Semi-supervised Clustering via Pairwise Constrained Optimal Graph

Feiping Nie[1,*] , **Han Zhang**[1] , **Rong Wang**[2,1] and **Xuelong Li**[1]

[1]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an, 710072, Shaanxi, P. R. China
[2]School of Cybersecurity, Northwestern Polytechnical University, Xi'an, 710072, P. R. China
{feipingnie,zhanghan9937}@gmail.com, wangrong07@tsinghua.org.cn, li@nwpu.edu.cn

Abstract

In this paper, we present a technique of definitely addressing the pairwise constraints in the semi-supervised clustering. Our method contributes to formulating the cannot-link relations and propagating them over the affinity graph flexibly. The pairwise constrained instances are provably guaranteed to be in the same or different connected components of the graph. Combined with the Laplacian rank constraint, the proposed model learns a Pairwise Constrained structured Optimal Graph (PCOG), from which the specified c clusters supporting the known pairwise constraints are directly obtained. An efficient algorithm invoked by the label propagation is designed to solve the formulation. Additionally, we also provide a compact criterion to acquire the key pairwise constraints for prompting the semi-supervised graph clustering. Substantial experimental results show that the proposed method achieves the significant improvements by using a few prior pairwise constraints.

1 Introduction

Traditional clustering targets at grouping instances according to their inherent similarities without any supervision. However, the attributes of each cluster obtained from unsupervised clustering are mostly unpredictable (e.g., Figure 1). In light of this, semi-supervised clustering attracts lots of interests, which not only guides the clustering results according to preference, but also improves the clustering performance significantly. Semi-supervised clustering frequently uses the prior information of pairwise constraints, including must-link constraints (ML, specify that the pair of instances must be in the same cluster), and the cannot-link constraints (CL, specify that the pair of instances must be in different clusters), since the pairwise constraints are much easier to acquire and more flexible than labels in practice when the cluster number is unavailable. For instance, to distinguish portraits, the images of one person are must linked, whereas the images of different persons are cannot linked. A lot of researches on pairwise constrained clustering have been made and implemented in many applications, such as medical diagnosis [Thangavel and Mohideen, 2010], image segmentation [Saha *et al.*, 2016], and information networks [Li *et al.*, 2017], etc.

Graph based clustering is always an important branch in the clustering and attracts more and more attentions in recent years, due to the high tractability of graph representations of data, such as the well-known normalized cut (N-cut) [Ng *et al.*, 2002; Shi and Malik, 2000], and ratio cut [Hagen and Kahng, 1992]. The performance of graph clustering depends on a given graph where the pairwise relations of samples are depicted by an affinity matrix, thus it is natural to incorporate pairwise constraints into graphs. Many constrained graph clustering [Lu and Carreira-Perpinan, 2008; Kulis *et al.*, 2009; Śmieja *et al.*, 2018] were raised accordingly that first refined the affinity matrix by given pairwise constraints, then performed spectral clustering on the modified graph. However, due to the cannot-link's non-transitive property, they only dealt with the must-link constraints effectively. Instead of modifying graphs, Wang et al. [Wang *et al.*, 2014] solved a constrained spectral clustering under the constraint that the number of satisfied pairwise constraints was greater than a threshold value. Later, Cucuringu [Cucuringu *et al.*, 2016] put forward that traditional spectral clustering was a special case of constrained clustering with implicit pairwise constraints, and they modified the denominator of Ncut's objective function to be the graph Laplacians associated with the defined cannot-link matrix, such that the pairwise constraints were considered into partitioning.

Nevertheless, existing approaches still encounter two major issues: **i)** the cannot-link problem, how to provably ensure the instances under the cannot-link constraint to be in different clusters; **ii)** the multi-class problem, how to flexibly incorporate pairwise constraints into the affinity matrix for direct multi-class clustering. In other words, **how to perform the multi-clustering task under the provably valid supervision of cannot-link constraints remains a crucial challenge.**

In this paper, we simultaneously address the cannot-link problem as well as the multi-class problem in the constrained graph clustering. To be specific, we contribute to

- presenting a novel cannot-link graph regularization that provably guarantees each cannot-link constrained instance to be in different clusters, as demonstrated in Theorem 1 and illustrated in Figure 2(a);

[*]Contact Author

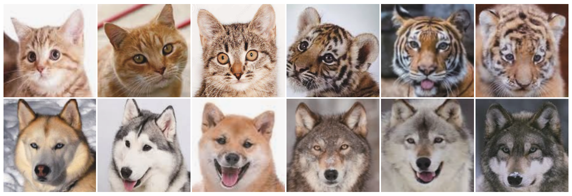


Figure 1: Grouping images into two clusters: from the view of aggressivity, the three left columns form a cluster, and the rest columns form the other cluster; from biology, each row is more like a cluster.

- optimizing the graph to support the given must-link as well as cannot-link constraints and has the specified c connected components for direct multi-class clustering, as illustrated in Figure 2(b);
- providing a simple yet efficient pairwise constraints selection criterion that specifically for improving semi-supervised graph clustering.

