Mach Learn (2010) 81: 37–51
DOI 10.1007/s10994-010-5201-z

On the eigenvectors of p-Laplacian

Dijun Luo · Heng Huang · Chris Ding · Feiping Nie

Received: 30 April 2010 / Accepted: 20 June 2010 / Published online: 22 July 2010

Abstract Spectral analysis approaches have been actively studied in machine learning and data mining areas, due to their generality, efficiency, and rich theoretical foundations. As a natural non-linear generalization of Graph Laplacian, p-Laplacian has recently been proposed, which interpolates between a relaxation of normalized cut and the Cheeger cut. However, the relaxation can only be applied to two-class cases. In this paper, we propose full eigenvector analysis of p-Laplacian and obtain a natural global embedding for multi-class clustering problems, instead of using greedy search strategy implemented by previous researchers. An efficient gradient descend optimization approach is introduced to obtain the p-Laplacian embedding space, which is guaranteed to converge to feasible local solutions. Empirical results suggest that the greedy search method often fails in many real-world applications with non-trivial data structures, but our approach consistently gets robust clustering results. Visualizations of experimental results also indicate our embedding space preserves the local smooth manifold structures existing in real-world data.

Keywords p-Laplacian · Graph Laplacian · Clustering · Cheeger cut · Normalized cut

1 Introduction

Graph-based methods, such as spectral embedding (Belkin and Niyogi 2001), spectral clustering (Shi and Malik 2000; Belkin and Niyogi 2001), and semi-supervised learning (Zhou et al. 2003; Kulis et al. 2009; Belkin et al. 2004), have recently received much attention from the machine learning community. Due to their generality, efficiency, and rich theoretical foundations (Chung 1997; Belkin and Niyogi 2001; Zhou et al. 2003; Hein et al. 2005; Robles-Kelly and Hancock 2007; Guattery 1998), these methods have been widely explored and applied into various machine learning related research areas, including computer vision

Editors: José L. Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag.

D. Luo · H. Huang (✉) · C. Ding · F. Nie
Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA
e-mail: heng@uta.edu

(Shi and Malik 2000; Jain and Zang 2007; Chen and Lerman 2009), data mining (Jin et al. 2005), speech recognition (Bach and Jordan 2006), social networking (White and Smyth 2005), bioinformatics (Liu et al. 2008), and even commercial usage (Anastasakos et al. 2009; Cheng et al. 2007). More Recently, as a nonlinear generalization of the standard graph Laplacian, graph p-Laplacian starts to attract attentions from machine learning community, such as Bühler et al. (2009) proved the relationship between graph p-Laplacian and Cheeger cuts. Meanwhile, discrete p-Laplacian has also been well studied in mathematics community and solid properties have been investigated by previous work (Amghibech 2003; Allegretto and Huang 1998; Bouchala 2003).

Bühler (2009) provided a rigorous proof of the approximation of the second eigenvector of p-Laplacian to the Cheeger cut. Unlike other graph-based approximation/relaxation techniques (*e.g.* (Ding and He 2005)), the approximation to the optimal Cheeger cut is guaranteed to be arbitrarily exact. This discovery theoretically and practically starts a direction for graph cut based applications. Unfortunately, the p-Laplacian eigenvector problem leads to an untractable optimization, which was solved (see Bühler and Hein 2009) by a somewhat complicated way. Moreover, they only solved the problem for the second eigenvector and provided a direct approach to solve two-class clustering problems. For multi-class problems, they employed hierarchical strategy, which often leads to poor clustering quality in real world data with complicated structures due to its intrinsically greedy property.

Putting the nice theoretical foundations of p-Laplacian and its difficulties together, one might immediately raise a question: can we obtain a full eigenvector space of p-Laplacian, similar to other regular spectral techniques, and easily derive a complete clustering analysis using p-Laplacian? To solve this question, in this paper, we investigate the whole eigenvector space of p-Laplacian and provide (1) an approximation of the whole eigenvectors which lead to a tractable optimization problems, (2) a proof to show that our approximation is very close to the true eigenvector solutions of p-Laplacian, and (3) an efficient algorithm to solve the resulting optimization problems, which is guaranteed to converge to feasible solutions.

After introducing several important research results from mathematics community, we further explore the new properties of the full eigenvector space of p-Laplacian. Our main theoretical contributions are summarized in Theorems 2 and 3. Through our theoretical analysis and practical algorithm, the p-Laplacian based clustering method can naturally and optimally find the cluster structures in multi-class problems. Empirical studies in real world data sets reveal that greedy search often fails in complicated structured data, and our approach consistently obtains high clustering qualities. Visualizations of images data also demonstrate that our approach extracts the intrinsic smooth manifold reserved in the embedding space.

2 Discrete p-Laplacian and eigenvector analysis

Given a set of similarity measurements, the data can be represented as a weighted, undirected graph $G = (V, E)$, where the vertices in V denote the data points and positive edge weights in W encode the similarity of pairwise data points. We denote the degree of node $i \in V$ by $d_i = \sum_j w_{ij}$. Given function $f : V \to \mathcal{R}$, the *p-Laplacian operator* is defined as follows:

$$(\Delta_p^W f)_i = \sum_j w_{ij} \phi_p(f_i - f_j), \tag{1}$$

where $\phi_p(x) = |x|^{p-1}\text{sign}(x)$. Note that $\phi_2(x) = x$, which becomes the standard graph Laplacian. In general, the p-Laplacian is a nonlinear operator. The eigenvector of p-Laplacian is defined as following:

Definition 1 $f : V \to \mathcal{R}$ is an eigenvector of p-Laplacian Δ_p^W, if there exists a real number λ, such that

$$(\Delta_p^W f)_i = \lambda \phi_p(f_i), \quad i \in V. \tag{2}$$

λ is called as eigenvalue of Δ_p^W associated with eigenvector f.

