Semi-supervised Clustering by Input Pattern Assisted Pairwise Similarity Matrix Completion

Jinfeng Yi YIJINFEN@MSU.EDU
Lijun Zhang ZHANGLIJ@MSU.EDU
Rong Jin RONGJIN@CSE.MSU.EDU
Qi Qian QIANQI@MSU.EDU
Anil K. Jain JAIN@CSE.MSU.EDU
Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 USA

Abstract

Many semi-supervised clustering algorithms have been proposed to improve the clustering accuracy by effectively exploring the available side information that is usually in the form of pairwise constraints. However, there are two main shortcomings of the existing semi-supervised clustering algorithms. First, they have to deal with non-convex optimization problems, leading to clustering results that are sensitive to the initialization. Second, none of these algorithms is equipped with theoretical guarantee regarding the clustering performance. We address these limitations by developing a framework for semi-supervised clustering based on *input pattern assisted matrix completion*. The key idea is to cast clustering into a matrix completion problem, and solve it efficiently by exploiting the correlation between input patterns and cluster assignments. Our analysis shows that under appropriate conditions, only $O(\log n)$ pairwise constraints are needed to accurately recover the true cluster partition. We verify the effectiveness of the proposed algorithm by comparing it to the state-of-the-art semi-supervised clustering algorithms on several benchmark datasets.

1. Introduction

Data clustering is an important task that has found numerous applications in many domains, including computer vision (Frigui & Krishnapuram, 1999), information retrieval (Bhatia & Deogun, 1998; Liu & Croft, 2004), recommender systems (Li & Kim, 2003), etc. But, on the other hand, data clustering is also an ill-posed problem due to its unsupervised nature (Jain, 2010). Semi-supervised clustering (Basu et al., 2002) addresses this limitation by effectively exploring the available side information that is often cast in the form of pairwise constraints: must-links for pairs of data points that belong to the same cluster and cannot-links for pairs of data points that belong to different clusters. The key idea of these algorithms is to search for the optimal data partition that is consistent with both the given pairwise constraints and the input data points to be clustered.

Despite the progress, there are two main shortcomings with the available semi-supervised clustering algorithms. First, most semi-supervised clustering algorithms have to deal with non-convex optimization problems, leading to clustering results that are only locally optimal and sensitive to the initialization. Second, although many computational algorithms have been proposed for semi-supervised learning, none of them is equipped with a theoretical guarantee on clustering performance. In particular, it is unknown how the clustering performance is improved with increasing number of pairwise constraints, an issue that is usually referred to as *sample complexity* in supervised learning (Bartlett, 1998).

In this work, we aim to address these limitations by developing a new framework for semi-supervised learning based on the theory of matrix completion (Candès & Tao, 2010). The proposed framework aims to reconstruct the pairwise similarity matrix, that gives 1 for any two data points in the same cluster and 0 otherwise, based on the given constraints and the input patterns of the objects to be clustered. The proposed

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

framework results in a convex optimization problem and, consequentially, globally optimal solutions. More importantly, the proposed work is equipped with a strong theoretical guarantee: with a high probability, the proposed algorithm can accurately recover the true data partition provided (i) the cluster membership vectors can be well approximated by the top singular vectors of the data matrix, and (ii) the number of pairwise constraints is sufficiently large. In particular, we show that under appropriate conditions, the true data partition can be *perfectly* recovered by the proposed algorithm with $O(rk \log n)$ pairwise constraints, where n is the number of data points to be clustered, r is the number of clusters, and k is the number of singular vectors used to approximate the cluster memberships. The logarithmic dependence on n makes the proposed algorithm particularly suitable for clustering large data sets.

2. Related work

Most semi-supervised clustering algorithms can be classified into two categories (Bilenko et al., 2004): constrained clustering and distance metric based semi-supervised clustering.

The key idea of constrained clustering is to directly incorporate the pairwise constraints into the existing clustering algorithms. The hard constrained clustering algorithms (Wagstaff et al., 2001; Shental et al., 2003; Allab & Benabdeslem, 2011) only consider the cluster assignments that are consistent with *all* the pairwise constraints, while the soft constrained clustering algorithms (Basu et al., 2004a; Lu & Leen, 2004; Basu et al., 2004a; Lu & Leen, 2004; Law et al., 2005; Davidson & Ravi, 2005; Law et al., 2005; Bekkerman & Sahami, 2006) penalize the clustering results based on the number of violated constraints.

The second group of semi-supervised clustering algorithms is based on distance metric learning (Xing et al., 2002). These algorithms first learn a distance metric from the given pairwise constraints. It then derives a linear transform from the learned distance metric, and applies it to generate a new vector representation for the data points to be clustered. The final data partition is computed by applying the existing clustering algorithms to the transformed vector representation. Various distance metric learning algorithms have been applied to semi-supervised clustering (Xing et al., 2002; Bar-Hillel et al., 2005; Hoi et al., 2006; Weinberger et al., 2006; Davis et al., 2007). Finally, several hybrid approaches have been developed for semi-supervised clustering that aim to combine the strength of constrained clustering with that of distance metric learning (Bilenko et al., 2004; Basu et al., 2004b).

Matrix completion (Candès & Tao, 2010) was originally proposed for collaborative filtering (Goldberg et al., 1992), where the goal is to predict the ratings of users for all the items given the ratings for a subset of randomly sampled items. It was recently exploited for graph-based clustering (Jalali et al., 2011), ensemble clustering (**?**) and crowdsourced clustering (Yi et al., 2012). The key difference between this work and the existing studies of clustering by matrix completion is that we explicitly incorporate the input patterns of objects into matrix completion, which not only improves the computational efficiency but, more importantly, reduces the number of pairwise constraints required for semi-supervised clustering. In addition, we address semi-supervised clustering in this work while the other studies focused on the standard clustering problem.