We further conduct substantial experiments and verify that the proposed method reaches the excellent performance by using a few key pairwise constraints.

Notations. Throughout the paper, $Tr(M)$ denotes the trace of a matrix M. $\|v\|_2$ is the ℓ_2-norm of a vector v. $\|M\|_F$ denotes the Frobenius norm of M. M^T is the transpose of M, and $rank(M)$ is the rank of M.

2 Preliminary

The prevalent graph based spectral clustering is a two-step process that first seeks the intrinsic low-dimensional embedding from the pre-constructed affinity graph, and then performs k-means on the embedding to obtain the cluster labels, since the graphs built from the original feature subspace lack of the explicit cluster structure. To deal with it, in this section, we revisit a constrained Laplacian rank algorithm proposed by [Nie *et al.*, 2016] for direct multi-class graph clustering, formulated as:

$$\begin{aligned} &\min_{S} \|S - A\|_F^2, \\ &s.t.\ \ S \succeq 0, S\mathbf{1} = \mathbf{1}, rank(L_S) = n - c, \end{aligned} \tag{1}$$

where $\mathbf{1} = [1, \cdots, 1]^T \in \mathbb{R}^n$. $A \in \mathbb{R}^{n\times n}$ is the pre-constructed affinity matrix of n data points, and $S \in \mathbb{R}^{n\times n}$ is the optimized affinity matrix. $L_S = D_S - S$ is the Laplacian matrix of S, and D_S is the degree matrix of S with $d_{ii} = \sum_j s_{ij}$. With inputting a rough similarity matrix A, this model obtains a non-negative and normalized approximation S which possesses the exact c connected components. As a result, the quality of original graph is improved by removing excrescent connections and the instances are exactly partitioned into c clusters. This formulation provides how to achieve a c-connected graph. However, whether the incorrect connections are removed and the valid connections are preserved is uncontrolled. In the next section, we elaborate how to overcome this deficiency by tactically incorporating the pairwise constraints and obtaining the desired clusters.

3 Methodology

3.1 Problem Formulation

Suppose an affinity matrix $A = [a_{ij}]_{n\times n}$ associated with the dataset $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$, where x_i is a d-dimensional data point, a_{ij} represents the similarity between i-th sample and j-th sample. We utilize a few pairwise constraints to guide the partition of $\mathcal{X}$, denoted as:

$$\begin{aligned} \mathcal{M} &= \{(x_i, x_j)|\forall i, \forall j > i, x_i \text{ must link } x_j\}; \\ \mathcal{C} &= \{(x_i, x_j)|\forall i, \forall j > i, x_i \text{ cannot link } x_j\}. \end{aligned} \tag{2}$$

From the viewpoint of the graph connectivity, when the must-link constrained instances locate in one connected component while the cannot-link constrained instances locate in different connect components, the constraints can be satisfied intuitively. Moreover, if the edge of must-linked instances exists, they naturally belong to one connected component. However, removing the connections between cannot-linked instances does not work since they still probably connect to each other through intermediate nodes. To address it, we first put forward a theorem to definitely isolate the cannot-link constrained instances from each other, as below.

Theorem 1. *Suppose a nonnegative graph S, and there exists a cannot-link constraint between x_a and x_b. $y \in \mathbb{R}^n$ is the cannot-link indicator where $y_a = 1$ and $y_b = -1$. When*

$$y^T L_S y = 0, \tag{3}$$

x_a and x_b must be in different connected components of S.

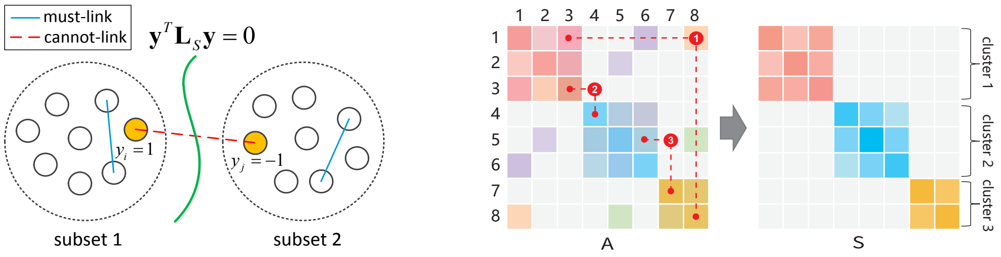
Proof. ***By reduction to absurdity***. If x_a and x_b are in the same connected component of S, there is at least a path $\mathcal{P}^* = \{x_a, x_{k_1}, \cdots, x_{k_t}, x_b\}$ from x_a to x_b. Denote $\mathcal{J} = y^T L_S y = \sum_i \sum_j (y_i - y_j)^2 s_{ij}$. We split $\mathcal{J} = \mathcal{J}(\mathcal{P}^*) + \mathcal{J}(\widetilde{\mathcal{P}}^*)$, where $\mathcal{J}(\mathcal{P}^*) \geq 0$ is the objectives associated with $\mathcal{P}^*$ and $\mathcal{J}(\widetilde{\mathcal{P}}^*) \geq 0$ is the objectives of all paths apart from $\mathcal{P}^*$. When $\mathcal{J} = 0$ is required, we have $\mathcal{J}(\mathcal{P}^*) = 0$ as well. Since $\mathcal{J}(\mathcal{P}^*) = (y_a - y_{k_1})^2 s_{ak_1} + \cdots + (y_{k_t} - y_b)^2 s_{k_t b}$, and $s_{a_{k_1}} > 0, \cdots, s_{k_t b} > 0$, it is inferred that $y_a = y_{k_1} = \cdots = y_{k_t} = y_b$, which is contradictory to the conditions of $y_a = 1, y_b = -1$. As a result, x_a and x_b must be in different connected components of S. □

