Adaptive Consistency Propagation Method for Graph Clustering

Xuelong Li, *Fellow, IEEE,* Mulin Chen, and Qi Wang, *Senior Member, IEEE*

Abstract—Graph clustering plays an important role in data mining. Based on an input data graph, data points are partitioned into clusters. However, most existing methods keep the data graph fixed during the clustering procedure, so they are limited to exploit the implied data manifold and highly dependent on the initial graph construction. Inspired by the recent development on manifold learning, this paper proposes an Adaptive Consistency Propagation (ACP) method for graph clustering. In order to utilize the features captured from different perspectives, we further put forward the Multi-view version of the ACP model (MACP). The main contributions are threefold: (1) the manifold structure of input data is sufficiently exploited by propagating the topological connectivities between data points from near to far; (2) the optimal graph for clustering is learned by taking graph learning as a part of the optimization procedure; (3) the negotiation among the heterogeneous features is captured by the multi-view clustering model. Extensive experiments on real-world datasets validate the effectiveness of the proposed methods on both single- and multi-view clustering, and show their superior performance over the state-of-the-arts.

Index Terms—Clustering, Manifold Learning, Graph Learning, Consistency Propagation

1 Introduction

Clustering is a fundamental task in the field of data mining with various applications, and has attracted many researchers in the past several decades. The objective of clustering is to divide the data points into different clusters. To achieve this goal, plenty of methods have been proposed, including k-means clustering [1], hierarchical clustering [2], spectral clustering [3], spectral embedded clustering [4], maximum margin clustering [5], support vector clustering [6], normalized cut [7], multi-view clustering [8], Nonnegative Matrix Factorization [9], *etc*. Among the existing approaches, graph-based clustering methods (e.g., Ratiocut [10], Normalized-cut [7]) have achieved relatively good performance because of the utilization of manifold information, and been widely used in various applications, such as image segmentation [11] and protein sequence clustering [12].

Most of existing graph-based clustering methods [3], [13], [14], [15], [16], [17] first construct a data graph according to the pairwise similarities of points, and then perform graph-theoretic optimization on the data graph. The two-stage processing brings three major drawbacks. First, in the data graph, the similarity is large only for the neighbors. However, for data with manifold structure, the far away points may also keep high consistency if they are linked by consecutive neighbors. Therefore, these methods are limited to discover the underlying data structure. Second, once the data graph is constructed, they are fixed during the clustering. Then traditional methods are unable to learn the optimal graph for clustering, and tend to fail if the initial graph is constructed with low quality. Third, the graph-theoretic optimization can not produce the clustering results directly, so a post-processing (e.g., k-means) has to be followed, which makes the result deviated from the optimal solution. More recently, some methods [18], [19], [20], [21] are proposed to tackle the last two problems. During the optimization stage, they update the data graph adaptively. In this way, graph learning is successfully integrated into the clustering procedure. Benefited from graph learning, these method are more robust to the initial graph quality. However, these methods still suffer from the first problem.

In this paper, a new graph clustering method, namely Adaptive Consistency Propagation (ACP) is developed to tackle the above issues. The multi-view version of the ACP method is also developed to deal with the data obtained from different feature extractors. The main contributions of this study are summarized as follows.

(1) The topological consistency of points are fully captured to investigate the data manifold. By propagating the consistency through neighbors, the proposed method is suitable to handle data with manifold structures.

(2) Graph learning is jointly combined into the clustering framework. The data graph is optimized adaptively in the optimization stage, so the clustering is less affected by the quality of the initial graph.

(3) An multi-view version of the proposed model is designed, which learns the correlation between the multi-view data and integrates them with the optimal combination.

The rest of this paper is organized as follows. Section 2 introduces the Adaptive Consistency Propagation method, and describes an efficient alternative algorithm to optimize the proposed problem. Section 5 provides the experimental results on several datasets. The parameter sensitivity are discussed in Section 6. And the conclusions are summarized in Section 7.

Notations: Throughout the paper, vectors are written

• *The authors are with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China. Email: xuelong_li@nwpu.edu.cn; chenmulin001@gmail.com; crabwq@gmail.com. Q. Wang is the corresponding author.*

as lowercase, and matrices are written as uppercase. For matrix $\mathbf{M}$, the (i,j)-th element of $\mathbf{M}$ is denoted by $\mathbf{M}_{ij}$, and the trace of $\mathbf{M}$ is denoted by $\mathrm{Tr}(\mathbf{M})$. For vector $\mathbf{z}$, its i-th element is denoted as $\mathbf{z}_i$, and the L2-norm of $\mathbf{z}$ is denoted by $||\mathbf{z}||_2$. $\mathbf{M} \geq 0$ and $\mathbf{z} \geq 0$ mean that all the elements of $\mathbf{M}$ and $\mathbf{z}$ are equal to or larger than zero. The transpose of $\mathbf{M}$ and $\mathbf{z}$ are denoted by $\mathbf{M}^T$ and $\mathbf{z}^T$ respectively. $\mathbf{I}$ indicates the identity matrix.