3. Semi-supervised Clustering by Input Pattern Assisted Matrix Completion

We first present a matrix completion based framework for semi-supervised clustering. We then present the proposed algorithm for semi-supervised clustering.

3.1. A Matrix Completion Framework for Semi-supervised Clustering

Let $\mathcal{D} = \{O_1, \ldots, O_n\}$ be the set of n objects to be clustered, and let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be their feature representation, where $\mathbf{x}_i \in \mathbb{R}^d$ is a vector of d dimensions. Let $\mathcal{M}$ denote the set of must-link constraints where $(i, j) \in \mathcal{M}$ implies that $\mathbf{x}_i$ and $\mathbf{x}_j$ should be in the same cluster, and $\mathcal{C}$ denote the set of cannot-link constraints, where $(i, j) \in \mathcal{C}$ implies that $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different clusters. For the convenience of presentation, we also define set $\Omega = \mathcal{M} \cup \mathcal{C}$ to include all the pairwise constraints. Let r be the number of clusters, and $n_{\min}$ be the size of the smallest cluster. The objective of semi-supervised clustering is to partition n data points into r clusters that are consistent with (i) the pairwise constraints in $\mathcal{M}$ and $\mathcal{C}$, and (ii) the data matrix X such that data points with similar input patterns are put into the same cluster.

Let $\mathbf{u}_i \in \{0, 1\}^n$ be the membership vector of the i-th cluster, where $u_{i,j} = 1$ if $\mathbf{x}_j$ is assigned to the i-th cluster and zero, otherwise. Define the pairwise similarity matrix $S \in \{0, +1\}$ as

$$S = \sum_{i=1}^{r} \mathbf{u}_i \mathbf{u}_i^\top$$

Evidently, $S_{i,j} = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are assigned to the

same cluster, and zero, otherwise. It is easy to verify the rank of matrix S is r. The given must-links in $\mathcal{M}$ and cannot-links in $\mathcal{C}$ provide partial observations for M, i.e. $S_{i,j} = 1$ if $(i,j) \in \mathcal{M}$ and $S_{i,j} = 0$ if $(i,j) \in \mathcal{C}$. Since finding the best data partition is equivalent to recovering the binary similarity matrix S, following (Jalali et al., 2011; Yi et al., 2012), we cast the semi-supervised clustering problem into a matrix completion problem, i.e. filling out the missing entries in binary similarity matrix S based on the pairwise constraints in $\mathcal{M}$ and $\mathcal{C}$ (i.e. the partial observations of S) and the data matrix X.

Similar to the standard theory for matrix completion (Candès & Tao, 2010), we can accurately recover the binary similarity matrix S because S is of low rank. We, however, note that the matrix completion problem discussed in this work is different from the previous studies of using matrix completion for clustering (Jalali et al., 2011; Yi et al., 2012) in that we aim to complete the binary similarity matrix S by utilizing both the observed entries in S and the input patterns in X. It will be shown later, both theoretically and empirically, that by effectively exploring the input patterns in X, the proposed algorithm is able to reduce the sample complexity for matrix completion from $O(n[\log n]^2)$ to $O(\log n)$, making it possible to apply the proposed algorithm to cluster very large data sets.

3.2. Input Pattern Assisted Matrix Completion

In this subsection, we first present input pattern assisted matrix completion for semi-supervised clustering. We then describe an efficient algorithm for solving the related optimization problem.

In the standard matrix completion theory (Candès & Tao, 2010), to reconstruct a matrix P of size $n \times n$ from a subset of observed entries in $\Delta \subseteq [n] \times [n]$, we solve the following optimization problem

$$\min_{P \in \mathbb{R}^{n \times n}} |P|_{tr} \text{ s. t. } \mathcal{R}_\Delta(P) = \mathcal{R}_\Delta(S) \tag{1}$$

where $|\cdot|_{tr}$ is the trace norm, and $\mathcal{R}_\Delta(S) : \mathbb{R}^{n \times n} \mapsto \mathbb{R}^{n \times n}$ is a linear operator that maps a matrix S to a new matrix $\mathcal{R}_\Delta(S)$ given by

$$[\mathcal{R}_\Delta(S)]_{i,j} = \begin{cases} S_{i,j} & (i,j) \in \Delta \\ 0 & (i,j) \notin \Delta \end{cases}$$

According to (Candès & Tao, 2010), with a high probability, matrix P can be perfectly recovered by solving the optimization problem in (1) if the number of observed entries in Δ is $O(\mu(P)^2 r(P) n[\log n]^2)$, where $r(P)$ is the rank of P and $\mu(P)$ is the coherence measure of P. In the case of binary similarity matrix S, it is easy to verify that the coherence measure $\mu(S)$ is bounded by $\sqrt{n/[n_{\min} r]}$ and the rank of S equals to the number of clusters r. As a result, the number of pairwise constraints required for perfectly recovering the binary similarity matrix S is $O(\kappa n[\log n]^2)$, where $\kappa = n/n_{\min}$. When data points are evenly distributed over clusters, we observe that the number of pairwise constraints required by matrix completion increases at least linearly in the number of data points to be clustered, making it unscalable to large data sets.