One can easily verify that when $p = 2$, the operator Δ_p^W becomes the regular graph Laplacian $\Delta_2^W = L = D - W$, where D is a diagonal matrix with $D_{ii} = d_i$, and the eigenvectors of Δ_p^W become the eigenvectors of L. The eigenvector of p-Laplacian is also called p-eigenfunction.

2.1 Properties of eigenvalues of p-Laplacian

Proposition 1 (Amghibech 2006) *If W represents a connected graph, and if λ is an eigenvalue of Δ_p^W, then*

$$\lambda \leq 2^{p-1} \max_{i \in V} d_i.$$

This indicates that the eigenvalues of p-Laplacian are bounded by the largest volume. It is easy to check that for connected bipartite regular graph, the equality is achieved.

2.2 Properties of eigenvectors of p-Laplacian

Starting from previous research results on p-Laplacian, we will introduce and prove our main theoretical contributions in Theorems 2 and 3. The eigenvectors of p-Laplacian have the following properties:

Theorem 1 (Bühler and Hein 2009) *f is an eigenvector of p-Laplacian Δ_p^W, if and only if f is a critical point of the following function*

$$F_p(f) = \frac{\sum_{ij} w_{ij} |f_i - f_j|^p}{2\|f\|_p^p}, \tag{3}$$

where

$$\|f\|_p^p = \sum_i |f_i|^p.$$

The above theorem provides an equivalent statement of eigenvector and eigenvalue of p-Laplacian. It also serves as the foundation of analysis of eigenvector. Notice that $F_p(\alpha f) = F_p(f)$ which indicates the following property of p-Laplacian:

Corollary 1 *If f is an eigenvector of Δ_p^W associated with eigenvalue λ, then for any $\alpha \neq 0$, αf is also an eigenvector of Δ_p^W associated with eigenvalue λ.*

Notice that Δ_p^W is not a linear operator, *i.e.* $\Delta_p^W(\alpha f) \neq \alpha \Delta_p^W f$, if $p \neq 2$. However, Corollary 1 shows that the linear transformation of a single eigenvector remains an eigenvector of the p-Laplacian. Also note that $\Delta_p^W f = \Delta_p^W (f + d)$ for any constant vector d. Thus, Δ_p^W is translation invariant, and we have

Corollary 2 *$c\mathbf{1}$ is an eigenvector of Δ_p^W for constant $c \neq 0$, associated with eigenvalue 0, where $\mathbf{1}$ is a column vector with all elements 1 and proper size.*

In the supplement (Lemma 3.2) of Bühler and Hein (2009), authors also provided the following property of the non-trivial eigenvector of p-Laplacian.

Proposition 2 *If f is a non-trivial eigenvector of Δ_p^W, then*

$$\sum_i \phi_p(f_i) = 0. \tag{4}$$

The non-trivial eigenvectors refer to those eigenvectors associated with non-zero eigenvalues. Inspired by the above properties of eigenvectors of p-Laplacian, we propose the following new theoretical analysis on eigenvectors of p-Laplacian.

Definition 2 We call $f \neq \mathbf{0}$ and $g \neq \mathbf{0}$ as p-orthogonal if the following condition holds

$$\sum_i \phi_p(f_i)\phi_p(g_i) = 0. \tag{5}$$

As one of the main results in this paper, the following property of the full eigenvectors of p-Laplacian is proposed,

Theorem 2 *If f and g are two eigenvectors of p-Laplacian Δ_p^W associated with different eigenvalues λ_f and λ_g, and W is symmetric, and $p \geq 1$, then f and g are p-orthogonal up to the second order Taylor expansion.*

Proof By definitions, we have

$$(\Delta_f^W)_i = \lambda_f \phi(f_i), \tag{6}$$

$$(\Delta_g^W)_i = \lambda_g \phi(g_i). \tag{7}$$

Multiplying $\phi_p(g_i)$ and $\phi_p(f_i)$ on both sides of (6) and (7), respectively, we have

$$(\Delta_f^W)_i \phi(g_i) = \lambda_f \phi(f_i)\phi(g_i), \tag{8}$$

$$(\Delta_g^W)_i \phi(f_i) = \lambda_g \phi(g_i)\phi(f_i). \tag{9}$$

By summing over i and taking the difference of both sides of (8) and (9), we get

$$(\lambda_f - \lambda_g) \sum_i \phi_p(f_i)\phi_p(g_i) = \sum_i [(\Delta^W f)_i \phi_p(g_i) - (\Delta^W g)_i \phi_p(f_i)].$$

Notice that for any $p > 1, a, b \in \mathcal{R}$,

$$\begin{aligned} \phi_p(a)\phi_p(b) &= |a|^{p-1}\operatorname{sign}(a)|b|^{p-1}\operatorname{sign}(b) \\ &= |a|^{p-1}|b|^{p-1}\operatorname{sign}(a)\operatorname{sign}(b) \\ &= |ab|^{p-1}\operatorname{sign}(ab) = \phi_p(ab). \end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\sum_i [(\Delta^W f)_i \phi_p(g_i) - (\Delta^W g)_i \phi_p(f_i)] \\
&\quad = \sum_{ij} w_{ij} [\phi_p(f_i - f_j)\phi_p(g_i) - \phi_p(g_i - g_j)\phi_p(f_i)] \\
&\quad = \sum_{ij} w_{ij} [\phi_p(f_i g_i - f_j g_i) - \phi_p(g_i f_i - g_j f_i)].
\end{aligned}$$