2 Preliminary

For a better clustering performance, it's essential to investigate the manifold structure implied within the input data [22]. Recently, a Propagation-Based Manifold Learning (PBML) method [23] is presented for crowd motion detection, and has shown satisfying performance. In this section, the PBML method is first revisited as a preliminary.

2.1 Propagation-Based Manifold Learning Method Revisited

The Propagation-Based Manifold Learning method (PBML) proposed by Wang et al. [23] aims to learn the topological relationship of individuals in crowd scene. Given a similarity graph $\mathbf{G} \in \mathbb{R}^{n\times n}$ (n is the number of individuals) of individuals, PBML assumes that individuals with large similarity should share similar topological relevance to any other point. And the objective function of PBML is defined as

$$\min_{\mathbf{S}} \frac{1}{2} \sum_{i,j,k=1}^{n} \mathbf{G}_{jk}(\mathbf{S}_{ij} - \mathbf{S}_{ik})^2 + \beta||\mathbf{S} - \mathbf{I}||_F^2, \tag{1}$$

where $\mathbf{S} \in \mathbb{R}^{n\times n}$ is the desired topological relationship matrix. In the above equation, the first term ensures that individual j and k share similar topological relationship with individual i if j and k are similar. And the second term prevents the trivial solution, where all the elements in $\mathbf{S}$ share the same value.

According to Eq. (1), the topological consistency is propagated through neighbors with high similarities, then the far away individuals will keep close relationship if they are linked by consecutive neighbors. And it performs well on crowd motion detection. However, the obtained topological relationship matrix $\mathbf{S}$ does not indicate the explicit cluster structures, so post-processing is necessary to divide the points into clusters. Moreover, it can not utilize the prior of cluster number, which is always given in the graph clustering literature. Thus, the PBML method can not be used for clustering tasks directly.

3 Adaptive Consistency Propagation

In this section, we extend PBML method to the domain of graph clustering and propose the Adaptive Consistency Propagation (ACP) method.

3.1 Methodology

As pointed out by Mohar et al. [24], the graph $\mathbf{S} \in \mathbb{R}^{n\times n}$ will contain exactly c connected components if the rank of its Laplacian matrix $\mathbf{L}_S$ is $n - c$. Supposing the number of points is n, and the desired cluster number is c, the data graph $\mathbf{S}$ can be considered as an indicator matrix, where the points from the same cluster are connected into the same component. According to the recent graph clustering methods [18], [19], [20], [21], if we impose the constraint $\mathrm{rank}(\mathbf{L}_S) = n - c$ on Eq. (1), the clustering task can be accomplished once the optimal $\mathbf{S}$ is obtained, without the need of performing post-processing. So the objective can be defined as

$$\begin{aligned} &\min_{\mathbf{S}} \frac{1}{2} \sum_{i,j,k=1}^{n} \mathbf{G}_{jk}(\mathbf{S}_{ij} - \mathbf{S}_{ik})^2 + \beta||\mathbf{S} - \mathbf{I}||_F^2, \\ &\text{s.t.} \mathrm{rank}(\mathbf{L}_S) = n - c, \mathbf{S} \geq 0, \sum\nolimits_j \mathbf{S}_{ij} = 1 \end{aligned} \tag{2}$$

where $\mathbf{L}_S$ is the Laplacian matrix of $\mathbf{S}$. Then the cluster number prior c is successfully utilized in problem (2). We also constrain that the sum of each row of $\mathbf{S}$ is one, and each element of $\mathbf{S}$ is non-negative. In problem (2), if point j is connected with many similar neighbors, it will affect the objective value to a large extent. In order to treat each point equally, we propose normalized version of Eq. (2) as follows:

$$\begin{aligned} &\min_{\mathbf{S}} \frac{1}{2} \sum_{i,j,k=1}^{n} \mathbf{G}_{jk}(\frac{\mathbf{S}_{ij}}{\sqrt{\mathbf{D}_{jj}}} - \frac{\mathbf{S}_{ik}}{\sqrt{\mathbf{D}_{kk}}})^2 + \beta||\mathbf{S} - \mathbf{I}||_F^2, \\ &\text{s.t.} \mathrm{rank}(\mathbf{L}_S) = n - c, \mathbf{S} \geq 0, \sum\nolimits_j \mathbf{S}_{ij} = 1, \end{aligned} \tag{3}$$

where $\mathbf{D}$ is the degree matrix of $\mathbf{G}$.

Problem (3) is difficult to solve since the rank constraint depends on $\mathbf{S}$. As Nie et al. [18] pointed out, $\mathrm{rank}(\mathbf{L}_S) = n - c$ is equivalent to $\sum_{i=1}^{c} \sigma_i(\mathbf{L}_S) = 0$, where $\sigma_i(\mathbf{L}_S)$ is the i-th smallest eigenvalue of $\mathbf{L}_S$. Then problem (3) is transformed into

$$\begin{aligned} &\min_{\mathbf{S}} \frac{1}{2} \sum_{i,j,k=1}^{n} \mathbf{G}_{jk}(\frac{\mathbf{S}_{ij}}{\sqrt{\mathbf{D}_{jj}}} - \frac{\mathbf{S}_{ik}}{\sqrt{\mathbf{D}_{kk}}})^2 + \beta||\mathbf{S} - \mathbf{I}||_F^2 \\ &\quad + 2\lambda \sum_{i=1}^{c} \sigma_i(\mathbf{L}_S), \\ &\text{s.t.} \mathbf{S} \geq 0, \sum\nolimits_j \mathbf{S}_{ij} = 1, \end{aligned} \tag{4}$$

where λ is a large enough parameter to enforce $\sum_{i=1}^{c} \sigma_i(\mathbf{L}_S) = 0$.