According to Theorem 1, we present the cannot-link graph regularization to learn the pairwise constrained structured optimal graph $S \in \mathbb{R}^{n\times n}$ from the given affinity A under the supervision of the pairwise constraints, formulated as:

$$\min_{S\in\Omega, Y\in\Psi} \|S - A\|_F^2 + \gamma \sum_{k=1}^{p} y_k^T L_S y_k, \tag{4}$$

where $\gamma > 0$ is a regularization parameter, and p is the number of the constraints in $\mathcal{C}$. $\Omega = \Omega^\dagger \cap \Omega^\ddagger$ is the feasible zone of the learned graph S, where $\Omega^\dagger$ inherits from model (1) and $\Omega^\ddagger$ is designed for addressing must-link constraints. $Y \in \Psi$ is the zone feasible of all y_k equipped for the cannot-link regularization, i.e.,

$$\begin{aligned} &\Omega^\dagger : \{S|S \succeq 0, S\mathbf{1} = \mathbf{1}, rank(L_S) = n - c\}, \\ &\Omega^\ddagger : \{S|\forall (x_i, x_j) \in \mathcal{M} : s_{ij} = \tau, s_{ji} = \tau\}, \\ &\Psi : \{Y|\forall (x_{i_k}, x_{j_k}) \in \mathcal{C} : y_k(i_k) = 1, y_k(j_k) = -1\}, \end{aligned} \tag{5}$$



(a) (b)

Figure 2: The illustration to the proposed method. (a) The cannot-link graph regularization. Suppose a cannot-link constraint between x_i and x_j. According to Theorem 1, when $y^T L_S y = 0$, the paths from x_i to x_j in the graph S are removed thoroughly; (b) The multi-class clustering. Given three cannot-link constraints between $\{x_1, x_8\}$, $\{x_3, x_4\}$ and $\{x_5, x_7\}$. By virtue of the Laplacian rank constraint and the cannot-link graph regularization, the graph S learned from A would has three connected components satisfying cannot-link constraints.

Invoked by model (1), $S \in \Omega^\dagger$ compels that S has exact c connected components. $S \in \Omega^\ddagger$ sets the connections between the must-linked instances by a small value τ (where $\tau = 0.1$ simply). $Y \in \Psi$ fixes $y_k(i_k) = 1$ and $y_k(j_k) = -1$ for each y_k to "label" the k-th couple of cannot-link constrained x_{i_k} and x_{j_k}. Benefited from label propagation [Zhu and Ghahramani, 2002], the rest entries in y_k are learned and thus the cannot-link relations are propagated over the affinity graph, as shown in Figure 2(a). According to Theorem 1, when γ is large enough, minimizing the regularization term in model 4 makes the cannot-link constraints satisfied simultaneously.

By optimizing model (4), we can obtain the Pairwise Constrained Optimal Graph (PCOG) that directly indicates c clusters, where the must-link constrained instances must be in the same cluster and the cannot-link constrained instances must be in different clusters (see Figure 2(b)). In this way, two aforementioned problems of semi-supervised clustering are addressed simultaneously. To our best knowledge, this is the first work to provably address the cannot-linked instances. Next, how to solve model (4) takes the first priority.

3.2 Optimization of Model (4)

Following the optimization of model (1), the non-convex rank constraint in model (4) is tackled by solving its counterpart. Denote the i-th smallest eigenvalue of L_S as $\sigma_i(L_S)$. Since L_S is positive semi-definite, we have $\sigma_i(L_S) \geq 0$ for each i. When the first c smallest $\sigma_i(L_S)$ equal zero, the constraint $rank(L_S) = n - c$ is actually achieved. Based on this, model (4) is transformed into:

$$\min_{S,Y} \|S - A\|_F^2 + \gamma \sum_{k=1}^{p} y_k^T L_S y_k + \lambda \sum_{i=1}^{c} \sigma_i(L_S), \tag{6}$$
$$s.t.\ \ S \succeq 0, S\mathbf{1} = \mathbf{1}, S \in \Omega^\ddagger, Y \in \Psi,$$

where λ is large enough, and $\sum_{i=1}^{c} \sigma_i(L_S)$ is minimized to be zero. According to Ky Fan's Theorem [Fan, 1950], problem (6) could be further equivalent to:

$$\min_{S,F,Y} \|S - A\|_F^2 + \gamma Tr(Y^T L_S Y) + \lambda Tr(F^T L_S F),$$
$$s.t.\ \ S \succeq 0, S\mathbf{1}^T = \mathbf{1}, S \in \Omega^\ddagger, F^T F = I, Y \in \Psi, \tag{7}$$

where $F = \{f^1, f^2, \cdots, f^n\} \in \mathbb{R}^{n\times c}$ is a manifold embedding of data. Subsequently, we optimize three variables $\{S, F, Y\}$ in an iterative manner, known as block-coordinate descent method [Tseng, 2001].

