2,p}$ loss not only suppresses the inevitable noise but also enhances the flexibility for adapting different noise levels.

By substituting the robust learning module in Eq.(4) into Eq.(3), we have

$$\begin{aligned} \min_{\mathbf{F},\mathbf{Q},\mathbf{Y},\mathbf{W}} \; & Tr(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \alpha\|\mathbf{Y}-\mathbf{F}\mathbf{Q}\|_F^2 + \beta\mathcal{L}_{2,p}(\mathbf{W};\mathbf{X},\mathbf{Y}), \\ & \text{s.t. } \mathbf{F}^T\mathbf{F} = \mathbf{I}_c \wedge \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_c \wedge \mathbf{Y} \in Idx, \end{aligned} \quad (7)$$

where $\alpha > 0$ and $\beta > 0$ are balance parameters.

2.3 Solution

Due to the existence of $\ell_{2,p}$ loss, direct optimizing the model (7) turns out to be difficult. In this part, we propose an efficient and effective algorithm which iteratively solves an alternative optimization problem and guarantees the obtained solution is the optimal solution to the original problem in (7). Denote the loss residual as

$$\mathbf{R} = \mathbf{Y} - \mathbf{X}^T\mathbf{W}, \quad (8)$$

then we introduce the alternative problem as follows:

$$\begin{aligned} \min_{\mathbf{F},\mathbf{Q},\mathbf{Y},\mathbf{W}} \; & Tr(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \alpha\|\mathbf{Y}-\mathbf{F}\mathbf{Q}\|_F^2 \\ & + \beta(Tr(\mathbf{R}^T\mathbf{D}\mathbf{R}) + \gamma\|\mathbf{W}\|_F^2), \\ & \text{s.t. } \mathbf{F}^T\mathbf{F} = \mathbf{I}_c \wedge \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_c \wedge \mathbf{Y} \in Idx, \end{aligned} \quad (9)$$

where $\mathbf{D}$ is a diagonal matrix with its i-th diagonal element computed as

$$\mathbf{D}_{ii} = \frac{1}{\frac{2}{p}\|\mathbf{r}_i\|_2^{2-p}} \quad (10)$$

where $\mathbf{r}_i$ is the i-th row of $\mathbf{R}$.

We show how to solve the above alternative problem first, and then explain the obtained solution is the optima of Eq.(7).

Update $\mathbf{F}$: With $\mathbf{Q}, \mathbf{W}, \mathbf{Y}$ fixed, the problem is reduced to

$$\begin{aligned} \min_{\mathbf{F}} \; & Tr(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \alpha\|\mathbf{Y}-\mathbf{F}\mathbf{Q}\|_F^2, \\ & \text{s.t. } \mathbf{F}^T\mathbf{F} = \mathbf{I}_c. \end{aligned} \quad (11)$$

The above problem with orthogonal constraint can be efficiently solved using the algorithm [Wen and Yin, 2013].

Update $\mathbf{Y}$: When $\mathbf{F}, \mathbf{Q}, \mathbf{W}$ are fixed, we have

$$\min_{\mathbf{Y}\in Idx} \; \alpha\|\mathbf{Y}-\mathbf{F}\mathbf{Q}\|_F^2 + \beta Tr((\mathbf{Y}-\mathbf{X}^T\mathbf{W})^T\mathbf{D}(\mathbf{Y}-\mathbf{X}^T\mathbf{W})), \quad (12)$$

Given the facts that $Tr(\mathbf{Y}^T\mathbf{Y}) = n$ and $Tr(\mathbf{Y}^T\mathbf{D}\mathbf{Y}) = Tr(\mathbf{D})$, we can rewrite the above sub-problem as below:

$$\max_{\mathbf{Y}\in Idx} \; Tr(\mathbf{Y}^T\mathbf{P}), \quad (13)$$

Algorithm 1 Algorithm for optimizing the proposed spectral clustering model.