Since any constant vector $c\mathbf{1}$ is a valid eigenvector of p-Laplacian, we write $\phi_p(x)$ as

$$\phi_p(x) = \phi_p(c) + \phi_p'(c)(x - c) + o_2,$$

where o_2 is the sum of high order Taylor expansion terms (starting from the second order) at constant c. Note that both $\phi_p(c)$ and $\phi_p'(c)$ are constants. Because of $w_{ij} = w_{ji}$, the above equation becomes

$$\begin{aligned}
&\sum_{ij} w_{ij} [\phi_p(c) + \phi_p'(c)(f_i g_i - f_j g_i - c) \\
&\qquad - \phi_p(c) - \phi_p'(c)(g_i f_i - g_j f_i - c)] + o_2 \\
&\quad = \sum_{ij} w_{ij} [\phi_p'(c)(f_i g_i - g_i f_i) - \phi_p'(c)(f_j g_i - g_j f_i) \\
&\qquad + \phi_p(c) - c\phi_p'(c) - \phi_p(c) + c\phi_p'(c)] + o_2 \\
&\quad = o_2.
\end{aligned}$$

All 0th and 1st order Taylor expansion terms are canceled explicitly. This leads to

$$(\lambda_f - \lambda_g) \sum_i \phi_p(f_i)\phi_p(g_i) \approx 0.$$

Since $\lambda_f \neq \lambda_g$, we have

$$\sum_i \phi_p(f_i)\phi_p(g_i) \approx 0.$$

If $p = 2$, the second order term of Taylor expansion is 0, then the approximately equal becomes exactly equal. This property of p-Laplacian is significant different from those in existing literacy, in the sense that it explores the relationship of the full eigenvectors space. □

Theorem 3 *If $f^{*1}, f^{*2}, \ldots, f^{*n}$ are n eigenvectors of operator Δ_p^W associated with unique eigenvalues $\lambda_1^*, \lambda_2^*, \ldots, \lambda_n^*$, then $f^{*1}, f^{*2}, \ldots, f^{*n}$ are local solution of the following optimization problem*

$$\min_{\mathcal{F}} J(\mathcal{F}) = \sum_k F_p(f^k), \tag{10}$$

$$s.t. \quad \sum_i \phi_p(f_i^k)\phi_p(f_i^l) = 0, \quad \forall k \neq l, \tag{11}$$

where $\mathcal{F} = (f^1, f^2, \ldots, f^n)$.

Proof We do the derivative of $J(\mathcal{F})$ w.r.t. f^k as:

$$\frac{\partial J(\mathcal{F})}{\partial f^k} = \frac{\partial F_p(f^k)}{\partial f^k} = \frac{\partial \frac{\sum_{ij} w_{ij}|f_i^k - f_j^k|^p}{2\|f^k\|_p^p}}{\partial f^k}$$
$$= \frac{1}{\|f^k\|_p^p}\left[\Delta_p^W(f^k) - \frac{\sum_{ij} w_{ij}|f_i^k - f_j^k|^p}{\|f^k\|_p^p}\phi_p(f^k)\right].$$

From Theorem 3.1 in Bühler and Hein (2009),

$$\lambda_k^* = \frac{\sum_{ij} w_{ij}|f_i^{*k} - f_j^{*k}|^p}{\|f^{*k}\|_p^p},$$

and by definition,

$$\Delta_p^W(f^{*k}) - \lambda_k^*\phi_p(f^{*k}),$$

thus we have,

$$\frac{\partial J(\mathcal{F})}{\partial f^{*k}} = 0,$$

and according to Theorem 2, the constraints in (11) are satisfied. Thus $f^{*k}, k = 1, 2, \ldots, n$ are local solutions for (10). □

On the other hand, one can show the following relationship between the Cheeger cut and the second eigenvector of p-Laplacian when $K = 2$.

Definition 3 Given a undirected graph W and a partition of the nodes $\{C_1, C_2, \ldots, C_K\}$, the Cheeger cut of the graph is

$$CC = \sum_{k=1}^{K} \frac{\mathrm{Cut}(C_k, \bar{C}_k)}{\min_{1\le l\le K}|C_l|}, \tag{12}$$

where

$$\mathrm{Cut}(A, B) = \sum_{i\in A, j\in B} W_{ij}, \tag{13}$$

and $\bar{C}_k$ is the complement of $C_k, k = 1, 2, \ldots, K$.

Proposition 3 *Denoted by CC_c^*, the Cheeger cut value is obtained by thresholding the second eigenvector of Δ_p^W, and CC^* is the global optimal value of (12) with $K = 2$, then the following holds*

$$CC^* \le CC_c^* \le p\left(\max_{i\in V} d_i\right)^{\frac{p-1}{p}} (CC^*)^{\frac{1}{p}}. \tag{14}$$

This property of the second eigenvector of Δ_p^W indicates that when $p \to 1$, $CC_c^* \to CC^*$. Notice that this approximation can be achieved arbitrarily accurate, which is different from other relaxation-based spectral clustering approximation. Thus, it opens a total new direction of spectral clustering.

However, this relationship holds only in the case of $K = 2$. In previous research, a greedy search strategy is applied to obtain Cheeger cut results for multi-class clustering (Bühler and Hein 2009). In the algorithm, they first split data in to two parts and recursively dichotomize the data till a desired number of clusters are achieved. In our study, we discover that in many real world data sets, this greedy search strategy isn't efficient and effective. This limitation inspires us to explore the whole eigenvectors of p-Laplacian to obtain better solution of Cheeger cut.