We address this limitation by developing a matrix completion approach that explicitly incorporates the data matrix X into the matrix completion process. Let $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_k)$ include the first k left singular vectors of X, where $k \geq r$. We make the following crucial assumption about the relationship between X and S:

A1 $\{\mathbf{u}_i\}_{i=1}^r$ lie in the subspace spanned by $\{\mathbf{z}_i\}_{i=1}^k$,

a similar assumption used by the spectral clustering algorithm (Ng et al., 2001). Using assumption **A1**, we can write S as $S = ZMZ^\top$, where $M \in \mathbb{R}^{k \times k}$. Following the theory of matrix completion, we obtain the optimal M by solving the following optimization problem:

$$\min_{M \in \mathbb{R}^{k \times k}} |M|_{tr} \text{ s. t. } \mathcal{R}_\Omega(ZMZ^\top) = \mathcal{R}_\Omega(S) \tag{2}$$

where $\Omega \subseteq [n] \times [n]$ includes all the observed entries in S derived from the pairwise constraints in $\mathcal{M}$ and $\mathcal{C}$.

The following theorem shows the perfect recovery result for (2).

Theorem 1. *Let $\mu(Z)$ be the coherence measure for matrix Z given by*

$$\mu(Z) = \max_{1 \leq i \leq n} \frac{n}{k} |[ZZ^\top]_{i,i}|^2 \tag{3}$$

Define

$$\mu_0 = \max\left(\mu(Z), \sqrt{\frac{n}{r n_{\min}}}\right). \tag{4}$$

For fixed $\beta > 2$, define a and B as

$$a = \frac{1}{2}(1 + \log_2 k - \log_2 r) \tag{5}$$

$$B = \frac{512\beta}{3} \mu_0 r k \ln n \tag{6}$$

Then, under assumption **A1** *with a probability $1 - 4(a+1)n^{-\beta+1} - 2an^{-\beta+2}$, $M_* = Z^\top S Z$ is the unique optimizer to (2) provided $|\Omega| \geq aB$.*

Remark: Compared to the standard matrix completion theory, the sample complexity of input pattern assisted matrix completion is reduced from $O(rn[\log n]^2)$ to $O(k \log n \log k)$ if $\mu_0 = O(1)$. Thus, if $k = \Omega(r)$ and the number of clusters r is small, Theorem 1 implies that $O(\ln n)$ pairwise constraints are needed in order to obtain the perfect clustering result, provided assumption **A1** holds and the coherence measure μ_0 is small.

Evidently, **A1** is a strong assumption that usually does not hold in real world applications. We thus relax this assumption by assuming that the cluster membership vectors $\{\mathbf{u}_i\}_{i=1}^r$ can be well approximated by the top k singular vectors of X. More specifically, we define the projection operator P_k as $P_k = ZZ^\top$, and the projection errors for the cluster membership vectors as

$$\mathcal{E}^2 = \max_{1 \le i \le r} \frac{1}{n^2} \|\mathbf{u}_i - P_k \mathbf{u}_i\|_F^2 \tag{7}$$

Instead of assuming $\mathcal{E} = 0$ as assumption **A1**, we assume that $\mathcal{E}$ is small enough to allow for an accurate recovery of the binary similarity matrix S. Under this assumption, we modify the optimization problem in (2) as follows

$$\min_{M \in \mathbb{R}^{k \times k}} |M|_{tr} + \frac{C}{2} \left\| \mathcal{R}_\Omega(ZMZ^\top) - \mathcal{R}_\Omega(S) \right\|_F^2 \tag{8}$$

where parameter $C > 0$ is introduced to balance the tradeoff between finding the low rank matrix M and fitting the observed pairwise constraints. The following theorem shows that the binary similarity matrix S can be accurately recovered by (8) if (i) the approximation error $\mathcal{E}$ is small and (ii) $|\Omega|$, the number of pairwise constraints, is sufficiently large.

Theorem 2. *Let $\widehat{M}$ be the optimal solution to (8) and $\widehat{S} = Z\widehat{M}Z$ be the reconstructed similarity matrix. For a fixed $\beta > 2$, with a probability $1 - 4(a+1)n^{-\beta+1} - 2an^{-\beta+2}$, we have*

$$\|\widehat{S} - S\|_F \le \nu(k, r)\mathcal{E}$$

where

$$\nu(k, r) = 6\left(\sqrt{2k} + 4\sqrt{r}\right)\left(3 + \sqrt{r}\right)$$

provided $|\Omega| \ge aB$ and $C \ge 1/[\sqrt{r}\mathcal{E}]$

As indicated by the above theorem, with a sufficiently large number of pairwise constraints, we have $\|\widehat{S} - \mathcal{S}\|_F \propto \mathcal{E}$, implying a small difference between $\widehat{S}$ and S when the cluster membership vectors can be well approximated by the top k singular vectors of X.