(i). When updating S with the fixed F and Y, problem (7) is transformed into:

$$\min_{S} \|S - A\|_F^2 + \gamma Tr(Y^T L_S Y) + \lambda Tr(F^T L_S F),$$
$$s.t.\ \ S \succeq 0, S\mathbf{1} = \mathbf{1}, S \in \Omega^\ddagger. \tag{8}$$

To address problem above, we introduce an important equation in the spectral analysis [Ng *et al.*, 2002], described as:

$$Tr(Y^T L_S Y) = \frac{1}{2}\sum_{i,j} \left\|y^i - y^j\right\|_2^2 s_{ij}. \tag{9}$$

Then, we could reformulate problem (8) as

$$\min_{S} \sum_{i,j} (s_{ij} - a_{ij})^2 + \sum_{i,j} \left(\frac{\gamma}{2} d_{ij}^y + \frac{\lambda}{2} d_{ij}^f\right) s_{ij}, \tag{10}$$
$$s.t.\ \ S \succeq 0, S\mathbf{1} = \mathbf{1}, S \in \Omega^\ddagger,$$

where $d_{ij}^y = \left\|y^i - y^j\right\|_2^2$ and $d_{ij}^f = \left\|f^i - f^j\right\|_2^2$. Note that the optimization of each column in S is independent, so we decompose problem (10) into solving

$$\min_{s_i} \left\|s_i - \left(a_i - \frac{1}{2} d_i\right)\right\|_2^2, \tag{11}$$
$$s.t.\ \ s_i \geq 0, s_i^T \mathbf{1} = 1, \forall j, (x_i, x_j) \in \mathcal{M} : s_{ij} = \tau,$$

where $d_i = \gamma d_i^y + \lambda d_i^f$. According to $\mathcal{M}$, we denote the must-linked objects of the i-th samples as $\kappa_i = \{j | (x_i, x_j) \in \mathcal{M}\}$. Fix $s_{i,j:j\in\kappa_i} = \tau$, $s_{j:j\in\kappa_i,i} = \tau$, and let $\tilde{s}_i = s_{i,j:j\notin\kappa_i}$, $\tilde{a}_i = a_{i,j:j\notin\kappa_i}$, $\tilde{d}_i = d_{i,j:j\notin\kappa_i}$, then problem (11) could be transformed into

$$\min_{\tilde{s}_i} \left\|\tilde{s}_i - \left(\tilde{a}_i - \frac{1}{2}\tilde{d}_i\right)\right\|_2^2, s.t.\ \tilde{s}_i \geq 0, \tilde{s}_i^T \mathbf{1} = 1 - |\kappa_i|\tau, \tag{12}$$

where $|\kappa_i|$ denotes the number of elements in κ_i. Problem (12) could be efficiently addressed by referring to problem (9) in Reference [Nie *et al.*, 2016].

Algorithm 1 Algorithm to solve problem (4)

```
Require: the pairwise constraints $\mathcal{M}$ and $\mathcal{C}$, an affinity matrix $A$, a constant value $\tau$, the parameters $\lambda$ and $\gamma$.
Ensure: the graph $S$ with $c$ connected components that supports the given constraints.
  Initialize $F \in \mathbb{R}^{n\times c}$ with the eigenvectors of $L_A = D_A - \frac{A^T+A}{2}$ corresponding to its $c$ smallest eigenvalues; $Y \in \mathbb{R}^{n\times p}$ with a random matrix ranging in $[-1,1]$ and let $y_k(i_k) = 1$, $y_k(j_k) = -1$ for $k$-th pair of cannot-linked samples $(x_{i_k}, x_{j_k}) \in \mathcal{C}$; $\kappa_i$ records the must-linked samples of $i$-th sample according to $\mathcal{M}$; Let $\tilde{a}_i = a_{i,j:j\notin\kappa_i}$.
  For each $i$, set $s_{i,j:j\in\kappa_i} = \tau$, $s_{j:j\in\kappa_i,i} = \tau$.
  while not converge do
    For each $i,j$, compute $d_{ij} = \gamma d_{ij}^y + \lambda d_{ij}^f$ where $d_{ij}^y = \left\|y^i - y^j\right\|_2^2$, $d_{ij}^f = \left\|f^i - f^j\right\|_2^2$;
    For each $i$, update $\tilde{s}_i$ by solving problem (12);
    For each $k$, update $y_u^{(k)}$ by solving problem (14);
    Update $F$ by solving problem (15);
  end while
```

(ii). When updating Y with the fixed F and S, the columns of Y in problem (7) are also independent to each other and could be achieved by solving

$$\min_{y_k} y_k^T L_S y_k, \quad s.t.\ \forall(x_{i_k}, x_{j_k}) \in \mathcal{C} : y_k(i_k) = 1, y_k(j_k) = -1. \tag{13}$$