3 Solving complete eigenfunctions for p-Laplacian

In previous section, we derive a single optimization problem for full eigenvectors of p-Laplacian. However, the optimization problem remains intractable. In this section, we propose an approximation algorithm to obtain full eigenvectors of p-Laplacian. We also provide a proof to show how good our approximation is.

3.1 Orthogonal p-Laplacian

Instead of solving (10), we solve the following problem:

$$\min_{\mathcal{F}} J_o(\mathcal{F}) = \sum_k \sum_{ij} w_{ij} |f_i^k - f_j^k|^p, \tag{15}$$

$$\text{s.t.} \quad \mathcal{F}^T \mathcal{F} = I, \quad \|f^k\|_p^p = 1, \quad k = 1, 2, \ldots, n. \tag{16}$$

3.2 The approximation evaluation

Here we show that the approximation is tight. By introducing Lagrangian multiplier, we obtain,

$$\mathcal{L} = \sum_k Q_p^W(f^k) - \operatorname{Tr} \mathcal{F}^T \mathcal{F} \Lambda - \sum_k \xi_k (\|f^k\|_p^p - 1), \tag{17}$$

where $Q_p^W(f) = \sum_{ij} w_{ij} |f_i - f_j|^p$. Taking the derivative of $\mathcal{L}$ w.r.t. f^k and set it to be zeros, we have,

$$p \sum_j w_{ij} \phi_p(f_i^k - f_j^k) - \lambda_k f_i^k - p \xi_k \phi_p(f_i^k) = 0, \quad i = 1, 2, \ldots, n, \tag{18}$$

which leads to

$$\lambda_k = \frac{p[\Delta_p^W(f^k) - \xi_k \phi_f^k]_i}{f_i^k},$$

or

$$\frac{\lambda_k}{\xi_k} = \frac{p[\Delta_p^W(f^k)/\xi_k - \phi_f^k]_i}{f_i^k}. \tag{19}$$

Denote $\eta_i = [\Delta_p^W(f^k)/\xi_k - \phi_f^k]_i$, from Amghibech (2006), we know that η_i is a constant w.r.t. i. Notice that (19) holds for all i, thus $\eta_i \approx 0$, indicating that compared to ξ_k, λ_k can

be ignored. Thus, (18) becomes

$$p\sum_j w_{ij}\phi_p(f_i^k - f_j^k) - p\xi_k\phi_p(f_i^k) = 0, \quad i = 1, 2, \ldots, n,$$

and by definition, f^k is an eigenvector of Δ_p^W associate with eigenvalue ξ_k.

4 p-Laplacian embedding

Since $F_p(f) = F_p(\alpha f)$ for $\alpha \neq 0$, we can always scale f without any change. Thus, we propose the following p-Laplacian Embedding problem.

$$\min_{\mathcal{F}} J_E(\mathcal{F}) = \sum_k \frac{\sum_{ij} w_{ij}|f_i^k - f_j^k|^p}{\|f^k\|_p^p}, \tag{20}$$

$$\text{s.t.} \quad \mathcal{F}^T\mathcal{F} = I. \tag{21}$$

4.1 Optimization

The gradient of J_E w.r.t. f_i^k can be written as,

$$\frac{\partial J_E}{\partial f_i^k} = \frac{1}{\|f^k\|_p^p}\left[\sum_j w_{ij}\phi_p(f_i^k - f_j^k) - \frac{\phi_p(f_i^k)}{\|f^k\|_p^p}\right]. \tag{22}$$

If we simply use the gradient descend approach, the solution f^k might not be orthogonal. We modify the gradient as following to enforce the orthogonality,

$$\frac{\partial J_E}{\partial \mathcal{F}} \leftarrow \frac{\partial J_E}{\partial \mathcal{F}} - \mathcal{F}\left(\frac{\partial J_E}{\partial \mathcal{F}}\right)^T \mathcal{F}.$$

We summarize the p-Laplacian embedding algorithm in Algorithm 1.
The parameter α is the step length, which is set to be

$$\alpha = 0.01\frac{\sum_{ik}|\mathcal{F}_{ik}|}{\sum_{ik}|G_{ik}|}.$$

Input: Pairwise graph similarity W, number of embedding dimension K
Output: Embedding space $\mathcal{F}$
Compute $L = D - W$, where D is a diagonal matrix with $D_{ii} = d_i$.
Compute eigenvector decomposition of L: $L = USU^T$,
Initialize $\mathcal{F} \leftarrow U(:, 1:K)$
while not converged do
  $G \leftarrow \frac{\partial J_E}{\partial \mathcal{F}} - \mathcal{F}(\frac{\partial J_E}{\partial \mathcal{F}})^T\mathcal{F}$, where $\frac{\partial J_E}{\partial \mathcal{F}}$ is computed using (22)
  $F \leftarrow F - \alpha G$.
end

Algorithm 1: The p-Laplacian embedding algorithm

One can easily see that if $\mathcal{F}^T\mathcal{F} = I$, then using the simple gradient descend approach can guarantee to give a feasible solution. More explicitly, we have the following theorem:

Theorem 4 *The solution obtained from Algorithm 1 satisfies the constraint in* (21).