Input: Training data $\mathbf{X}$;
Output: $\mathbf{F}, \mathbf{Q}, \mathbf{W}, \mathbf{Y}$;
1: Randomly initialize $\mathbf{F}, \mathbf{Q}, \mathbf{W}, \mathbf{Y}$;
2: Construct Laplacian matrix $\mathbf{L}$;
3: **repeat**
4: Update $\mathbf{D}$ according to Eq.(10);
5: Update $\mathbf{F}$ by solving the problem in Eq.(11);
6: Update $\mathbf{Y}$ according to Eq.(14);
7: Update $\mathbf{Q}$ by solving the problem in Eq.(15);
8: Update $\mathbf{W}$ according to Eq.(17);
9: **until** there is no change to $\mathbf{F}, \mathbf{Q}, \mathbf{W}, \mathbf{Y}$
10: **return** $\mathbf{F}, \mathbf{Q}, \mathbf{W}, \mathbf{Y}$

where $\mathbf{P} = \alpha\mathbf{F}\mathbf{Q} + \beta\mathbf{D}\mathbf{X}^T\mathbf{W}$. The optimal solution of the above sub-problem can be easily generated as follows:

$$\mathbf{Y}_{ij} = \begin{cases} 1, & j = \arg\max_k \mathbf{P}_{ik} \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

Update $\mathbf{Q}$: When $\mathbf{F}, \mathbf{W}, \mathbf{Y}$ are fixed, the problem becomes

$$\min_{\mathbf{Q}} \; \|\mathbf{Y}-\mathbf{F}\mathbf{Q}\|_F^2, \quad \text{s.t. } \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_c, \quad (15)$$

which can be efficiently solved using the algorithm [Wen and Yin, 2013].

Update $\mathbf{W}$: Fixing $\mathbf{F}, \mathbf{Q}, \mathbf{Y}$, we arrive at

$$\min_{\mathbf{W}} \; Tr((\mathbf{Y}-\mathbf{X}^T\mathbf{W})^T\mathbf{D}(\mathbf{Y}-\mathbf{X}^T\mathbf{W})) + \gamma\|\mathbf{W}\|_F^2. \quad (16)$$

Setting the derivative of Eq.(16) w.r.t. $\mathbf{W}$ to zero, we have

$$\mathbf{W} = (\mathbf{X}\mathbf{D}\mathbf{X}^T + \gamma\mathbf{I}_d)^{-1}\mathbf{X}\mathbf{D}\mathbf{Y}, \quad (17)$$

where $\mathbf{I}_d$ is an identity matrix of size $d \times d$.

We summarize the optimization process of Eq.(7) in Algorithm 1. We will show that Algorithm 1 converges to an optima of the original problem in Eq.(7).

We first introduce two lemmas.

Lemma 1. *Let $\mathbf{r}_i$ be the i-th row of the residual $\mathbf{R}$ in previous iteration, and $\tilde{\mathbf{r}}_i$ be the i-th row of the residual $\tilde{\mathbf{R}}$ in current iteration, then the following inequality holds:*

$$\|\tilde{\mathbf{r}}_i\|^p - \frac{p\|\tilde{\mathbf{r}}_i\|^2}{2\|\mathbf{r}_i\|^{2-p}} \le \|\mathbf{r}_i\|^p - \frac{p\|\mathbf{r}_i\|^2}{2\|\mathbf{r}_i\|^{2-p}}. \quad (18)$$

Proof. Given the following unary function

$$h(x) = px^2 - 2x^p + (2-p), \quad 0 < p < 2. \quad (19)$$

We have $h'(x) = 2px - 2px^{p-1}$ and $h''(x) = 2p - 2p(p-1)x^{p-2}$. Clearly, $h'(x) = 0$ only when x equals to 1. Besides, when $x \in (0,1)$, $h'(x) < 0$ and when $x > 1$, $h'(x) > 0$, which indicates that $h(x)$ is monotonically decreasing as $0 < x < 1$ and monotonically increasing when $x > 1$. Furthermore, we know $h''(1) = 2p(2-p) > 0$. Hence, we have the conclusion that for $\forall x > 0$, $h(x) \ge h(1) = 0$.