With Ky Fan's Theorem [25], we have

$$\sum\nolimits_{i=1}^{c} \sigma_i(\mathbf{L}_S) = \min_{\mathbf{F}\in\mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F}=\mathbf{I}} \mathrm{Tr}(\mathbf{F}^T\mathbf{L}_S\mathbf{F}), \tag{5}$$

where $\mathbf{F}$ is an orthonormal vector that minimizes the value of $\mathrm{Tr}(\mathbf{F}^T\mathbf{L}_S\mathbf{F})$.

Combining Eq. (4) and Eq. (5), we get the following problem

$$\begin{aligned} &\min_{\mathbf{S},\mathbf{F}} \frac{1}{2} \sum_{i,j,k=1}^{n} \mathbf{G}_{jk}(\frac{\mathbf{S}_{ij}}{\sqrt{\mathbf{D}_{jj}}} - \frac{\mathbf{S}_{ik}}{\sqrt{\mathbf{D}_{kk}}})^2 + \beta||\mathbf{S} - \mathbf{I}||_F^2 \\ &\quad + 2\lambda \mathrm{Tr}(\mathbf{F}^T\mathbf{L}_S\mathbf{F}), \\ &\text{s.t.} \mathbf{S} \geq 0, \sum\nolimits_j \mathbf{S}_{ij} = 1, \mathbf{F} \in \mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F} = \mathbf{I}, \end{aligned} \tag{6}$$

which is much easier to solve compared with problem (4).

In Eq. (6), the graph $\mathbf{S}$ propagates the consistency through neighbors, and pulls the far away points together if they are linked by the similar points. Moreover, the cluster structure is represented explicitly in $\mathbf{S}$. So the desired $\mathbf{S}$

can be treated as the cluster indicator. Once the optimal $\mathbf{S}$ is learned, the final clustering results are obtained. In the following part, an alternating approach is proposed to solve problem (6) and learn the optimal $\mathbf{S}$ adaptively.

3.2 Optimization Strategy

In this work, the input data graph $\mathbf{G}$ is firstly constructed with an efficient method [18], which builds a sparse and scale invariant affinity matrix according to the points' Euclidean distances. In problem (6), both $\mathbf{S}$ and $\mathbf{F}$ need to be optimized. Then we propose to fix one variable and solve another one iteratively.

Fix S update F

When $\mathbf{S}$ is fixed, problem (6) becomes

$$\min_{\mathbf{F}\in\mathbb{R}^{n\times c},\mathbf{F}^T\mathbf{F}=\mathbf{I}} \mathrm{Tr}(\mathbf{F}^T\mathbf{L}_S\mathbf{F}). \tag{7}$$

Because $\mathbf{F}$ is orthogonal, we have $\min \mathrm{Tr}(\mathbf{F}^T\mathbf{L}_S\mathbf{F})$ equal to the sum of the c smallest eigenvalues of $\mathbf{L}_S$. Therefore, the optimal $\mathbf{F}$ is consist of the c eigenvectors of $\mathbf{L}_S$ associated with the c smallest eigenvalues.

Fix F update S

According to spectral analysis theory [3], we have

$$2\mathrm{Tr}(\mathbf{F}^T\mathbf{L}_S\mathbf{F}) = \sum_{i,j} ||\mathbf{f}_i - \mathbf{f}_j||_2^2\mathbf{S}_{ij}, \tag{8}$$

where $\mathbf{f}_i \in \mathbb{R}^{n\times 1}$ is a vector with its j-th element equal to $\mathbf{F}_{ij}$. Then the objective function (6) becomes

$$\begin{aligned}\min_{\mathbf{S}} \sum_{i=1}^{n}[\frac{1}{2}\sum_{j,k=1}^{n}\mathbf{G}_{jk}(\frac{\mathbf{S}_{ij}}{\sqrt{\mathbf{D}_{jj}}} - \frac{\mathbf{S}_{ik}}{\sqrt{\mathbf{D}_{kk}}})^2 + \beta\sum_{j=1}^{n}(\mathbf{S}_{ij}-\mathbf{I}_{ij})^2 \\ + \lambda\sum_{j=1}^{n}||\mathbf{f}_i - \mathbf{f}_j||_2^2\mathbf{S}_{ij}], \\ s.t. \mathbf{S}_{ij}\geq 0, \sum\nolimits_j \mathbf{S}_{ij} = 1.\end{aligned} \tag{9}$$