Proof Since Laplacian L is symmetric, we have $\mathcal{F}^T\mathcal{F} = I$ for initialization, and

$$\begin{aligned} &G^T\mathcal{F}^t + \left(\mathcal{F}^t\right)^T G \\ &= \left(\frac{\partial J_E}{\partial \mathcal{F}} - \mathcal{F}\left(\frac{\partial J_E}{\partial \mathcal{F}}\right)^T \mathcal{F}\right)^T \mathcal{F}^t + \left(\mathcal{F}^t\right)^T \left[\frac{\partial J_E}{\partial \mathcal{F}} - \mathcal{F}\left(\frac{\partial J_E}{\partial \mathcal{F}}\right)^T \mathcal{F}\right] \\ &= \left(\frac{\partial J_E}{\partial \mathcal{F}}\right)^T \mathcal{F}^t - \left(\mathcal{F}^t\right)^T \frac{\partial J_E}{\partial \mathcal{F}} - \left(\frac{\partial J_E}{\partial \mathcal{F}}\right)^T \mathcal{F}^t + \left(\mathcal{F}^t\right)^T \frac{\partial J_E}{\partial \mathcal{F}} \\ &= 0. \end{aligned}$$

By Algorithm 1 we have,

$$\mathcal{F}^{t+1} = \mathcal{F}^t - \alpha G.$$

Thus

$$(\mathcal{F}^{t+1})^T\mathcal{F}^{t+1} = (\mathcal{F}^t - \alpha G)^T(\mathcal{F}^t - \alpha G) = (\mathcal{F}^t)^T\mathcal{F}^t - \alpha[G^T\mathcal{F}^t + (\mathcal{F}^t)^T G] = I.$$

□

This technique is a special case of *Natural Gradient*, which can be found in Amari (1998). Since $J_E(\mathcal{F})$ is bounded as $J_E(\mathcal{F}) \geq 0$, our algorithm also has the following obvious property:

Theorem 5 *Algorithm 1 is guaranteed to converge.*

5 Experimental results

In this section, we will evaluate the efficiency of our proposed p-Laplacian Embedding algorithm. To demonstrate the results, we use eight benchmark data sets: AT&T, MNIST, PIE, UMIST, YALEB, ECOLI, GLASS, and DERMATOLOGY.

5.1 Data set descriptions

In the AT&T database[1], there are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expression, and facial details. All images were taken against a dark homogeneous back-ground with the subjects in an upright, frontal, position (with tolerance for some side movement).

MNIST hand-written digits data set consists of 60,000 training and 10,000 test digits (Cun et al. 1998). The MNIST data set can be downloaded from website[2] with 10 classes,

[1] See http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[2] See http://yann.lecun.com/exdb/mnist/.

Table 1 Detailed information of data sets used in our experiments

Data set	#samples	#Attribute	#class
AT&T	400	644	40
MNIST	150	784	10
PIE	680	1,024	68
UMIST	360	644	20
YALEB	1,984	2,016	31
ECOLI	336	343	8
GLASS	214	9	6
DERMATOLOGY	366	34	6

from digit "0" to "9". In the MNIST data set, each image is centered (according to the center of mass of the pixel intensities) on a 28×28 grid. We select 15 images for each digit in our experiment.

UMIST faces is for multi-view face recognition, which is challenging in computer vision because the variations between the images of the same face in viewing direction are almost always larger than image variations in face identity. This data set contains 20 persons with 18 images for each. All these images of UMIST database are cropped and resized into 28×23 images. Due to the multi-view characteristics, the images shall lie in a smooth manifold. We further use this data set to visually test our embedding smoothness.

CMU PIE face database contains 68 subjects with 41,368 face images. Preprocessing to locate the faces was applied. Original images were normalized (in scale and orientation) such that two eyes were aligned at the same position. Then, the facial areas were cropped into the final images for matching. The size of each cropped image is 64×64 pixels, with 256 grey levels per pixel. In our experiment, we randomly pick 10 different combinations of pose, face expression, and illumination condition. Finally we have $68 \times 10 = 680$ images.

Another images benchmark used in our experiment is the combination of extended and original Yale database (Georghiades et al. 2001). These two databases contain single light source images of 38 subjects (10 subjects in original database and 28 subjects in extended one) under 576 viewing conditions (9 poses $\times$ 64 illumination conditions). Thus, for each subject, we got 576 images under different lighting conditions. The facial areas were cropped into the final images for matching (Georghiades et al. 2001). The size of each cropped image in our experiments is 192×168 pixels, with 256 gray levels per pixel. We randomly pick up 20 images for each person and also sub-sample the images down to 48×42. To visualize the quality of the embedding space, we pickup the images such that they come from different illumination conditions.

Three other data sets (ECOLI, GLASS, and DERMATOLOGY) come from UCI Repository (Asuncion and Newman 2007). The detailed information of eight benchmark data sets can be found in Table 1.

For all data sets used in our experiments, we directly use the original space without any processing. More specifically, for images data sets, we use the raw gray level values as features.

5.2 Experimental settings

We construct the pairwise similarity of data points as follows.

$$W_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{r_i r_j}\right), & x_i, x_j \text{ are neighbors} \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

where r_i and r_j are the average distances of K-nearest neighbors of data points i and j, respectively. K is set to 10 in all our experiments, which is the same as in Bühler and Hein (2009). By neighbors here we mean x_i is a K-nearest neighbors of x_j or x_j is a K-nearest neighbors of x_i.

For our method (Cheeger cut Embedding or CCE), we first obtain the embedding space using Algorithm 1 Then a standard K-means algorithm is applied to further determine the clustering assignments. For visualization, we use the second and third eigenvectors as the x-axis and y-axis, respectively.

In direct comparison and succinct presentation, we compare our results to greedy search Cheeger cut algorithm (Bühler and Hein 2009) in terms of three clustering quality measurements. We download their codes and directly use them with default settings. For both methods, we set $p = 1.2$, which is suggested in previous research (Bühler and Hein 2009).