Then, by substituting $x = \frac{\|\tilde{\mathbf{r}}_i\|}{\|\mathbf{r}_i\|}$ into Eq.(19), we obtain the conclusion

$$\begin{aligned}
& p\frac{\|\tilde{\mathbf{r}}_i\|^2}{\|\mathbf{r}_i\|^2} - 2\frac{\|\tilde{\mathbf{r}}_i\|^p}{\|\mathbf{r}_i\|^p} + (2-p) \geq 0, \\
\Leftrightarrow\ & p\|\tilde{\mathbf{r}}_i\|^2 - 2\|\tilde{\mathbf{r}}_i\|^p\|\mathbf{r}_i\|^{2-p} + (2-p)\|\mathbf{r}_i\|^2 \geq 0, \\
\Leftrightarrow\ & p\|\tilde{\mathbf{r}}_i\|^2\|\mathbf{r}_i\|^{p-2} - 2\|\tilde{\mathbf{r}}_i\|^p + (2-p)\|\mathbf{r}_i\|^p \geq 0, \\
\Leftrightarrow\ & 2\|\tilde{\mathbf{r}}_i\|^p - p\|\tilde{\mathbf{r}}_i\|^2\|\mathbf{r}_i\|^{p-2} \leq (2-p)\|\mathbf{r}_i\|^p, \\
\Leftrightarrow\ & \|\tilde{\mathbf{r}}_i\|^p - \frac{p\|\tilde{\mathbf{r}}_i\|^2}{2\|\mathbf{r}_i\|^{2-p}} \leq \|\mathbf{r}_i\|^p - \frac{p\|\mathbf{r}_i\|^p}{2\|\mathbf{r}_i\|^{2-p}}.
\end{aligned}$$

□

Lemma 2. *Given* $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2 \ldots, \mathbf{r}_n]^T$*, then we have the following conclusion:*

$$\sum_{i=1}^{n} \|\tilde{\mathbf{r}}_i\|^p - \sum_{i=1}^{n} \frac{p\|\tilde{\mathbf{r}}_i\|^2}{2\left\|\mathbf{r}_i\right\|^{2-p}} \leq \sum_{i=1}^{n} \|\mathbf{r}_i\|^p - \sum_{i=1}^{n} \frac{p\left\|\mathbf{r}_i\right\|^p}{2\left\|\mathbf{r}_i\right\|^{2-p}}. \tag{20}$$

Proof. By summing up the inequalities of all $\mathbf{r}_i$, $i = 1, 2, \ldots, n$ according to Lemma 1, we can easily reach the conclusion of Lemma 2. □

Theorem 1. *Each iteration (line 4 to line 8) of Algorithm 1 decreases the value of the objective function in Eq.(7) monotonically.*

Proof. Suppose $\tilde{\mathbf{Y}}, \tilde{\mathbf{F}}, \tilde{\mathbf{Q}}, \tilde{\mathbf{W}}$ are the optimized solution of the alternative problem (9), and we denote

$$\begin{cases} \mathcal{J} = Tr(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \alpha\left\|\mathbf{Y} - \mathbf{F}\mathbf{Q}\right\|_F^2 + \beta\gamma\|\mathbf{W}\|_F^2, \\ \tilde{\mathcal{J}} = Tr(\tilde{\mathbf{F}}^T\mathbf{L}\tilde{\mathbf{F}}) + \alpha\|\tilde{\mathbf{Y}} - \tilde{\mathbf{F}}\tilde{\mathbf{Q}}\|_F^2 + \beta\gamma\|\tilde{\mathbf{W}}\|_F^2, \end{cases}$$