Algorithm 1 Efficient Stochastic Subgradient Descent for Solving the Optimization Problem (8)

Input:
- $Z \in \mathbb{R}^{n \times k}$: first k left singular vectors of X
- $C > 0$: loss function parameter
- r: number of clusters
- T: number of iterations
- η_t: step size

Initialization: $U_0 = \mathbf{0}_{k \times r}$, $\Sigma_0 = \mathbf{0}_{r \times r}$, $V_0 = \mathbf{0}_{k \times r}$
for $t = 0, \ldots, T-1$ **do**
  Generate a $k \times r$ probing matrix H
  Set $\hat{U}_{t+1} = [U_t \Sigma_t, B_t]$, where $B_t = (U_t V_t^\top + C \cdot Z^\top(\mathcal{R}_\Omega(ZMZ^\top - S))Z)H$.
  Set $\hat{V}_{t+1} = [V_t \quad -\eta_t H]$
  QR factorization of $\hat{U}_{t+1}$: $\hat{U}_{t+1} = Q_U R_U$
  QR factorization of $\hat{V}_{t+1}$: $\hat{V}_{t+1} = Q_V R_V$
  Compute $K = R_U R_V^\top$
  SVD decomposition of K: $K = M\bar{\Sigma}_{t+1}N^\top$
  Set $\bar{U}_{t+1} = Q_U M$ and $\bar{V}_{t+1} = Q_V N$
  $U_{t+1} = \bar{U}_{t+1}(1{:}k, 1{:}r)$
  $\Sigma_{t+1} = \bar{\Sigma}_{t+1}(1{:}r, 1{:}r)$
  $V_{t+1} = \bar{V}_{t+1}(1{:}k, 1{:}r)$
  $M^{(t+1)} = \Pi(U_{t+1}\Sigma_{t+1}V_{t+1}^\top)$
end for

Let $\widehat{M}$ be the optimal solution for (8). The estimated binary similarity matrix is given by $\widehat{S} = Z\widehat{M}Z^\top$. Since $|\widehat{S} - S|_F$ is small and the eigenvectors of S correspond to the cluster membership vectors, we expect the first r eigenvectors of $\widehat{S}$ reveal the clustering structure of the data. As a result, we apply the spectral clustering algorithm to find the best data partition, i.e. we first compute the top r eigenvectors of $\widehat{S}$, and then run the k-means algorithm over the computed eigenvectors. To improve the computational efficiency, we apply the spectral clustering algorithm proposed in (Chen et al., 2011) that reduces computational cost by the matrix sparsification technique (von Luxburg, 2007) and the Nystrom approximation (Fowlkes et al., 2004).

We finally discuss how to efficiently solve the optimization problem in (8). We exploit the fast stochastic subgradient descent (FSGD) method developed in (Avron et al., 2012). Define

$$\mathcal{L}(M) = \frac{C}{2} \left\| \mathcal{R}_\Omega(ZMZ^\top) - \mathcal{R}_\Omega(S) \right\|_F^2.$$

At each iteration, the proposed algorithm samples a subset of rows from the binary similarity matrix S by introducing a probing matrix H. It then computes an unbiased estimate of the gradient $\nabla\mathcal{L}(M_t)$, denoted by $\tilde{\nabla}\mathcal{L}(M_t)$, based on the sampled rows. Given the unbiased estimate of gradient, solution M_t is updated

by $M_{t+1} = \Pi\left(M'_{t+1} = M_t - \eta\tilde{\nabla}\mathcal{L}(M_t)\right)$. Here, $\Pi(A)$ is a soft thresholding function and is defined as $\Pi(A) = \sum_{i=1}^{r}\max(\lambda_i - 1, 0)\mathbf{a}_i\mathbf{a}_i^\top$, where $(\mathbf{a}_i, \lambda_i), i = 1, \ldots, r$ are the top r singular vectors and singular values of A. Algorithm 1 shows the detailed steps of the proposed algorithm, where the notation $U(1{:}k, 1{:}r)$ represents the sub-matrix of U that includes the first k rows and the first r columns of U.

4. Analysis

In this analysis, we will focus on the result for the noisy case, namely where the cluster membership vectors can be well approximated by the top k singular vectors of X although they do not lie in the subspace spanned by the top k singular vectors. This is a more general case and the perfect recovery result in Theorem 1 follows immediately from Theorem 2 by setting $\mathcal{E} = 0$.

We need to define a few notations before presenting our analysis. We define two linear operators $P_T : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ and $P_{T^\perp} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ as follows:

$$P_T(A) = P_U A + AP_U - P_U AP_U \quad (9)$$
$$P_{T^\perp}(A) = (P_k - P_U)A(P_k - P_U) \quad (10)$$

where $P_U = UU^\top$ and $P_k = ZZ^\top$. The coherence measurement μ for binary similarity matrix S is given by

$$\mu(S) = \frac{n}{r}\max_{1\le i\le n}|P_U\mathbf{e}_i|^2 \le \frac{n}{rn_{\min}} \quad (11)$$

As a result, we have the following inequality for μ_0 defined in (4)

$$\mu_0 = \max\left(\mu(Z), \frac{n}{rn_{\min}}\right) \ge \max(\mu(Z), \mu(S))$$

Our strategy is to first identify the deterministic conditions for the optimal solution $M_* = Z^\top SZ$ to be close to $\widehat{M}$, and then confirm that these deterministic conditions will hold with high probability.