Note that the above problem is independent for different i, denoting $\mathbf{s}_i \in \mathbb{R}^{n\times 1}$ with its j-th element as $\mathbf{S}_{ij}$, and denoting $\mathbf{e}_i \in \mathbb{R}^{n\times 1}$ with the j-th element as $\mathbf{I}_{ij}$, we can solve the following problem for each i separately:

$$\min_{\mathbf{s}_i\geq 0,\mathbf{s}_i^T\mathbf{1}=1} \mathbf{s}_i^T(\mathbf{I}-\mathbf{D}^{-\frac{1}{2}}\mathbf{G}\mathbf{D}^{-\frac{1}{2}})\mathbf{s}_i + \beta||\mathbf{s}_i-\mathbf{e}_i||_2^2 + \mathbf{s}_i^T\mathbf{m}_i, \tag{10}$$

where $\mathbf{1} \in \mathbb{R}^{n\times 1}$ is a vector with all its elements as 1, and $\mathbf{m}_i \in \mathbb{R}^{n\times 1}$ is a vector with the j-th element equal to $\lambda||\mathbf{f}_i - \mathbf{f}_j||_2^2$. Denoting matrix $(\beta+1)\mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{G}\mathbf{D}^{-\frac{1}{2}}$ as $\mathbf{A}$, and denoting vector $2\beta\mathbf{e}_i - \mathbf{m}_i$ as $\mathbf{b}$, problem (10) can be easily transformed into

$$\min_{\mathbf{s}_i\geq 0,\mathbf{s}_i^T\mathbf{1}=1} \mathbf{s}_i^T\mathbf{A}\mathbf{s}_i - \mathbf{s}_i^T\mathbf{b}. \tag{11}$$

Problem (11) can be solved by optimizing the following problem

$$\min_{\mathbf{s}_i\geq 0,\mathbf{s}_i^T\mathbf{1}=1,\mathbf{z}=\mathbf{s}_i} \mathbf{s}_i^T\mathbf{A}\mathbf{z} - \mathbf{s}_i^T\mathbf{b}. \tag{12}$$

Since $\mathbf{A}$ is a positive definite matrix, according to the Augmented Largrange Method (ALM) [26], the above problem is equivalent to the following problem

$$\min_{\mathbf{s}_i^T\mathbf{1}=1,\mathbf{s}_i\geq 0,\mathbf{z}} \mathbf{s}_i^T\mathbf{A}\mathbf{z} - \mathbf{s}_i^T\mathbf{b} + \frac{\mu}{2}||\mathbf{s}_i - \mathbf{z} + \frac{1}{\mu}\alpha||_2^2, \tag{13}$$

where $\mu \in \mathbb{R}^{1\times 1}$ and $\alpha \in \mathbb{R}^{n\times 1}$ are parameters. We propose to update $\mathbf{s}_i$ and $\mathbf{z}$ iteratively.

When fixing $\mathbf{s}_i$, problem (13) becomes an unconstrained optimization problem. Taking the derivative of Eq. (13) w.r.t. $\mathbf{z}$ to zero, we have

$$\mathbf{z} = \mathbf{s}_i - \frac{1}{\mu}(\mathbf{A}^T\mathbf{s}_i + \alpha). \tag{14}$$

When fixing $\mathbf{z}$, problem (13) is simplified into the following problem

$$\min_{\mathbf{s}_i^T\mathbf{1}=1,\mathbf{s}_i\geq 0} ||\mathbf{s}_i + \frac{1}{\mu}\alpha - \mathbf{z} + \frac{\mathbf{A}\mathbf{z}-\mathbf{b}}{\mu}||_2^2, \tag{15}$$

which is a close form problem and can be readily solved by the optimization algorithm in [15]. The detailed algorithm to solve problem (13) is described in Algorithm 1.

Algorithm 1 Algorithm to solve problem (13)

1: Set $1 < \rho < 2$, initialize $\mu > 0$, α.
2: **repeat**
3: Update $\mathbf{z}$ with Eq. (14).
4: Update $\mathbf{s}_i$ by solving problem (15).
5: Update μ by $\mu = \rho\mu$.
6: Update α by $\alpha = \alpha + \mu(\mathbf{s}_i - \mathbf{z})$.
7: **until** Converge

4 Multi-view Adaptive Consistency Propagation

In real world applications, objects could be represented from multiple views. For example, in computer vision, an image may be described by different features, such as SIFT [27], HOG [28] and CENT [29]. Each feature captures a specific statistical property, and it is necessary to integrate these heterogeneous features and utilize the complementary information. In this section, we propose the Multi-view version of the ACP model (MACP).

4.1 Methodology

Supposing there are features captured from n_v views, we construct an affinity graph for each view and denote them as $\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \cdots, \mathbf{G}^{(n_v)}$. We propose to integrate these graphs to learn an optimal similarity matrix $\mathbf{S}$, so the objective function is rewritten as

$$\begin{aligned}\min_{\mathbf{S},\mathbf{F},\mathbf{w}} \frac{1}{2}\sum_{v=1}^{n_v} w_v^2 \sum_{i,j,k=1}^{n}\mathbf{G}_{jk}^{(v)}(\frac{\mathbf{S}_{ij}}{\sqrt{\mathbf{D}_{jj}^{(v)}}} - \frac{\mathbf{S}_{ik}}{\sqrt{\mathbf{D}_{kk}^{(v)}}})^2 \\ + \beta||\mathbf{S}-\mathbf{I}||_F^2 + 2\lambda\mathrm{Tr}(\mathbf{F}^T\mathbf{L}_S\mathbf{F}), \\ \text{s.t.}\mathbf{S}\geq 0, \sum\nolimits_j \mathbf{S}_{ij} = 1, \mathbf{F}\in\mathbb{R}^{n\times c}, \mathbf{F}^T\mathbf{F} = \mathbf{I}, \\ \mathbf{w}\geq 0, \sum\nolimits_v w_v = 1,\end{aligned} \tag{16}$$