then we have

$$\begin{aligned}
& \tilde{\mathcal{J}} + \beta Tr(\tilde{\mathbf{R}}^T\mathbf{D}\tilde{\mathbf{R}}) \leq \mathcal{J} + \beta Tr(\mathbf{R}^T\mathbf{D}\mathbf{R}) \\
\Rightarrow\ & \tilde{\mathcal{J}} + \beta\sum_{i=1}^{n} \frac{p\|\tilde{\mathbf{r}}_i\|^2}{2\|\mathbf{r}_i\|^{2-p}} \leq \mathcal{J} + \beta\sum_{i=1}^{n} \frac{p\|\mathbf{r}_i\|^2}{2\|\mathbf{r}_i\|^{2-p}} \\
\Rightarrow\ & \tilde{\mathcal{J}} + \beta\sum_{i=1}^{n} \|\tilde{\mathbf{r}}_i\|^p - \beta(\sum_{i=1}^{n} \|\tilde{\mathbf{r}}_i\|^p - \sum_{i=1}^{n} \frac{p\|\tilde{\mathbf{r}}_i\|^2}{2\|\mathbf{r}_i\|^{2-p}}) \leq \\
& \mathcal{J} + \beta\sum_{i=1}^{n} \|\mathbf{r}_i\|^p - \beta(\sum_{i=1}^{n} \|\mathbf{r}_i\|^p - \sum_{i=1}^{n} \frac{p\|\mathbf{r}_i\|^2}{2\|\mathbf{r}_i\|^{2-p}}).
\end{aligned}$$

Using Lemma 2, we have

$$\tilde{\mathcal{J}} + \beta\sum_{i=1}^{n} \|\tilde{\mathbf{r}}_i\|^p \leq \mathcal{J} + \beta\sum_{i=1}^{n} \|\mathbf{r}_i\|^p.$$

which indicates the monotonic decreasing trend of the objective function in Eq. (7) in each iteration. □

To sum up, we can see that Algorithm 1 will eventually converge to the optima of the original problem in Eq. (7) according to Theorem 1.

3 Experiments

3.1 Experimental Settings

In the experiments, we evaluate our proposed approach on six UCI datasets [Lichman, 2013], including Image Segmentation, Vehicle, Vote, Ecoli, Solar and Wine. The statistics of datasets are summarized in Table 1.

Table 1: Statistics of the evaluated data sets.

	Number of Samples	Number of Dimensions	Number of Classes
Segment	2310	19	7
Vehicle	846	18	4
Vote	435	16	2
Ecoli	336	343	8
Solar	323	12	6
Wine	178	13	3

We choose several existing clustering approaches for comparison, including k-means clustering (KM), discriminative k-means clustering (DKM) [Ye *et al.*, 2007b], Spectral Clustering (SC), Local Learning Clustering (LLC) [Wu and Scholkopf, 2007], CLGR [Wang *et al.*, 2009] and Spectral Embedding Clustering (SEC) [Nie *et al.*, 2011]. We set the number of neighbors k to 5 for all spectral clustering methods. The parameters of all comparison algorithms are tested in $\{10^{-6}, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}, 10^{6}\}$. We set p of $\ell_{2,p}$ loss in $\{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75\}$. We randomly choose 50% of samples for training and the rest are used for test.

We employ conventional **Normalized Mutual Information (NMI)** and **Accuracy (ACC)** as evaluation metrics.

- **NMI**: We first define normalized mutual information of two distributions A and B as below:

$$NMI(A, B) = \frac{\mathcal{I}(A, B)}{\sqrt{\mathcal{H}(A)\mathcal{H}(B)}}, \tag{21}$$

  where $\mathcal{I}(A, B)$ computes the mutual information of A and B. $\mathcal{H}(\cdot)$ is the entropy of a distribution. Denote n_i as the number of datums in the i-th cluster $\mathbf{C}_i$ generated by a clustering algorithm, $\hat{n}_j$ as the number of data points in the j-th ground truth class $\mathbf{G}_j$ and $n_{i,j}$ as the number of data occurring in both $\mathbf{C}_i$ and $\mathbf{G}_j$. Then, NMI is calculated as follows:

$$NMI = \frac{\sum_{i=1}^{c}\sum_{j=1}^{c} n_{i,j}\log(\frac{n\times n_{i,j}}{n_i\hat{n}_j})}{\sqrt{(\sum_{i=1}^{c} n_i\log\frac{n_i}{n})(\sum_{j=1}^{c}\hat{n}_j\log\frac{\hat{n}_j}{n})}}. \tag{22}$$

  Larger NMI values indicate better clustering performance.