5.3 Measurements

We use three metrics to measure the performance in our experiments: the value of objective in (3), the Cheeger cut defined in (12), and clustering accuracy. Clustering accuracy (ACC) is defined as:

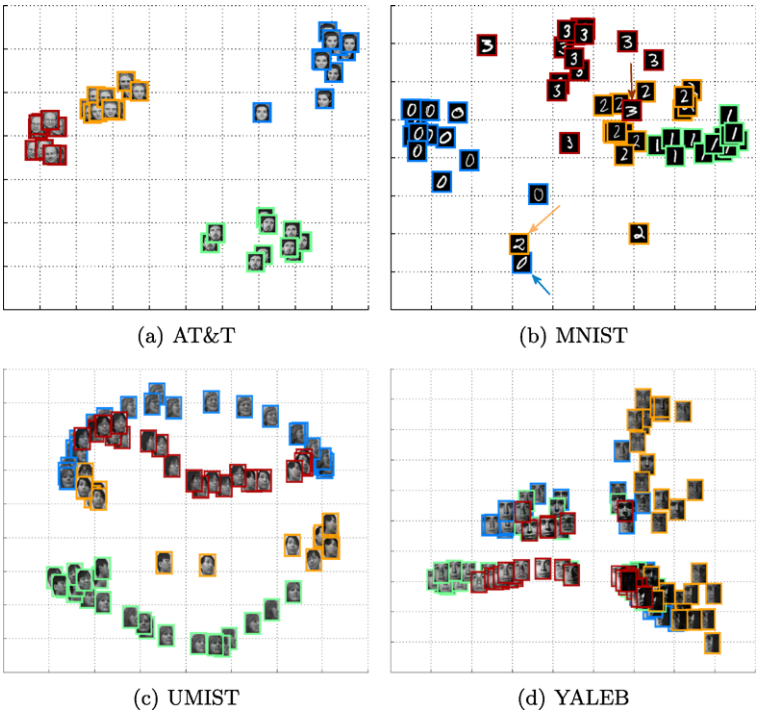
$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} \delta(l_i, \mathrm{map}(c_i))}{n}, \tag{24}$$

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. And a lower value of objective in (3) or lower Cheeger cut suggests better clustering quality.

5.4 Evaluation results

Embedding results We use 4 data sets (AT&T, MNIST, UMIST, YALEB) to visualize the embedding results obtained by our method. For each data set, we select samples in four different clusters. We use the second and third eigenvector as x-axis and y-axis, respectively. The embedding results are shown in Fig. 1(a)–(d). For AT&T data, the four persons are well separated. For MNIST data, the four digits are separated in most of the images. Three images ("3", "2", and "0" as highlighted in Fig. 1(b)) are visually different from other images of the same group. The embedding results also show that these three images are far way from the other objects in the same group. This result indicates that our embedding space reserves the visual characteristics. For UMIST and YALEB data, since the images from the same group are taken under different face expression or illumination conditions, they are arranged in a smooth manifold. This structure also remains in our embedding space, see Fig. 1(c) and (d).

Clustering analysis on confusion matrices We select 10 groups for AT&T, MNIST, PIE, UMIST, and YALEB, 6 for GLASS and DERMATOLOGY, and 8 for ECOLI. We compare the greedy search Cheeger cut (Bühler and Hein 2009) (GSCC) to our method (CCE). The confusion matrices are shown in Fig. 2. In AT&T, MNIST, and ECOLI data, our method



(a) AT&T (b) MNIST

(c) UMIST (d) YALEB

Fig. 1 Embedding results on four image data sets using the second and third eigenvectors of p-Laplacian as x-axis and y-axis, respectively, where $p = 1.2$. Different colors indicate different groups according to ground truth. In (**b**) the highlighted are images which are visually far away from other images in the same group

obviously outperforms GSCC, because the diagonals of our confusion matrices are much stronger than those in GSCC results.

Clustering quality analysis We use three metrics mentioned above to measure the quality of clustering results. We compare our method to greedy search Cheeger cut in various experimental settings. For AT&T, MNIST, PIE, and UMIST, we choose $k = 2, 3, 4, 5, 6, 8, 10$, where k is the number of clusters. Typically, a larger k leads to a more difficult clustering task and a lower clustering accuracy. For ECOLI, GLASS, YALEB, and DERMATOLOGY data, we set $k = 2, 3, 4, 5, 6, 8$, $k = 2, 3, 4, 5, 6, 7$, $k = 2, 4, 5, 6, 8, 10$ and $k = 2, 3, 4, 5, 6$, respectively. We set these numbers of k according to the size of the original data sets and also for convenient presentation. All results are shown in Table 2. Notice that for greedy search, if $k > 2$, there is no way to calculate the objective function values defined in (3).

In Table 2, when the data set is simple (*i.e.* k is small), the accuracy of the two methods is close to each other. However, if the data is complex (*i.e.* when k is large), our method has much better clustering results than greedy search. For example, in AT&T, when $k = 10$, our approach remains high (78%) in clustering accuracy, while greedy search only achieves 38%. Also we can see that when k is large, our algorithm obtains much lower values in both objective and Cheeger cut than greedy search. One should notice that the setting of MNIST