where $\mathbf{D}^{(v)}$ is the degree matrix of graph $\mathbf{G}^{(v)}$. Each graph $\mathbf{G}^{(v)}$ is assigned with a weight w_v, and we aim to learn the optimal weight vector $\mathbf{w} = [w_1, w_2, \cdots, w_{n_v}]^T \in \mathbb{R}^{n_v\times 1}$ to combine these graphs. Through the learning of $\mathbf{w}$, the optimal linear combination of the graphs is exploited. Then the clustering consistency across different views can be achieved.

4.2 Optimization Strategy

In MACP, the construction of graphs $\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \cdots, \mathbf{G}^{(n_v)}$ is similar to that of ACP. In the optimization, we solve $\mathbf{S}$, $\mathbf{F}$ and $\mathbf{w}$ alternatively. The optimization of $\mathbf{F}$ is the same as the solution of problem (7). We mainly describe the optimization of $\mathbf{S}$ and $\mathbf{w}$.

Fix w, F update S

Similar to problem (9), the optimization of $\mathbf{S}$ is independent for different i, so the problem can be rewritten as

$$\begin{aligned}\min_{\mathbf{s}_i\geq 0,\mathbf{s}_i^T\mathbf{1}=1} & \mathbf{s}_i^T[\sum_{v=1}^{n_v} w_v^2(\mathbf{I}-\mathbf{D}^{(v)^{-\frac{1}{2}}}\mathbf{G}^{(v)}\mathbf{D}^{(v)^{-\frac{1}{2}}})]\mathbf{s}_i \\ & +\beta||\mathbf{s}_i-\mathbf{e}_i||_2^2+\mathbf{s}_i^T\mathbf{m}_i,\end{aligned} \tag{17}$$

where the definition of $\mathbf{m}_i$ is the same as that in problem (10). After removing the irrelevant terms, problem (17) is with the similar form to problem (11), and can be readily solved by Algorithm 1.

Fix S, F update w

When updating $\mathbf{w}$, problem (16) is simplified into

$$\begin{aligned}&\min_{\mathbf{w}}\sum_{v=1}^{n_v} w_v^2\sum_{i,j,k=1}^{n}\mathbf{G}_{jk}^{(v)}(\frac{\mathbf{S}_{ij}}{\sqrt{\mathbf{D}_{jj}^{(v)}}}-\frac{\mathbf{S}_{ik}}{\sqrt{\mathbf{D}_{kk}^{(v)}}})^2 \\ &\text{s.t. } \mathbf{w}\geq 0,\sum\nolimits_v w_v=1.\end{aligned} \tag{18}$$

Denoting $\sum_{i,j,k=1}^{n}\mathbf{G}_{jk}^{(v)}(\frac{\mathbf{S}_{ij}}{\sqrt{\mathbf{D}_{jj}^{(v)}}}-\frac{\mathbf{S}_{ik}}{\sqrt{\mathbf{D}_{kk}^{(v)}}})^2$ as p_v, the above problem is transformed into

$$\begin{aligned}&\min_{\mathbf{w}}\sum_{v=1}^{n_v} w_v^2 p_v \\ &\text{s.t. } \mathbf{w}\geq 0,\sum\nolimits_v w_v=1.\end{aligned} \tag{19}$$

Denoting a diagonal matrix $\mathbf{P}\in\mathbb{R}^{n_v\times n_v}$, problem (19) is equivalent to

$$\begin{aligned}&\min_{\mathbf{w}}\mathbf{w}^T\mathbf{P}\mathbf{w} \\ &\text{s.t. } \mathbf{w}\geq 0,\mathbf{w}^T\mathbf{1}=1.\end{aligned} \tag{20}$$

Removing the constraint $\mathbf{w}\geq 0$, the Lagrangian function of problem (20) is

$$\mathcal{L}(\mathbf{w},\eta)=\mathbf{w}^T\mathbf{P}\mathbf{w}-\eta(\mathbf{w}^T\mathbf{1}-1), \tag{21}$$

where η is the Lagrangian multiplier. According to the Karush-Kuhn-Tucker (KKT) condition, we have $\frac{\partial\mathcal{L}(\mathbf{w},\eta)}{\partial\mathbf{w}}=0$. Together with the constraint $\mathbf{w}^T\mathbf{1}=1$, the optimal $\mathbf{w}$ can be obtained as

$$w_v=\frac{1}{p_v}\times(\sum_{r=1}^{n_v}\frac{1}{p_r})^{-1}, \tag{22}$$

which satisfies the removed constraint $\mathbf{w}\geq 0$ definitely.