As stated before, the optimization of y_k could be regarded as the label propagation process over the graph S [Zhu and Ghahramani, 2002]. To be specific, we rearrange all samples as $\mathcal{X}^{(k)} = \{x_{i_k}, x_{j_k}, x_1, \cdots, x_n\}$. Accordingly, the rearranged y_k is expressed as $y^{(k)} = [y_l^{(k)}, y_u^{(k)}]^T \in \mathbb{R}^n$, where $y_l^{(k)} = [1, -1]^T$. The similarity S is rearranged into $S^{(k)}$, whose Laplacian matrix is $L_S^{(k)}$. So, problem (13) is reformulated as

$$\min_{y^{(k)}} y^{(k)T} L_S^{(k)} y^{(k)}, \quad s.t.\ y_l^{(k)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \text{ where } L_S^{(k)} = \begin{bmatrix} L_{ll}^{(k)} & L_{lu}^{(k)} \\ L_{ul}^{(k)} & L_{uu}^{(k)} \end{bmatrix}. \tag{14}$$

According to the label propagation algorithm [Zhu and Ghahramani, 2002], the closed-form solution to problem (14) is $y_u^{(k)} = -{L_{uu}^{(k)}}^{-1} L_{ul}^{(k)} y_l^{(k)}$.

(iii). When updating F with the fixed S and Y, problem (7) is equivalent to

$$\min_F Tr(F^T L_S F), \quad s.t.\ F^T F = I. \tag{15}$$

Problem (15) degrades to a spectral clustering problem [Ng *et al.*, 2002], whose solution consists of the eigenvectors of L_S corresponding to its c smallest eigenvalues.

3.3 Key Pairwise Constraints Selection

In semi-supervised clustering approaches, a compromised way to acquire the pairwise constraints is to query pairs of samples as much as possible. Since excessive constraints is labor-intensive, many constraint selection approaches [Craenendonck *et al.*, 2017; Xiong *et al.*, 2014] have been proposed. Different from them, we put forward a compact scheme specifically for semi-supervised graph clustering. Considering that A is sparse that each sample only connects to its k neighbors in the feature space, the rough "neighbors" could be obtained easily. Based on this, we query pairwise constraints following that: i) the must-link constraints are obtained by querying the "non-neighbors" of the samples; ii) the cannot-link constraints are obtained by querying the "neighbors" of the samples. In this way, we can find the key pairwise samples which are easily mis-linked or mis-unlinked in the original feature space. Compared to the random querying, the key querying simply makes the best of the pre-constructed graph and has a great significance in prompting clustering.

3.4 Theoretical Analysis

Firstly, we discuss the computational complexity. Since A is sparse and our algorithm only computes the k nonzero connections of each row in A, updating S requires $\mathcal{O}(n^2k)$. In terms of solving y_k, the inverse operation and the label propagation in model (14) requires the complexity of $\mathcal{O}((n-2)^2k)$. The optimization of F is an eigen-decomposition, requiring $\mathcal{O}(n^2c)$. Totally, the main computational complexity of our algorithm is $\mathcal{O}(n^2ct + (n-2)^2pt)$, where p is the number of cannot-link constraints and t is the iteration number. Secondly, we talk about the effect of the parameters λ, τ and γ respectively. Actually, τ is a constant that sets the edges between must-linked instances. Since we partition data points according to the graph connectivity which is independent of the intensity of edges, the value of τ does not effect the performance of our algorithm. In terms of λ, as we said, a large enough λ ensures that S possesses c connected components exactly. However, how large λ should be is difficult to seek. Thus, we adopt a widely used manner to determine λ heuristically [Nie *et al.*, 2014]. Specifically, we first initialize λ with a small value like 0.1, and update it according to the number of eigenvalue zero of L_S in the iterations. If this number is smaller than c, λ is multiplied by 2; or if it is greater than $c+1$, λ is divided by 2, otherwise we terminate the iterations. In terms of γ, when $\gamma \to \infty$, the regularization term infinitely approaches to zero and all of the cannot-link constraints would be satisfied. However, it cannot be neglected that when the cannot-link regularization is addressed too seriously, it would cause meticulous cluster results, and the clustering performance could be degraded. So, γ should be appropriate that is neither too small nor too large.

4 Experiments

We conducted extensive experiments to validate the advantages of the proposed PCOG and the proposed key pairwise constraints selection strategy.

Settings. The proposed PCOG is evaluated in two aspects: one is the clustering performance, and the other is the ability of satisfying the given pairwise constraints. The clustering performance is calculated by clustering ACCuracy (ACC) and the Normalized Mutual Information (NMI) [Strehl and

	Nam.	Siz.	Dim.	Cla.
UCI Datasets	Dermatology	366	34	6
	Control	600	60	6
	Monk1	432	6	2
	Glass	214	9	6
Image Datasets	ORL	400	1024	40
	COIL20	1440	1024	20
	UMIST	1400	1024	200
	USPS	2007	256	10
	YALE	165	1024	15

Table 1: The numerical introduction to real datasets.