- **ACC**: Denote $\mathbf{o}_i$ as the resultant clustering label of $\mathbf{x}_i$ using certain clustering method and $\mathbf{g}_i$ as the ground truth of $\mathbf{x}_i$, then we have

$$ACC = \frac{\sum_i \delta(\mathbf{o}_i, map(\mathbf{g}_i))}{n}, \tag{23}$$

  where $\delta(x, y) = 1$ if $x = y$; $\delta(x, y) = 0$ otherwise, and $map(\mathbf{g}_i)$ is the best mapping function that permutes clustering labels to match the ground truth labels. Larger ACC values indicate better clustering performance.

Table 2: Comparison of our proposed approach and existing methods on six evaluated datasets. (a) ACC on Training set; (b) NMI on Training set; (c) ACC on Test set; and (d) NMI on Test set.

(a) ACC on Training set.

	KM	DKM	SC	LL	CLGR	SEC	Ours
Solar	0.4568	0.5679	0.4444	0.4074	0.5370	0.5802	**0.6111**
Vehicle	0.4539	0.4539	0.4941	0.5532	0.5177	0.4681	**0.6312**
Vote	0.7844	0.8624	0.8394	0.8716	0.8440	0.7936	**0.8807**
Ecoli	0.6488	**0.6667**	0.3631	0.4286	0.5179	0.6429	0.6607
Segment	0.5610	0.7100	0.5700	0.5960	0.6190	0.6620	**0.7720**
Wine	0.7079	0.7079	0.7079	0.5843	0.7079	0.7416	**0.9101**

(b) NMI on Training set.

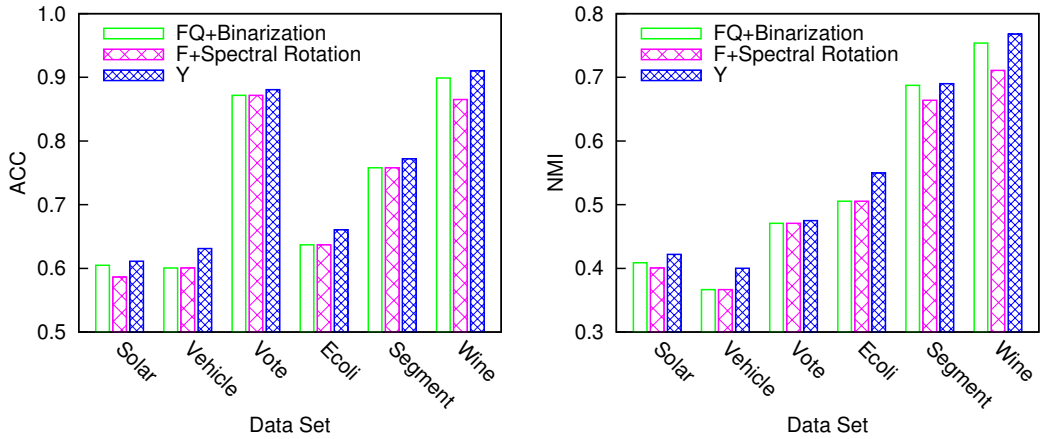
	KM	DKM	SC	LL	CLGR	SEC	Ours
Solar	0.3654	0.4196	0.2361	0.3289	0.3947	0.4194	**0.4219**
Vehicle	0.2179	0.2179	0.2562	0.3055	0.2562	0.2559	**0.4001**
Vote	0.2571	0.3943	0.3492	0.4077	0.3585	0.2834	**0.4750**
Ecoli	0.4630	0.4718	0.2759	0.3362	0.3332	0.4557	**0.5498**
Segment	0.5167	0.6785	0.5945	0.5360	0.6066	0.6186	**0.6899**
Wine	0.3743	0.3743	0.3643	0.2834	0.3643	0.4586	**0.7678**

(c) ACC on Test set.

	KM	DKM	SC	LL	CLGR	SEC	Ours
Solar	0.5155	0.5466	0.3043	0.5217	0.6025	0.5776	**0.6149**
Vehicle	0.4468	0.4586	0.3877	0.5272	0.4894	0.4539	**0.5957**
Vote	0.8065	0.8433	0.7926	0.8986	0.8525	0.8111	**0.9263**
Ecoli	0.7321	0.7262	0.6548	0.7202	0.6726	0.7381	**0.8036**
Segment	0.5855	0.7328	0.6809	0.6496	0.6916	0.7115	**0.7511**
Wine	0.7079	0.7079	0.6292	0.5393	0.7753	0.6966	**0.9101**

(d) NMI on Test set.