Theorem 3. *Under the assumptions*

1. *the number of pairwise constraints is sufficiently large, i.e.*
$$|\Omega| > \frac{512\mu_0^2 r(k-r)\ln n}{3} \quad (12)$$
2. *there exists a dual matrix $Y \in \mathbb{R}^{n\times m}$ satisfied the following condition*
$$\mathcal{R}_\Omega(Y) = Y,$$
$$\|P_T(Y) - UU^\top\| \le \sqrt{\frac{r}{2k}}, \quad (13)$$
$$\|P_{T^\perp}(Y)\| \le \frac{1}{2}$$
3. *for any nonzero $F \in \mathbb{R}^{n\times n}$ satisfying $F = P_k F P_k$, we have*
$$\|P_T(F)\|_F \le \gamma\|P_{T^\perp}(F)\|_F + 2\|\mathcal{R}_\Omega(F)\|_F, \quad (14)$$
where γ is given by
$$\gamma = 4\mu_0(k-r)\sqrt{\frac{2\log n}{3|\Omega|}} \quad (15)$$

Then, by setting $C = \frac{1}{\sqrt{r}\mathcal{E}}$,, we have

$$|S - \widehat{S}|_F \le \left[6\left(\sqrt{2k} + 4\sqrt{r}\right)\left(3 + \sqrt{r}\right)\right]\mathcal{E}$$

The proof can be found in the appendix. The following two theorems are developed to confirm that the conditions specified in Theorem 3 hold with a high probability.

Theorem 4. *With a probability $1 - 4n^{-\beta+1}$,for any $Z \ne 0$ satisfying $Z = P_U Z P_U$, we have*

$$\|P_T(Z)\|_F \le \gamma\|P_{T^\perp}(Z)\|_F + 2\|\mathcal{R}_\Omega(Z)\|_F$$

where γ is given in (15), provided $|\Omega| \ge \Omega_0$ and $|\Omega_1| \le \Omega_0$.

To verify if there exists a matrix Y that satisfies the condition in (14), we follow (Candès & Recht, 2011) and construct Y as follows. We randomly select $q\Omega_0$ entries from Ω, and divide the selected entries into q subsets of equal size, denoted by $\Omega_1, \ldots, \Omega_q$, with

$$|\Omega_i| = \Omega_0, \quad i = 1, \ldots, q.$$

We generate a sequence of $Y_t, t = 1, \ldots, q$ as follows

$$Y_t = \frac{n^2}{\Omega_0}\sum_{i=1}^{t}\mathcal{R}_{\Omega_i}(W_i)$$

where $W_1 = UU^\top$ and W_{t+1} is defined inductively as

$$\begin{aligned} W_{t+1} &= P_T(UU^\top - Y_t) \\ &= W_t - \frac{n^2}{\Omega_0}P_T\mathcal{R}_{\Omega_t}(W_t) \\ &= \left(P_T - \frac{n^2}{\Omega_0}P_T\mathcal{R}_{\Omega_t}P_T\right)W_t \end{aligned}$$

We construct Y as the last element of the sequence, i.e. $Y = Y_q$. Evidently, we have $Y = \mathcal{R}_\Omega(Y)$. The following theorems show that Y satisfies the other properties specified in (14)

Theorem 5. *With a probability $1 - 2qn^{-\beta+1}$, we have*

$$\|P_T(Y) - UU^\top\| \le \sqrt{\frac{r}{2k}}$$

if $q \ge a$.

Theorem 4 and Theorem 5 follows directly from the analysis from (Recht, 2011).

5. Experiments

In this section, we first conduct a simulated study to verify our theoretical claim, i.e. the sample complexity of the proposed semi-supervised clustering algorithm is only logarithmic dependence on n. We then compare the proposed algorithm to the state-of-the-art algorithms for semi-supervised clustering on several benchmark datasets.

5.1. Baselines, and Parameter Settings

Baselines. We compare the proposed semi-supervised clustering algorithm to the following six state-of-the-art algorithms for semi-supervised clustering, including three constrained clustering algorithms and three distance metric learning algorithms. The three constrained clustering algorithms are (a) **MPCK-means**, the metric pairwise constrained k-means algorithm (Bilenko et al., 2004), (b) **CCSKL**, constrained clustering by spectral kernel learning (Li & Liu, 2009), and (c) **PMMC**, pairwise constrained maximum margin clustering (Zeng & ming Cheung, 2012). The three state-of-the-art distance metric learning algorithms are (d) **DCA**, the discriminative component analysis (Hoi et al., 2006), (e) **LMNN**, the large margin nearest neighbor classifier (Weinberger et al., 2006), and (f) **ITML**, the information theoretic metric learning algorithm (Davis et al., 2007). In order to examine the effectiveness of pairwise constraints for clustering, we also include the baseline method, referred to as **Base**, that directly applies the spectral clustering algorithm to cluster data points without any constraints. We refer to the proposed semi-supervised clustering algorithm as Matrix Completion based Constraint Clustering, or **MCCC** for short.

Evaluation and Parameter Settings. Normalized mutual information (NMI for short) (Cover & Thomas, 2006) is used to measure the coherence between the inferred clustering and the ground truth categorization. To determine the parameter C in (8), we follow the heuristic used in (Yi et al., 2012; **?**) that chooses the best C that results in a balanced cluster distribution. Two criteria are used in determining the values for k. First, k should be small enough to make the Algorithm 1 efficient. Second, k should be reasonably large to make the projection errors relatively small. In our experiments, we set $k = \min(100, d)$, where d is the dimensionality of the datasets.