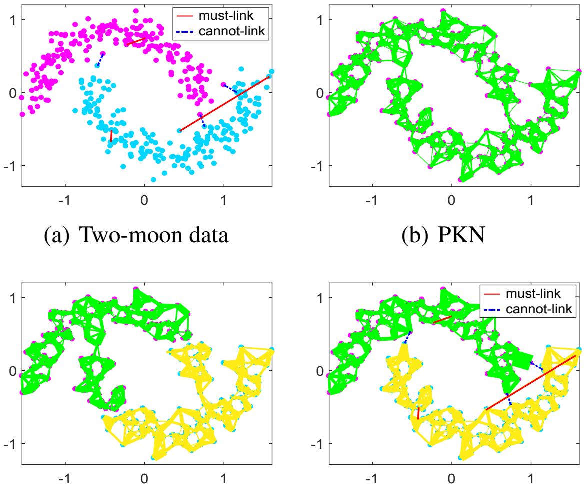
Ghosh, 2003]. ACC computes the percentage of correctly clustered samples. and NMI computes the mutual information between cluster labels and real labels.

Datasets. The proposed PCOG along with the compared approaches are tested on both toy data and real-world datasets. The real world datasets include four UCI datasets [Dua and Graff, 2017] (Dermatology, Control, Monk1 and Glass) and five image datasets (ORL [Samaria and Harter, 1994], COIL20 [Nene *et al.*, 1996], UMIST [Graham and Allinson, 1998], USPS [Hull, 1994] and YALE [Minear and Park, 2004], as described in Table 1.

Comparisons. We compare PCOG with five representative methods including three most related semi-supervised graph clustering approaches (SSGCK, CSCAP and CSP) [Lu and Carreira-Perpinan, 2008; Kulis *et al.*, 2009; Wang *et al.*, 2014]; the unsupervised graph clustering via Laplacian rank constraint (CLR) [Nie *et al.*, 2016], and the classical spectral clustering (Ncut) [Shi and Malik, 2000].

4.1 Experimental Results on Toy data

To verify the effectiveness of the proposed cannot-link graph regularization in PCOG, the toy experiment was Figure 3(a) illustrates the toy data distributed in the shape of two closed half-moons as well as the used must-link constraints and cannot-link constraints. Figure 3(b) illustrates the used graph A constructed by PKN ($k = 5$) (i.e., a graph construction manner proposed in CLR [Nie *et al.*, 2016] on the original feature space. Figure 3(c) and Figure 3(d) respectively represent the results of CLR and PCOG. It is concluded that 1) the achieved pairwise constraints are between either easily mis-linked points or easily mis-unlinked points as expected; 2) the input graph A constructed in the original feature space cannot reflect any cluster structure; 3) by only using the original features, the unsupervised graph clustering CLR fails in partitioning the pairs of points who are close in the geometrical space but different in terms of labels; 4) via the cannot-link graph regularization, the proposed PCOG ensures each cannot-link constraint, making the associated points into different clusters; 5) incorporated with the must-link constraints, PCOG dramatically enhances the clustering performance where a mass of connected points in the ambiguous areas could be correctly clustered.



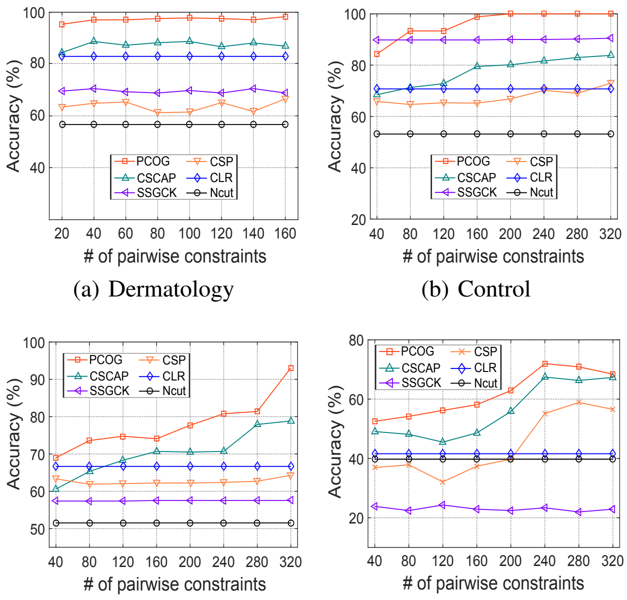
(a) Two-moon data

(b) PKN

(c) CLR

(d) PCOG

Figure 3: Clustering results on two-moon toy data: (a) original samples and pairwise constraints; (b) the given affinity graph A; (c) not using pairwise constraints; (d) using pairwise constraints.



(a) Dermatology

(b) Control

(c) Monk1

(d) Glass

Figure 4: Comparisons of clustering performance w.r.t. the number of pairwise constraints on four UCI datasets .