	KM	DKM	SC	LL	CLGR	SEC	Ours
Solar	0.3755	0.3885	0.2402	0.3301	0.3971	0.4055	**0.4625**
Vehicle	0.1955	0.1955	0.1231	0.2723	0.2401	0.2037	**0.3485**
Vote	0.3051	0.3678	0.3161	0.5189	0.4362	0.3179	**0.6441**
Ecoli	0.5242	0.5123	0.4779	0.5567	0.5159	0.6443	**0.6710**
Segment	0.5452	0.6736	0.6274	0.5697	0.6273	0.6203	**0.6991**
Wine	0.4601	0.4601	0.2516	0.2324	0.4978	0.5260	**0.7525**



(a) ACC (b) NMI

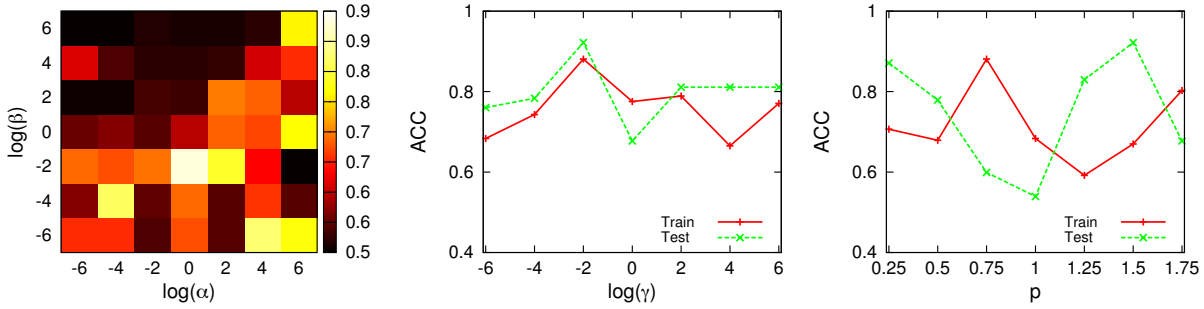
Figure 1: Comparison of discrete clustering labels and continuous clustering labels. (a) and (b) illustrate ACC performance and NMI performance, respectively.

3.2 Comparison

We compare our proposed discrete spectral clustering to several existing methods. Table 2 reports ACC and NMI performance on training and test sets of six evaluated datasets, from which we derive the following observation and analysis.

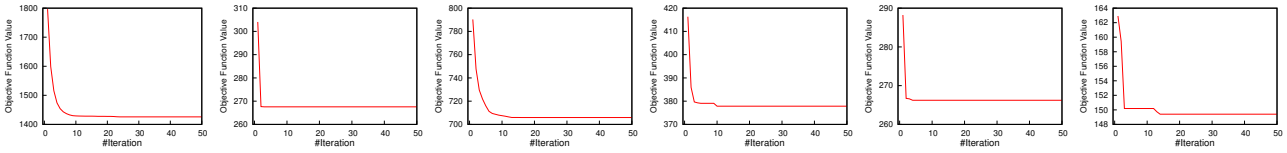
As we can see, our proposed approach significantly outperforms other comparison methods, including traditional clustering (KM and DKM), spectral clustering (SC, LL and CLGR) and clustering with out-of-sample extension (SEC). On the one hand, our solution is capable of searching "genuine" clustering membership of data, i.e., discrete labels, which helps to reach the optimal clustering without any extra discretizing manipulation, thereby leading to minimal loss. On the other hand, the optimization of the prediction function learning module helps to control the unreliable factors in the derived clustering labels, which in turn exerts positive influence on the learning of the discrete spectral clustering to further boost the quality of the resultant clustering labels.