5.2. Experiment with Synthesized Data

We first conduct experiments with simulated data to verify that under the assumption **A1**, the proposed semi-supervised clustering algorithm can perfectly recover the true data partition with only $O(\log n)$ sampled pairwise constraints. To this end, for a fixed n, the number of data points to be clustered, we create a partition of five clusters of equal size. Let $\mathbf{u}_i \in \{0,1\}^n, i = 1, \ldots, 5$ represent the cluster membership vectors. The target matrix to be recovered is $S = \sum_{i=1}^{5} \mathbf{u}_i \mathbf{u}_i^\top$. We construct the input pattern matrix X^{syn} by first generating a Gaussian random matrix $G \in \mathbb{R}^{5\times 15}$, with $G_{i,j}$ drawn independently from a Gaussian distribution $\mathcal{N}(0,1)$, and setting $X^{\text{syn}} = UG$, where $U = (\mathbf{u}_1, \ldots, \mathbf{u}_5)$. We vary n in range $\{5,000, 10,000, 20,000, 50,000, 10,0000\}$. For each n, we search for the smallest number of pairwise constraints that results in the perfect partition (i.e. NMI = 1). Figure 1 shows that the number of required constraints increases linearly in $\log n$, thus verifying that the sample complexity is logarithmic in the number of data points to be clustered.

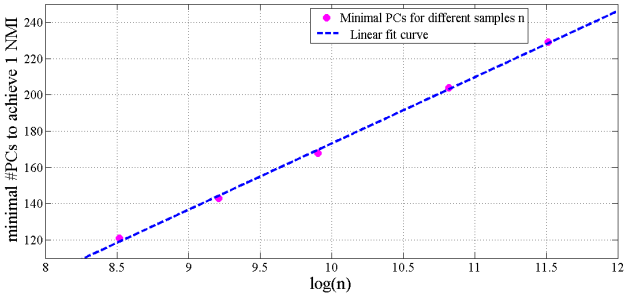

Figure 1. The plot of the smallest number of pairwise constraints (PCs) needed for perfect recovery. The correlation coefficient computed by the linear fit is 0.992, indicating a linear dependence of sample complexity in $\log n$.

Another advantage of the proposed algorithm is its scalability to large datasets since it only requires solving an optimization problem involving a small ($k \times k$, $k = \Omega(r)$) matrix. Table 1 summarizes the running time of recovering the synthetic data X^{syn} of different sizes, with the number of observed pairwise entries set to be the minimum required for perfect recovery. We observe that even for $n = 100,000$, it takes the proposed semi-supervised clustering algorithm less than an hour.

Table 1. Running time (in seconds) for recovering synthetic data of different size

n	5K	10K	20K	50K	100K
CPU time	24.0	77.1	217	1,086	3,429

5.3. Experiment with Benchmark Datasets

We evaluate the proposed semi-supervised clustering algorithm on several benchmark datasets. They are

Table 3. Average Clustering performance of the proposed semi-supervised clustering algorithm (MCCC) and the baseline algorithms (Base, MPCKmeans (MPCK) (Bilenko et al., 2004), CCSKL (Li & Liu, 2009), PMMC (Zeng & ming Cheung, 2012), DCA (Hoi et al., 2006), LMNN (Weinberger et al., 2006), and ITML (Davis et al., 2007)) on three datasets with 2,000, 4,000 and 6,000 randomly sampled pairwise constraints

Datasets	#pairwise constraints	**MCCC**	Base	MPCK	CCSKL	PMMC	DCA	LMNN	ITML
Mushrooms	2,000	**0.982**	0.540	0.645	0.652	0.876	0.873	0.980	0.971
	4,000	**0.991**	0.540	0.684	0.786	0.898	0.977	0.982	0.981
	6,000	**0.998**	0.540	0.713	0.754	0.923	0.988	0.983	0.984
USPS M2	2,000	0.979	0.866	0.950	0.979	0.976	0.971	0.976	**0.982**
	4,000	**0.984**	0.866	0.977	0.981	0.979	0.975	0.979	0.983
	6,000	**0.991**	0.866	0.989	0.982	0.987	0.981	0.985	0.986
Segment	2,000	**0.750**	0.651	0.693	0.721	0.718	0.723	0.714	0.706
	4,000	**0.755**	0.651	0.701	0.695	0.734	0.741	0.744	0.740
	6,000	**0.774**	0.651	0.718	0.684	0.748	0.760	0.751	0.743

Table 2. Description of Datasets

Name	#Instances	#Features	#Clusters
Mushrooms	8,124	112	2
USPS M2	2,822	256	2
Segment	2,310	19	7

(i) *Mushrooms database*[1] that contains 8,124 mushrooms belonging to 2 classes: poisonous or edible; (ii) *USPS M2 database*, that is comprised of 2,822 images belonging to the first two categories of USPS handwritten dataset (Hull, 1994); and (iii) *Segment database*[2] that contains 2,310 random segmentations of 7 outdoor images. Details of the three datasets are given in Table 2.

We vary the number of randomly sampled pairwise constraints from 2,000, 4,000 to 6,000 for each data sets. We note that we did not run experiments with smaller numbers of pairwise constraints because our theoretical analysis shows that the proposed algorithm is effective only when the number of constraints is sufficiently large. All the experiments are performed on a PC with Xeon 2.40 GHz processor and 64.0 GB memory. Each experiment is repeated five times, and the performance averaged over five trials is reported.

Table 3 summarizes the performance of the proposed semi-supervised clustering algorithm and the baseline algorithms. We first observed that although all the semi-supervised clustering algorithms significantly outperform the Base method with sufficiently large numbers of pairwise constraints, generally speaking, the distance metric based algorithms outperform the constrained clustering algorithms. We conjecture that this may be due to the fact that the number of pairwise constraints is large enough to learn a good distance metric such that data points of the same class will be separated by a small distance and data points from different classes are separated by a large distance. For the first two datasets Mushrooms and USPS M2, we observe that MCCC, the proposed semi-supervised clustering algorithm, achieves very high NMI value (close to 1) and outperforms all the baseline methods when the number of constraints is relatively large (i.e. 4,000 and 6,000). Among them, the experimental results for Mushrooms dataset is very encouraging since only 4,000 pairwise constraints are needed to achieve more than 0.99 NMI. This only accounts for about 0.012% of all possible pairwise constraints. For the Segment datasets, although the proposed algorithm is unable to achieve a (close to) perfect clustering, it still significantly outperforms all the baseline methods with all the number of pairwise constraints. The results in Table 3 demonstrates that the proposed algorithm is able to yield good clustering performance with sufficiently large numbers of pairwise constraints.