By updating $\mathbf{F},\mathbf{S}$ and $\mathbf{w}$ iteratively, the graphs from different views are integrated to obtain the consistent clustering result.

5 Experiments

In this section, we verify the clustering performance of the proposed ACP and MACP respectively. The evaluation is based on two widely used clustering metric: clustering accuracy (ACC) [8] and Normalized Mutual Information (NMI) [23]. For a fair comparison, we let all the competitors use their best parameters.

5.1 Experimental Results of ACP on Single-View Clustering

In this part, the performance of the proposed Adaptive Consistency Propagation (ACP) method is evaluated on real-world datasets.

Datasets: experiments are conducted on nine real-world benchmarks: one object dataset, i.e., Coil20 [30], two face datasets, i.e., Jaffe [31] and orlraws10P [32], three datasets form UCI Machine Learning Repository [33], i.e., Yeast, Mfeat-pix and Movement, and two biology datasets, i.e., Lung [34] and Carcinom [35].

Competitors: for a quantitative evaluation, the proposed ACP is compared with 7 competitors, including k-means [1], Non-negative matrix Matrix Factorization (NMF) [9], Normalized-cut (Ncut) [7], Simplex Sparse Representation (SSR) [15], Constrained Laplacian Rank L1-norm (CLR_L1) [18], Spectral Clustering with Single Kernel (SCSK1) [19] and Similarity and Clustering with a Single Kernel (SCSK2) [20]. For CLR_L1 and the proposed ACP, the neighborhood size is set as 5 when constructing the input data graph. And for Ncut, the data graph is built with the self-tune Gaussian method [16]. For SCSK1 and SCSK2, the kernel is constructed as $K(\mathbf{x},\mathbf{y})=\exp(-||\mathbf{x}-\mathbf{y}||_2^2/d_{max}^2)$, where d_{max} is the largest distance between the data points. In addition, since Ncut, SCSK1 and SCSK2 involve k-means as the post-processing, we repeat them for 30 times and report the average performance. In our method, the parameter β is set as 1 empirically, and the value of λ is chosen in a heuristic way according to the number of zero eigenvalues in $\mathbf{L}_S$ [18].

Performance: the clustering results of different methods are shown in Table 1, from which we can observe that ACP achieves the highest ACC and NMI in most cases. Particularly, on Yeast, Mfeat-pix and Carcinom, ACP outperforms the second best method a lot. NCut relies highly on the input affinity graph, so its performance may be adversely affected by the graph quality. NMF does not require the data graph as input, but it just emphasizes the global data structure and fails to capture the local data structure. SSR, SCSK1 and SCSK2 are robust to the graph quality and show good performance because they optimize the data graph during the clustering procedure. But the obtained graphs cannot be directly used as the indicator matrices, so they need spectral clustering as the post-processing (different results for every time of running). All the competitors neglect the connectivity between the far away points, which makes them fail to perceive the data manifold. The proposed ACP sufficiently explores the manifold structure without any post-processing, so it outperforms the competitors.

5.2 Experimental Results of MACP on Multi-View Clustering

We also evaluate the Multi-view ACP (MACP) on multi-view clustering, and compare its performance with the state-of-the-art multi-view clustering methods.

Datasets: four benchmark multi-view clustering datasets are used in the experiments, including MSRC-v1 [36], Handwritten [37], Caltech101-7 and Caltech101-20 [38]. For MSRC-v1 dataset, following Nie et al. [39], 210 images are chosen for clustering, which comes from 7 classes. We

TABLE 1
ACC/NMI of single-view clustering methods. The best results are in bold face.

	k-means	NMF	NCut	SSR	CLR	SCSK1	SCSK2	ACP
Coil20	0.55/0.71	0.43/0.55	0.48/0.64	0.69/0.82	0.82/0.90	0.68/0.80	0.66/0.79	**0.84**/**0.93**
Jaffe	0.74/0.80	0.66/0.67	0.74/0.79	0.61/0.69	0.96/0.95	0.95/0.94	0.81/0.85	**0.97**/**0.96**
orlraw10P	0.67/0.94	0.63/0.68	0.72/0.79	0.61/0.68	0.78/0.84	0.74/0.78	0.74/0.82	**0.81**/**0.87**
Yeast	0.36/0.23	0.30/0.14	0.31/0.22	0.23/0.11	0.34/0.13	0.37/0.24	0.37/0.25	**0.44**/**0.30**
Mfeatpix	0.71/0.71	0.42/0.35	0.71/0.70	0.66/0.74	0.87/0.88	0.66/0.72	0.77/0.79	**0.94**/**0.89**
Movement	0.44/0.57	0.41/0.47	0.43/0.57	0.21/0.34	0.43/0.59	**0.53**/0.60	0.46/0.58	0.51/**0.63**
Lung	0.67/0.50	0.57/0.37	0.55/0.42	0.72/0.53	0.79/0.49	0.67/0.44	0.84/**0.66**	**0.88**/0.64
Carcinom	0.64/0.66	0.73/0.74	0.67/0.71	0.45/0.56	0.65/0.67	0.78/0.76	0.77/0.79	**0.84**/**0.81**

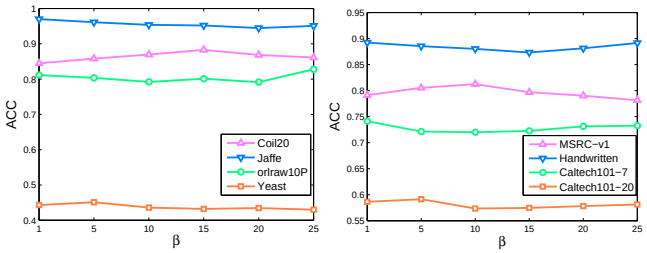
TABLE 2
ACC/NMI of multi-view clustering methods. The best results are in bold face.