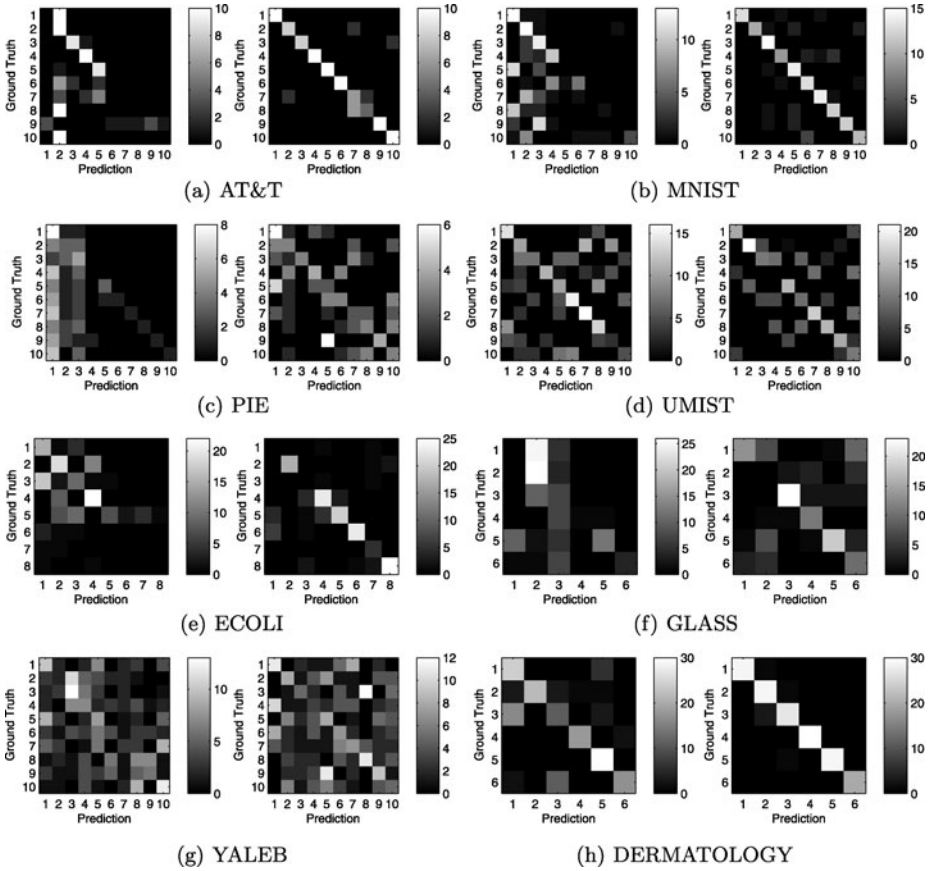


Fig. 2 Comparisons of confusion matrices of GSCC (*left* in *each panel*) and our CCE (*right* in *each panel*) on 8 data sets. *Each column* of the matrix represents the instances in a predicted class, and *each row* represents the instances in an actual class

data used in our experiment is different from the one used in previous research (Bühler and Hein 2009).

6 Conclusions

Spectral data analysis is important in machine learning and data mining areas. Unlike other relaxation-based approximation techniques, the solution obtained by p-Laplacian can approximate the global solution arbitrarily tight. Meanwhile, Cheeger cut favors the solutions which are more balanced. This paper is the first one to offer a full eigenvector analysis of p-Laplacian. We proposed an efficient gradient descend approach to solve the full eigenvector problem. Moreover, we provided new analysis of the properties of eigenvectors of p-Laplacian. Empirical studies show that our algorithm is much more robust in real world data sets clustering than the previous greedy search p-Laplacian spectral clustering. Therefore, both theoretical and practical results proposed by this paper introduce a promising direction to machine learning community and related applications.

Table 2 Clustering quality comparison of greedy search Cheeger cut and our method. Obj is the objective function value defined in (3), CC is the Cheeger cut objective defined in (12), and Acc is the clustering accuracy defined in (24). For greedy search, objective in (3) is not provided when $k > 2$

AT&T	Greedy search			Our method			MNIST	Greedy search			Our method		
k	Obj	CC	Acc	Obj	CC	Acc	k	Obj	CC	Acc	Obj	CC	Acc
2	31.2	29.1	100.0	31.2	29.1	100.0	2	46.0	42.9	100.0	46.0	42.9	100.0
3	–	124.9	80.0	187.4	91.3	100.0	3	–	182.9	56.0	268.1	132.3	98.0
4	–	385.9	60.0	333.2	176.4	100.0	4	–	534.0	47.0	459.7	252.9	97.0
5	–	1092.5	50.0	500.7	306.6	90.0	5	–	1129.9	45.0	680.7	402.8	92.0
6	–	3034.6	45.0	673.3	436.8	73.0	6	–	6356.0	40.0	923.2	582.7	89.0
8	–	5045.6	36.0	1134.9	862.6	80.0	8	–	10785.6	33.0	1608.9	1110.4	86.0
10	–	8012.7	38.0	1712.8	1519.7	78.0	10	–	16555.5	35.0	2461.2	1731.2	85.0
PIE	Greedy search			Our method			UMIST	Greedy search			Our method		
k	Obj	CC	Acc	Obj	CC	Acc	k	Obj	CC	Acc	Obj	CC	Acc
2	38.4	31.0	60.0	38.4	38.4	65.0	2	86.2	80.8	57.0	86.2	80.4	57.0
3	–	144.0	50.0	189.1	112.8	67.0	3	–	269.1	42.0	388.4	193.2	58.0
4	–	514.3	43.0	324.3	179.6	60.0	4	–	732.3	42.0	669.6	399.8	62.0
5	–	2224.7	34.0	477.1	336.6	58.0	5	–	973.5	42.0	1026.1	639.1	57.0
6	–	3059.9	28.0	673.6	489.7	62.0	6	–	1888.4	33.0	1426.0	898.2	60.0
8	–	5362.7	24.0	1146.1	985.9	45.0	8	–	8454.9	29.0	2374.8	1939.8	55.0
10	–	7927.2	22.0	1707.4	1776.3	45.0	10	–	4204.1	34.0	3793.1	2673.7	54.0
ECOLI	Greedy search			Our method			GLASS	Greedy search			Our method		
k	Obj	CC	Acc	Obj	CC	Acc	k	Obj	CC	Acc	Obj	CC	Acc
2	70.7	65.7	97.0	70.7	67.8	98.0	2	71.4	85.3	63.0	71.4	111.3	63.0
3	–	189.5	73.0	318.3	180.6	89.0	3	–	259.5	39.0	386.4	198.8	66.0
4	–	529.0	72.0	458.4	306.0	84.0	4	–	821.3	48.0	617.9	389.6	71.0
5	–	738.1	57.0	790.0	566.7	77.0	5	–	7659.5	43.0	862.7	498.9	57.0
6	–	1445.6	61.0	1083.5	993.1	80.0	6	–	12160.9	38.0	1253.7	814.5	62.0
8	–	16048.2	58.0	1969.5	1736.0	79.0	7	–	12047.2	37.0	1253.7	816.3	62.0
YALEB	Greedy search			Our method			DERMA	Greedy search			Our method		
k	Obj	CC	Acc	Obj	CC	Acc	k	Obj	CC	Acc	Obj	CC	Acc
2	73.5	68.6	50.0	73.5	68.6	50.0	2	74.1	69.1	100.0	74.1	69.1	100.0
4	–	425.7	33.0	740.1	405.0	39.0	3	–	275.0	78.0	426.7	192.0	100.0
6	–	1642.4	27.0	2300.6	1534.5	34.0	4	–	493.8	81.0	770.9	403.3	95.0
8	–	3953.5	27.0	3044.4	2109.8	31.0	5	–	1120.5	77.0	1206.6	661.2	96.0
10	–	4911.5	24.0	4690.6	3368.1	28.0	6	–	2637.5	45.0	1638.6	1115.1	96.0