6. Conclusions

In this paper, we propose a framework for semi-supervised clustering based on input pattern assisted matrix completion. The key idea is to cast clustering into a matrix completion problem, and solve it efficiently by exploiting the correlation between input patterns and class assignments. Under the assumption that cluster membership vectors can be well approximated by the top few singular vectors of the data

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[2] http://archive.ics.uci.edu/ml/datasets/Image+Segmentation/

matrix, we show that with an overwhelming probability, the proposed algorithm can accurately recover the true data partition with only $O(\log n)$ randomly sampled pairwise constraints. Our empirical study verifies the effectiveness of the proposed algorithm.

In the future, we plan to further improve the efficiency for solving the optimization problem in (8) by exploiting various optimization techniques. We also plan to explore active learning technique to further reduce the sample complexity by actively selecting a subset of pairwise constraints. Furthermore, we plan to develop hybrid approaches that combine the power of input pattern assisted matrix completion with the strength of the other approaches for more effective semi-supervised clustering.

Acknowledgement:

This work is partially supported by Office of Navy Research (ONR Award N00014-11-1-0100).

A. Proof of Theorem 3

Define $S_* = Z^\top M_* Z$ and $F = Z\widehat{M}Z^\top - S_*$. Evidently, we have $F = P_Z F P_Z$. Using the condition in (14), we have

$$\|P_T(F)\|_F \leq \gamma \|P_{T^\perp}(F)\|_F + 2\|\mathcal{R}_\Omega(F)\|_F$$

Let $U_\perp$ be the eigenvectors of $P_{T^\perp}(Z)$. Evidently, column vectors in $U_\perp$ are orthogonal to the column vectors in U. We have

$$\begin{aligned}
|\widehat{M}|_{tr} = |\widehat{S}|_{tr} &\geq \langle S_* + Z, UU^\top + U_\perp U_\perp^\top \rangle \\
&\geq |S|_{tr} - |S_* - S|_{tr} + \langle Z, -Y + UU^\top + U_\perp U_\perp^\top \rangle \\
&\geq |S_*|_{tr} + \langle F, UU^\top - P_T(Y) + U_\perp U_\perp^\top - P_{T^\perp}(Y) \rangle \\
&\quad -2\sqrt{2r}\|S_* - S\|_F \\
&\geq |M_*|_{tr} + \|P_T(F)\|_F \|UU^\top - P_T(Y)\|_F - 2\sqrt{2}r\mathcal{E} \\
&\quad + (1 - \|P_T(Y)\|)\, \|P_{T^\perp}(F)\|_F \\
&\geq |M_*|_{tr} - 2\sqrt{\frac{r}{2k}}\|\mathcal{R}_\Omega(F)\|_F - 2\sqrt{2}r\mathcal{E} \\
&\quad + \|P_{T^\perp}(F)\|_F \left(\frac{1}{2} - \gamma\sqrt{\frac{r}{2k}}\right)
\end{aligned}$$

When $|\Omega| > \frac{512\mu_0^2 r(k-r)\log n}{3}$, we have

$$\begin{aligned}
|\widehat{M}|_{tr} &\geq |M_*|_{tr} - 2\sqrt{\frac{r}{2k}}\|\mathcal{R}_\Omega(F)\|_F \\
&\quad + \frac{\|P_{T^\perp}(F)\|_F}{4} - 2\sqrt{2}r\mathcal{E}
\end{aligned}$$

Since

$$\begin{aligned}
\mathcal{L}(\widehat{M}) &= \frac{C}{2}\|\mathcal{R}_\Omega(Z\widehat{M}Z^\top - S)\|_F^2 \\
&\geq \frac{C}{2}\left(\|\mathcal{R}_\Omega(S_* - S)\|_F - \|\mathcal{R}_\Omega(Z)\|_F\right)^2
\end{aligned}$$

and $C \geq \frac{1}{\sqrt{r}\mathcal{E}}$, it is easy to verify that

$$\|\mathcal{R}_\Omega(Z)\|_F \leq (12 + 2\sqrt{r})\sqrt{r}\mathcal{E}$$

and therefore

$$\begin{aligned}
\|P_{T^\perp}(Z)\| &\leq 8\sqrt{\frac{r}{2k}}\|\mathcal{R}_\Omega(Z)\|_F + 8\sqrt{2}r\mathcal{E} + Cr\mathcal{E}^2 \\
&\leq 24\sqrt{r}\mathcal{E}\left(3 + \sqrt{r}\right)
\end{aligned}$$

As a result, we have

$$\begin{aligned}
\|Z\|_F &\leq \|P_T(Z)\|_F + \|P_{T^\perp}(Z)\|_F \\
&\leq (\gamma + 1)\|P_{T^\perp}(Z)\|_F + 2\|\mathcal{R}_\Omega(Z)\|_F \\
&\leq \left[6\left(\sqrt{2k} + 4\sqrt{r}\right)\left(3 + \sqrt{r}\right)\right]\mathcal{E}
\end{aligned}$$

References

Allab, Kais and Benabdeslem, Khalid. Constraint selection for semi-supervised topological clustering. In *ECML/PKDD*, pp. 28–43, 2011.