In order to further illustrate the efficacy of the discrete clustering, we compare clustering performance of discrete label matrix $\mathbf{Y}$, original continuous label matrix $\mathbf{F}$ with spectral rotation [Huang *et al.*, 2013] and the rotated continuous label matrix $\mathbf{FQ}$ with maximum value binarization. The ACC and NMI performance is reported in Figure 1. As we can see, the discrete clustering label matrix $\mathbf{Y}$ always achieves the best performance, and $\mathbf{FQ}$ performs slightly better than $\mathbf{F}$. This



(a) α and β (Train ACC) (b) γ (c) p

Figure 2: Parameter sensitivity on dataset Wine (Train ACC).



(a) Solar (b) Vehicle (c) Vote (d) Ecoli (e) Segment (f) Wine

Figure 3: Convergence analysis on six datasets.

phenomenon again implies that our discrete solution is closer to the real partition of data than the relaxed continuous clustering labels.

3.3 Parameter Sensitivity

We further study how our approach performs when using different settings of parameters. As stated before, we tune trade-off parameter α, β and γ in the range of $\{10^{-6}, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}, 10^{6}\}$, and p from 0.25 to 1.75. Figure 2 illustrates the experimental results.

- **Joint effects of α and β.** α and β controls the contributions of continuous label generation, label rotation and prediction learning. When α and β are neither too large nor too small (e.g., $\alpha = 1$ and $\beta = 0.01$), our approach can achieve satisfactory performance. This hints us that different components should be well balanced in order to maximize the joint effects.
- **Effects of γ.** γ controls the inner balance of prediction residual error $\|\mathbf{Y} - \mathbf{X}^T\mathbf{W}\|_{2,p}$ and regularizer $\|\mathbf{W}\|_F^2$. As γ increases from 10^{-6} to 10^{-2}, our approach performs gradually better. When γ keeps going up, we see a decreasing trend instead. If γ is small, the regularizer will be easily ignored, which may lead to overfitting problem. In contrary, when we use large γ, the prediction residual error will not be well controlled, which makes our approach to produce poor prediction functions and clustering labels.
- **Effects of p.** The performance of our approach fluctuates (from 0.5 to 0.9) when p varies from 0.25 to 1.75. For both training and test settings, the best performance is gained when p is not close to 2, which indeed reflects the ability of $\ell_{2,p}$ loss handling unreliable clustering labels.

3.4 Convergence Study

As aforementioned, we have shown that Algorithm 1 is guaranteed to converge to an optima when we iteratively solve the alternative problem (9). In this part, we empirically test the convergence and the efficiency of the proposed algorithm. As illustrated in Figure 3, we conduct experiments on six evaluated datasets. For illustration purpose, we consistently fix the values of all parameters to 1. As we can see, our algorithm is able to achieve a very rapid convergence within only a few iterations (less than 10). In this sense, it is reasonable for us to set the number of iterations to 10, which provides sufficient efficiency for the learning.

4 Conclusion

In this work, we coped with the problem existing in most traditional spectral clustering algorithms, i.e., relaxing discrete constraints to continuous one, which consists of two independent stages (i.e., first learning continuous labels and then rounding the learned labels into discrete ones). In order to reduce information loss and performance degradation, we proposed a unified spectral clustering approach to directly learn discrete clustering labels and robust out-of-sample prediction functions. To be more specific, our proposed approach can explicitly rotate continuous labels to discrete ones. Meanwhile, to the end of handling the noisy clustering labels, we integrated an adaptive robust module to learn prediction function for unseen data. Extensive experiments on six data sets demonstrated the promising performance of our proposal as compared to existing clustering approaches.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Project 61502081 and Project 61572108, and the Fundamental Research Funds for the Central Universities under Project ZYGX2014Z007.

References

[Belkin *et al.*, 2016] Mikhail Belkin, Luis Rademacher, and James Voss. The hidden convexity of spectral clustering. *AAAI*, 2016.

[Bengio *et al.*, 2004] Y. Bengio, J.F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *NIPS*, 16:177–184, 2004.

[Bühler and Hein, 2009] Thomas Bühler and Matthias Hein. Spectral clustering based on the graph p-laplacian. In *ICML*, pages 81–88, 2009.