4.2 Experimental Results on Real data

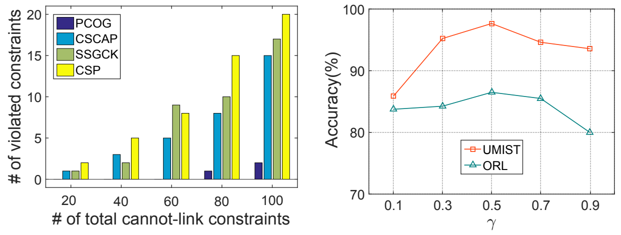
In this part, we firstly evaluate the clustering performance on four UCI datasets w.r.t. the pairwise constraints, as shown in Figure 4. We fix that a quarter of the total pairwise constraints are CL constraints, and the rest are ML constraints. From the results, it is observed that the proposed method outperforms other methods distinctly. As the number of the pairwise constraints increasing, the clustering accuracy of PCOG and CSCAP has remarkable improvement while CSP and SSGCK are steady. This is because PCOG and CSCAP propagate the cannot-link affinity in the graph, and thus they can achieve

	ACC				
Method	ORL	COIL20	UMIST	USPS	YALE
CLR	0.585	0.865	0.725	0.599	0.393
SSGCK	0.663	0.794	0.820	0.534	0.480
CSCAP	0.725	0.885	0.780	0.583	0.539
CSP	0.595	0.834	0.878	0.630	0.451
PCOG	**0.835**	**1.000**	**0.930**	**0.710**	**0.600**
	NMI				
Method	ORL	COIL20	UMIST	USPS	YALE
CLR	0.768	0.941	0.874	0.665	0.430
SSGCK	0.787	0.742	0.853	0.579	0.541
CSCAP	0.860	0.891	0.801	0.618	0.607
CSP	0.752	0.862	0.901	0.655	0.480
PCOG	**0.925**	**1.000**	**0.957**	**0.730**	**0.750**

Table 2: The clustering ACC and NMI on five image datasets.

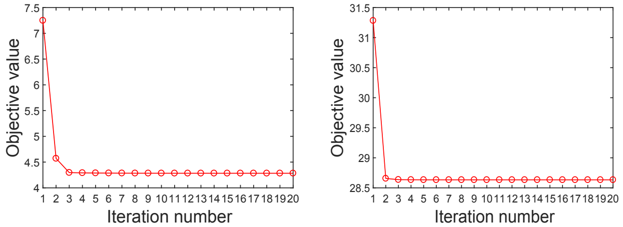
the valid supervision via a few constraints. The methods like CSP and SSGCK have to depend on a lot of given pairwise constraints. Noting that despite an unsupervised manner, the performance of CLR is dramatic compared to other semi-supervised methods, showing the validity of Laplacian rank constraint on the graph. Obviously, PCOG defeats CLR in the real datasets as well.

Secondly, we evaluate the proposed PCOG in five image datasets, covering the face images (i.e., ORL, UMIST and YALE), the object images (i.e., COIL20) and also the handwritten images (i.e., USPS). All of these algorithms use the same key pairwise constraint sets consisting of 80 cannot-link constraints and 240 must-link constraints, except for the unsupervised CLR. The clustering ACC and NMI of different approaches are recorded in Table 2. The best results are highlighted and the second best results are underlined. It is observed that the proposed PCOG outperforms all the competitors. Moreover, the clustering results in terms of ACC and NMI are improved about 10 percent compared to the second best results, showing the dramatic performance of the proposed algorithm. Additionally, to compare the ability of algorithms in satisfying the given CL constraints, Figure 5(a) recorded the number of violated cannot-link of four pairwise constrained clustering approaches by using ORL. It is obvious that the performance of PCOG surpasses others distinctly. Although CSCAP is the most well known affinity propagation algorithm, its performance of guaranteeing the constraints is very limited compared to the proposed method.



(a) (b)

Figure 5: (a) The number of the violated CLs w.r.t. the number of total CLs on ORL. (b) The clustering accuracy of PCOG w.r.t. γ.



(a) Glass (b) Dermatology

Figure 6: The convergence demonstration of Algorithm 1.

4.3 Parameter Selection and Convergence Study

We have theoretically analyzed that γ intensively impacts our clustering performance. So, we first seek γ in a large range of $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$, and we find that our algorithm works well in a small range $[0.1, 1]$. As a result, we further search γ from 0.1 to 1 with the interval of 0.2, as shown in Figure 5. The experimental curves obviously reflect the consistent conclusion with the theoretical analysis before. It is also shown that we can obtain the well performance in the range of [0.3, 0.7], and thus we search γ from 0.3 to 0.7 to obtain the best results in other experiments. We fix λ as a constant to investigate the convergence of Algorithm 1 experimentally. Figure 6 shows the objective value of model (7) within twenty iterations on two datasets. From the results, we observe that the proposed algorithm not only decreases the objective value but also converges fast, which demonstrates the high efficiency of the proposed algorithm.