Avron, H, Kale, S, Kasiviswanathan, S, and Sindhwani, V. Efficient and practical stochastic subgradient descent for nuclear norm regularization. In *ICML*, 2012.

Bar-Hillel, Aharon, Hertz, Tomer, Shental, Noam, and Weinshall, Daphna. Learning a mahalanobis metric from equivalence constraints. *JMLR*, 6:937–965, 2005.

Bartlett, P L. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

Basu, S, Banerjee, A, and Mooney, R J. Semi-supervised clustering by seeding. In *ICML*, pp. 27–34, 2002.

Basu, Sugato, Bilenko, Mikhail, and Mooney, Raymond J. A probabilistic framework for semi-supervised clustering. In *KDD*, pp. 59–68, 2004a.

Basu, Sugato, Bilenko, Mikhail, and Mooney, Raymond J. Probabilistic Semi-Supervised clustering with constraints. pp. 59–68, 2004b.

Bekkerman, R. and Sahami, M. Semi-supervised clustering using combinatorial MRFs. In *Proceedings of ICML-06 Workshop on Learning in Structured Output Spaces*, 2006.

Bhatia, Sanjiv K. and Deogun, Jitender S. Conceptual clustering in information retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 28 (3):427–436, 1998.

Bilenko, M, Basu, S, and Mooney, R J. Integrating constraints and metric learning in semi-supervised clustering. In *ICML*, 2004.

Candès, E J. and Tao, T. The power of convex relaxation: near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Candès, Emmanuel J. and Recht, Benjamin. Simple bounds for low-complexity model reconstruction. *CoRR*, abs/1106.1474, 2011.

Chen, Wen-Yen, Song, Yangqiu, Bai, Hongjie, Lin, Chih-Jen, and Chang, Edward Y. Parallel spectral clustering in distributed systems. *PAMI*, 33(3):568–586, 2011.

Cover, Thomas M. and Thomas, Joy A. *Elements of Information Theory (2nd ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9.

Davidson, Ian and Ravi, S. S. Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. In *PKDD*, pp. 59–70, 2005.

Davis, Jason V., Kulis, Brian, Jain, Prateek, Sra, Suvrit, and Dhillon, Inderjit S. Information-theoretic metric learning. In *ICML*, pp. 209–216, 2007.

Fowlkes, Charless, Belongie, Serge, Chung, Fan R. K., and Malik, Jitendra. Spectral grouping using the nyström method. *PAMI*, 26(2):214–225, 2004.

Frigui, Hichem and Krishnapuram, Raghu. A robust competitive clustering algorithm with applications in computer vision. *PAMI*, 21(5):450–465, 1999.

Goldberg, David, Nichols, David A., Oki, Brian M., and Terry, Douglas B. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, 1992.

Hoi, S.C.H., Liu, W., Lyu, M.R., and Ma, W.Y. Learning distance metrics with contextual constraints for image retrieval. In *CVPR*, pp. 2072–2078, 2006.

Hull, J.J. A database for handwritten text recognition research. *PAMI*, 16(5):550–554, 1994.

Jain, Anil K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

Jalali, Ali, Chen, Yudong, Sanghavi, Sujay, and Xu, Huan. Clustering partially observed graphs via convex optimization. In *ICML*, pp. 1001–1008, 2011.

Law, M, Topchy, A P., and Jain, A K. Model-based clustering with probabilistic constraints. In *SDM*, 2005.

Li, Q and Kim, B. Clustering approach for hybrid recommender system. In *Web Intelligence*, pp. 33–38, 2003.

Li, Zhenguo and Liu, Jianzhuang. Constrained clustering by spectral kernel learning. In *ICCV*, pp. 421–427, 2009.

Liu, Xiaoyong and Croft, W. Bruce. Cluster-based retrieval using language models. In *SIGIR*, pp. 186–193, 2004.

Lu, Zhengdong and Leen, Todd K. Semi-supervised learning with penalized probabilistic clustering. In *NIPS*, 2004.

Ng, Andrew Y., Jordan, Michael I., and Weiss, Yair. On spectral clustering: Analysis and an algorithm. In *NIPS*, pp. 849–856, 2001.

Recht, Benjamin. A simpler approach to matrix completion. *JMLR*, 12:3413–3430, 2011.

Shental, Noam, Bar-Hillel, Aharon, Hertz, Tomer, and Weinshall, Daphna. Computing gaussian mixture models with em using equivalence constraints. In *NIPS*, 2003.

von Luxburg, Ulrike. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Wagstaff, Kiri, Cardie, Claire, Rogers, Seth, and Schrödl, Stefan. Constrained k-means clustering with background knowledge. In *ICML*, pp. 577–584, 2001.

Weinberger, K.Q., Blitzer, J., and Saul, L.K. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.

Xing, E.P., Ng, A.Y., Jordan, M.I., and Russell, S. Distance metric learning, with application to clustering with side-information. *NIPS*, 15:505–512, 2002.

Yi, J., Jin, R., Jain, A. K., and Jain, S. Crowdclustering with sparse pairwise labels: A matrix completion approach. In *AAAI Workshop on Human Computation*, 2012.

Zeng, Hong and ming Cheung, Yiu. Semi-supervised maximum margin clustering with pairwise constraints. *TKDE*, 24(5):926–939, 2012.