	Co-reg	RMSC	MMSC	AMGL	IVA	SCMK1	SCMK2	MACP
MSRC	0.70/0.60	0.67/0.59	0.71/0.63	0.72/0.65	0.73/0.66	0.76/0.69	0.73/0.65	**0.79**/**0.71**
Handwritten	0.79/0.82	0.77/0.75	0.84/0.86	0.79/0.83	0.85/0.87	0.87/0.88	0.83/0.86	**0.89**/**0.91**
Caltech101-7	0.43/0.37	0.59/0.52	0.70/0.58	0.60/0.54	0.67/0.57	0.51/0.40	0.55/0.49	**0.74**/**0.62**
Caltech101-20	0.48/0.55	0.51/0.59	0.52/0.60	0.49/0.57	0.50/0.59	0.54/0.62	0.56/0.63	**0.59**/**0.66**

extract 5 features, including Color Moment (CM), HOG, GIST, LBP and CENT, for multi-view clustering. Handwritten dataset contains 2000 digit images from 10 classes. Each image is represented by 6 features: FOU, FAC, KAR, PIX, ZER and MOR. Caltech101-7 and Caltech101-20 contains 1474 and 2386 images respectively, and the features are Gabor, Wavelet Moments (WM), CENT, HOG, GIST, LBP.

Competitors: the proposed MACP is compared with 7 multi-view clustering methods, including Co-regularized spectral clustering (Co-reg) [40], Robust Multiview Spectral Clustering (RMSC) [41] and Multi-Modal Spectral Clustering (MMSC) [42], Autoweighted Multiple Graph Learning (AMGL) [39], Iterative Views Agreement (IVA) [43], Spectral Clustering with Multiple Kernels (SCMK1) [19] and Similarity learning and Clustering with Multiple Kernel (SCMK2) [20]. Note that, SCMK1 and SCMK2 are the multi-kernel version of SCSK1 and SCSK2 respectively, in experiments we use the data graphs as the kernels. To reduce the influence of initialization, some competitors are repeated for 30 times and the average results are reported. The parameter β in MACP is set as 1.

Performance: Table 2 show the quantitative results of different methods, it can be seen that the proposed MACP shows the best performance on all the datasets. Co-reg, MMSC, IVA simply assign the equal weight for each view, so they fail to find the optimal combination of the multi-view graphs. RMSC assumes that each view is sufficient to maintain most of the discriminative information, and then learns the optimal graph by removing the noise within each view. Therefore, it tends to be affected by the weak views. AMGL, SCMK1 and SCMK2 learn the weight of each view and find the desired linear combination, but they cannot produce stable results because spectral clustering is used for post-processing. The proposed MACP captures the consistency propagation between points, and integrates the graphs with the optimal combination to obtain the clustering results directly.



(a) ACP (b) MACP

Fig. 1. ACC curves of ACP and MACP with different values of β.

6 Discussion

In this section, we investigate the effect of parameter in ACP and MACP. As mentioned above, the parameter λ is determined automatically in a heuristic way, so we only discuss the impact of parameter β, which controls the balance of the smooth term and fitting term in Eq. (6) and Eq. (16). As we vary the value of β from $\{1, 5, 10, 15, 20, 25\}$, the variances of clustering accuracies are shown in Figure 1. As shown in the figure, the performance of ACP and MACP are not very sensitive to value of β. So we simply set β to 1 in the experiments.

7 Conclusion

In this paper, the Adaptive Consistency Propagation (ACP) and its multi-view version MACP are proposed for clustering. Most of the traditional methods only focus on the data points with neighboring relationship, and keep the data graph fixed during the optimization procedure. In our new methods, the local consistency is propagated adaptively from near to far, so the points from the same cluster can be all pulled together. In addition, with a reasonable constraint, ACP and MACP are able to learn the optimal graph for clustering, and accomplish clustering simultaneously without any post-processing. Comprehensive experiments on single-

and multi-view clustering show the superior performance of our methods on various kinds of datasets.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant 61871470, U1864204 and 61773316, and Project of Special Zone for National Defense Science and Technology Innovation.

REFERENCES

[1] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[2] F. J. Rohlf, "Adaptive hierarchical clustering schemes," *Systematic Zoology*, vol. 19, no. 1, pp. 58–82, 1970.

[3] Q. Wang, Z. Qin, F. Nie, and X. Li, "Spectral embedded adaptive neighbors clustering," *TNNLS*, vol. 30, no. 4, pp. 1265–1271, 2019.