5 Conclusion and Future Works

Pairwise constrained clustering is a vital technique in many realistic tasks. However, how to make the best of the cannot-link constraints is a long-term challenge since the cannot-link constraints are difficult to transform and propagate efficiently. In this paper, we present a method that addresses the cannot-link constraints problem via a specific cannot-link graph regularization, provably tackling the cannot-link constraints in the graph learning. We accordingly provide a matchable pairwise constraint selection for graph-based clustering methods. The superiority of the proposed method is verified in both toy data and nine real world datasets, improving the performance of semi-supervised clustering significantly. In the future, the research will proceed and we focus on two remained challenges: i) automatically determining γ in PCOG, releasing the algorithm from parameters; ii) enhancing the efficiency, such that the algorithm is applicable to large-scale data.

Acknowledgements

This work was supported in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX201918, in part by the National Key Research and Development Program of China under Grant 2018AAA0101902, in part by the National Natural Science Foundation of China under Grants 61936014, 61772427 and 61751202, in part by the Fundamental Research Funds for the Central Universities under Grant G2019KY0501.

References

[Craenendonck *et al.*, 2017] Toon Van Craenendonck, Sebastijan Dumancic, and Hendrik Blockeel. Cobra: A fast and simple method for active clustering with pairwise constraints. In *Proceedings of International Joint Conference on Artificial Intelligence*, 2017.

[Cucuringu *et al.*, 2016] Mihai Cucuringu, Ioannis Koutis, Sanjay Chawla, Gary Miller, and Richard Peng. Simple and scalable constrained clustering: a generalized spectral method. In *Artificial Intelligence and Statistics*, pages 445–454, 2016.

[Dua and Graff, 2017] Dheeru Dua and Casey Graff. Uci machine learning repository. http://archive.ics.uci.edu/ml, 2017.

[Fan, 1950] Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations: Ii. *Proceedings of the National Academy of Sciences of the United States of America*, 36(1):31, 1950.

[Graham and Allinson, 1998] Daniel B. Graham and Nigel M. Allinson. Characterising virtual eigensignatures for general purpose face recognition. https://www.sheffield.ac.uk/eee/research/iel/research/face, 1998.

[Hagen and Kahng, 1992] Lars Hagen and Andrew B Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 11(9):1074–1085, 1992.

[Hull, 1994] Jonathan J. Hull. A database for handwritten text recognition research. https://www.usps.com/nationalpremieraccounts/findzipcodes.htm, 1994.

[Kulis *et al.*, 2009] Brian Kulis, Sugato Basu, Inderjit Dhillon, and Raymond Mooney. Semi-supervised graph clustering: a kernel approach. *Machine learning*, 74(1):1–22, 2009.

[Li *et al.*, 2017] Xiang Li, Yao Wu, Martin Ester, Ben Kao, Xin Wang, and Yudian Zheng. Semi-supervised clustering in attributed heterogeneous information networks. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1621–1629. International World Wide Web Conferences Steering Committee, 2017.

[Lu and Carreira-Perpinan, 2008] Zhengdong Lu and Miguel A Carreira-Perpinan. Constrained spectral clustering through affinity propagation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[Minear and Park, 2004] Meredith Minear and Denise C Park. A lifespan database of adult facial stimuli. http://vision.ucsd.edu/content/yale-face-database, 2004.

[Nene *et al.*, 1996] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php, 1996.

[Ng *et al.*, 2002] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[Nie *et al.*, 2014] Feiping Nie, Xiaoqian Wang, and Heng Huang. Clustering and projected clustering with adaptive neighbors. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 977–986. ACM, 2014.

[Nie *et al.*, 2016] Feiping Nie, Xiaoqian Wang, Michael I. Jordan, and Heng Huang. The constrained laplacian rank algorithm for graph-based clustering. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[Saha *et al.*, 2016] Sriparna Saha, Abhay Kumar Alok, and Asif Ekbal. Brain image segmentation using semi-supervised clustering. *Expert Systems with Applications*, 52:50–63, 2016.

[Samaria and Harter, 1994] Ferdinando S Samaria and Andy C Harter. Parameterisation of a stochastic model for human face identification. https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html, 1994.

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

[Śmieja *et al.*, 2018] Marek Śmieja, Oleksandr Myronov, and Jacek Tabor. Semi-supervised discriminative clustering with graph regularization. *Knowledge-Based Systems*, 151:24–36, 2018.

[Strehl and Ghosh, 2003] Alexander Strehl and Joydeep Ghosh. *Cluster ensembles - a knowledge reuse framework for combining multiple partitions*. JMLR.org, 2003.

[Thangavel and Mohideen, 2010] K Thangavel and A Kaja Mohideen. Semi-supervised k-means clustering for outlier detection in mammogram classification. In *Trendz in Information Sciences & Computing (TISC2010)*, pages 68–72. IEEE, 2010.

[Tseng, 2001] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

[Wang *et al.*, 2014] Xiang Wang, Buyue Qian, and Ian Davidson. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, 28(1):1–30, 2014.

[Xiong *et al.*, 2014] Sicheng Xiong, Javad Azimi, and Xiaoli Z. Fern. Active learning of constraints for semi-supervised clustering. *IEEE Transactions on Knowledge & Data Engineering*, 26(1):43–54, 2014.

[Zhu and Ghahramani, 2002] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Citeseer, 2002.