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

[Ding and Li, 2007] C. Ding and T. Li. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *ICML*, pages 521–528, 2007.

[Elhamifar and Vidal, 2013] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE TPAMI*, 35(11):2765–2781, 2013.

[Felzenszwalb and Huttenlocher, 2004] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

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

[Galasso *et al.*, 2014] Fabio Galasso, Margret Keuper, Thomas Brox, and Bernt Schiele. Spectral graph reduction for efficient image and streaming video segmentation. In *CVPR*, pages 49–56, 2014.

[Gordon *et al.*, 2003] S. Gordon, H. Greenspan, and J. Goldberger. Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations. In *CVPR*, pages 370–377, 2003.

[Huang *et al.*, 2013] Jin Huang, Feiping Nie, and Heng Huang. Spectral rotation versus k-means in spectral clustering. In *AAAI*, 2013.

[Jain *et al.*, 1999] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[Jiang *et al.*, 2004] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. *IEEE TKDE*, 16(11):1370–1386, 2004.

[Lichman, 2013] M. Lichman. UCI machine learning repository, 2013.

[Liu *et al.*, 2015] Hongfu Liu, Tongliang Liu, Junjie Wu, Dacheng Tao, and Yun Fu. Spectral ensemble clustering. In *SIGKDD*, pages 715–724, 2015.

[Nie *et al.*, 2011] Feiping Nie, Zinan Zeng, Ivor W Tsang, Dong Xu, and Changshui Zhang. Spectral embedded clustering: a framework for in-sample and out-of-sample spectral clustering. *IEEE TNN*, 22(11):1796–1808, 2011.

[Rodriguez and Laio, 2014] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[Shen *et al.*, 2015] Fumin Shen, Chunhua Shen, Qinfeng Shi, Anton van den Hengel, Zhenmin Tang, and Heng Tao Shen. Hashing on nonlinear manifolds. *IEEE TIP*, 24(6):1839–1851, 2015.

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

[Wang *et al.*, 2008] X.J. Wang, L. Zhang, X. Li, and W.Y. Ma. Annotating images by mining image search results. *IEEE TPAMI*, 30(11):1919–1932, 2008.

[Wang *et al.*, 2009] F. Wang, C. Zhang, and T. Li. Clustering with local and global regularization. *IEEE TKDE*, 21(12):1665–1678, 2009.

[Wen and Yin, 2013] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.

[Wu and Scholkopf, 2007] M. Wu and B. Scholkopf. A local learning approach for clustering. *NIPS*, 19:1529–1536, 2007.

[Xia *et al.*, 2014] Rongkai Xia, Yan Pan, Lei Du, and Jian Yin. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *AAAI*, pages 2149–2155, 2014.

[Yang *et al.*, 2011] Yi Yang, Heng Tao Shen, Feiping Nie, Rongrong Ji, and Xiaofang Zhou. Nonnegative spectral clustering with discriminative regularization. In *AAAI*, 2011.

[Yang *et al.*, 2013] Yang Yang, Yi Yang, Heng Tao Shen, Yanchun Zhang, Xiaoyong Du, and Xiaofang Zhou. Discriminative nonnegative spectral clustering with out-of-sample extension. *IEEE TKDE*, 25(8):1760–1771, Aug 2013.

[Yang *et al.*, 2014] Yang Yang, Zheng-Jun Zha, Yue Gao, Xiaofeng Zhu, and Tat-Seng Chua. Exploiting web images for semantic video indexing via robust sample-specific loss. *IEEE TMM*, 16(6):1677–1689, 2014.

[Yang *et al.*, 2015] Yang Yang, Zhigang Ma, Yi Yang, Feiping Nie, and Heng Tao Shen. Multitask spectral clustering by exploring intertask correlation. *IEEE TCYB*, 45(5):1083–1094, 2015.

[Ye *et al.*, 2007a] J. Ye, Z. Zhao, and H. Liu. Adaptive distance metric learning for clustering. In *CVPR*, pages 1–7, 2007.

[Ye *et al.*, 2007b] J. Ye, Z. Zhao, and M. Wu. Discriminative k-means for clustering. *NIPS*, 20:1649–1656, 2007.