[4] F. Nie, Z. Zeng, I. Tsang, D. Xu, and C. Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," *IEEE Transactions on Neural Networks*, vol. 22, no. 11, pp. 1796–1808, 2011.

[5] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, "Maximum margin clustering," in *Advances in Neural Information Processing Systems*, 2004, pp. 1537–1544.

[6] A. Ben-Hur, D. Horn, H. Siegelmann, and V. Vapnik, "Support vector clustering," *Journal of Machine Learning Research*, vol. 2, pp. 125–137, 2001.

[7] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis on Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[8] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4147–4153.

[9] T. Li and C. Ding, "The relationships among various nonnegative matrix factorization methods for clustering," in *IEEE International Conference on Data Mining*, 2006, pp. 362–371.

[10] P. Chan, M. Schlag, and J. Zien, "Spectral k-way ratio-cut partitioning and clustering," *IEEE Transactions on CAD of Integrated Circuits and Systems*, vol. 13, no. 9, pp. 1088–1096, 1994.

[11] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis on Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[12] W. Pentney and M. Meila, "Spectral clustering of biological sequence data," in *National Conference on Artificial Intelligence*, 2005, pp. 845–850.

[13] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI: 10.1109/TPAMI.2018.2875002, 2018.

[14] D. Cai, X. He, J. Han, and T. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[15] J. Huang, F. Nie, and H. Huang, "A new simplex sparse learning model to measure data similarity for clustering," in *International Joint Conference on Artificial Intelligence*, 2015, pp. 3569–3575.

[16] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems*, 2004, pp. 1601–1608.

[17] C. Peng, Z. Kang, Y. Hu, J. Cheng, and Q. Cheng, "Nonnegative matrix factorization with integrated graph and feature learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 3, pp. 42:1–42:29, 2017.

[18] F. Nie, X. Wang, M. Jordan, and H. Huang, "The constrained laplacian rank algorithm for graph-based clustering," in *AAAI Conference on Artificial Intelligence*, 2016, pp. 1969–1976.

[19] Z. Kang, C. Peng, Q. Cheng, and Z. Xu, "Unified spectral clustering with optimal graph," in *AAAI Conference on Artificial Intelligence*, 2018.

[20] Z. Kang, C. Peng, and Q. Cheng, "Twin learning for similarity and clustering: A unified kernel approach," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 2080–2086.

[21] ——, "Clustering with adaptive manifold structure learning," in *IEEE International Conference on Data Engineering*, 2017, pp. 79–82.

[22] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 977–986.

[23] Q. Wang, M. Chen, and X. Li, "Quantifying and detecting collective motion by manifold learning," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4292–4298.

[24] B. Mohar, Y. Alavi, G. Chartrand, O. R. Oellermann, and A. J. Schwenk, "The laplacian spectrum of graphs," in *Graph Theory, Combinatorics, and Applications*, 2001, pp. 871–898.

[25] K. Fan, "On a theorem of weyl concerning eigenvalues of linear transformations i." *National Academy of Sciences of the United States of America*, vol. 35, no. 11, pp. 652–655, 1949.

[26] F. Nie, H. Wang, H. Huang, and C. Ding, "Joint schatten p -norm and ℓ_p -norm robust matrix completion for missing value recovery," *Knowledge and Information Systems*, vol. 42, no. 3, pp. 525–544, 2015.

[27] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[29] J. Wu and J. Rehg, "Where am I: place instance and category recognition using spatial PACT," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[30] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, p. 1548, 2011.

[31] M. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357–1362, 1999.

[32] J. Li, K. Cheng, S. Wang, F. Morstatter, T. Robert, J. Tang, and H. Liu, "Feature selection: A data perspective," *arXiv:1601.07996*, 2016.

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

[34] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, and J. P. Richie, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, p. 203, 2002.

[35] J. Li, K. Cheng, S. Wang, F. Morstatter, T. Robert, J. Tang, and H. Liu, "Feature selection: A data perspective," *arXiv:1601.07996*, 2016.

[36] J. Winn and N. Jojic, "LOCUS: learning object classes with unsupervised segmentation," in *IEEE International Conference on Computer Vision*, 2005, pp. 756–763.

[37] M. Breukelen, R. Duin, D. Tax, and J. Hartog, "Handwritten digit recognition by combined classifiers," *Kybernetika*, vol. 34, no. 4, pp. 381–386, 1998.

[38] F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[39] F. Nie, J. Li, and X. Li, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 1881–1887.

[40] A. Kumar, P. Rai, and H. Daum, "Co-regularized multi-view spectral clustering," in *Advances in Neural Information Processing Systems*, 2011, pp. 1413–1421.

[41] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *AAAI Conference on Artificial Intelligence*, 2014, pp. 2149–2155.

[42] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1977–1984.

[43] Y. Wang, W. Zhang, L. Wu, X. Lin, M. Fang, and S. Pan, "Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 2153–2159.

Xuelong Li (M'02-SM'07-F'12) is currently a Full Professor with the School of Computer Science and the Center for OPTical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China.

Mulin Chen received the B.E. degree in software engineering and the M.E. degree in computer application technology from Northwestern Polytechnical University, Xi'an, China, in 2014 and 2016 respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His current research intersts include computer vision and machine learning.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.