References

Allegretto, W., & Huang, Y. X. (1998). A picone's identity for the p-Laplacian and applications. *Nonlinear Analysis*, *32*, 819–830.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*, 251–276.

Amghibech, S. (2003). Eigenvalues of the discrete p-Laplacian for graphs. *Ars Comb*, *67*, 283–302.

Amghibech, S. (2006). Bounds for the largest p-Laplacian eigenvalue for graphs. *Discrete Mathematics*, *306*, 2762–2771.

Anastasakos, T., Hillard, D., Kshetramade, S., & Raghavan, H. (2009). A collaborative filtering approach to ad recommendation using the query-ad click graph. In D. W. L. Cheung, I. Y. Song, W. W. Chu, Hu, X., & J. J. Lin (Eds.), *CIKM* (pp. 1927–1930). New York: ACM.

Asuncion, A., & Newman, D. (2007). UCI machine learning repository.

Bach, F. R., & Jordan, M. I. (2006). Learning spectral clustering, with application to speech separation. *Journal of Machine Learning Research*, *7*, 1963–2001.

Belkin, M., & Niyogi (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS* (Vol. 14, pp. 585–591). Cambridge: MIT Press.

Belkin, Matveeva, & Niyogi (2004). Regularization and semi-supervised learning on large graphs. In *COLT: Proceedings of the workshop on computational learning theory*. San Mateo: Morgan Kaufmann.

Bouchala, J. (2003). Resonance problems for p-Laplacian. *Mathematics and Computers in Simulation*, *61*, 599–604.

Bühler, T., & Hein, M. (2009). Spectral clustering based on the graph p-Laplacian. In *ICML* (Vol. 382, pp. 81–88). New York: ACM.

Chen, G., & Lerman, G. (2009). Spectral curvature clustering (SCC). *International Journal of Computer Vision*, *81*, 317–330.

Cheng, H., Tan, P. N., Sticklen, J., & Punch, W. F. (2007). Recommendation via query centered random walk on K-partite graph. In *ICDM* (pp. 457–462). New York: IEEE Computer Society.

Chung, F. (1997). *Spectral graph theory*. Providence: AMS.

Cun, Y. L. L., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE*, *86*, 2278–2324.

Ding, C. H. Q., & He, X. (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*.

Georghiades, A., Belhumeur, P., & Kriegman, D. (2001). From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*, 643–660.

Guattery, Miller (1998). On the quality of spectral separators. *SIAM Journal on Matrix Analysis and Applications*, *19*.

Hein, Audibert, & von Luxburg (2005). From graphs to manifolds—weak and strong pointwise consistency of graph Laplacians. In Auer, P., Meir, R. (Eds.), *Proc. of the 18th conf. on learning theory (COLT)* (pp. 486–500). Berlin: Springer.

Jain, V., & Zang, H. (2007). A spectral approach to shape-based retrieval of articulated 3D models. *Computer-Aided Design*, *39*, 398–407.

Jin, R., Ding, C. H. Q., & Kang, F. (2005). A probabilistic approach for optimizing spectral clustering.

Kulis, B., Basu, S., Dhillon, I. S., & Mooney, R. J. (2009). Semi-supervised graph clustering: a kernel approach. *Machine Learning*, *74*, 1–22.

Liu, Y., Eyal, E., & Bahar, I. (2008). Analysis of correlated mutations in HIV-1 protease using spectral clustering. *Bioinformatics*, *24*, 1243–1250.

Robles-Kelly, A., & Hancock, E. R. (2007). A Riemannian approach to graph embedding. *Pattern Recognition*, *40*.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 888–905.

White, S., & Smyth, P. (2005). A spectral clustering approach to finding communities in graph. In *SDM*.

Zhou, D. B. O., Lal, T. N., Weston, J., & Schölkopf, B. (2003). Learning with local and global consistency. In *NIPS* (Vol. 16, pp. 321–328). Cambridge: MIT Press.