Multitask Spectral Clustering by Exploring Intertask Correlation

Yang Yang, Zhigang Ma, Yi Yang, Feiping Nie, and Heng Tao Shen

***Abstract*—Clustering, as one of the most classical research problems in pattern recognition and data mining, has been widely explored and applied to various applications. Due to the rapid evolution of data on the Web, more emerging challenges have been posed on traditional clustering techniques: 1) correlations among related clustering tasks and/or within individual task are not well captured; 2) the problem of clustering out-of-sample data is seldom considered; and 3) the discriminative property of cluster label matrix is not well explored. In this paper, we propose a novel clustering model, namely multitask spectral clustering (MTSC), to cope with the above challenges. Specifically, two types of correlations are well considered: 1) intertask clustering correlation, which refers the relations among different clustering tasks and 2) intratask learning correlation, which enables the processes of learning cluster labels and learning mapping function to reinforce each other. We incorporate a novel $\ell_{2,p}$-norm regularizer to control the coherence of all the tasks based on an assumption that related tasks should share a common low-dimensional representation. Moreover, for each individual task, an explicit mapping function is simultaneously learnt for predicting cluster labels by mapping features to the cluster label matrix. Meanwhile, we show that the learning process can naturally incorporate discriminative information to further improve clustering performance. We explore and discuss the relationships between our proposed model and several representative clustering techniques, including spectral clustering, k-means and discriminative k-means. Extensive experiments on various real-world datasets illustrate the advantage of the proposed MTSC model compared to state-of-the-art clustering approaches.**

***Index Terms*—Clustering, multitask, out-of-sample.**

I. Introduction

CLUSTERING has been extensively explored as one of the most fundamental techniques in machine learning and data mining [1]. Various applications, such as image segmentation [2], [3], gene expression analysis [4], document analysis [5], content based image retrieval [6], image annotation [7]–[9], similarity searches [10], have witnessed the practical effectiveness of clustering.

k-means is one of the most classic data clustering algorithms and has been extensively applied in practice due to its effectiveness and simplicity. The typical procedure of traditional k-means (TKM) clustering algorithm iteratively assigns each data point to its closest cluster and computes a new clustering center. However, the "curse of dimensionality" may degrade the performance of TKM significantly [11]. Several research endeavors have been made to handle this problem by seeking a low-dimensional projection through dimensionality reduction, e.g., PCA, and then performing TKM. To step further, discriminative analysis [11]–[15] has been injected into TKM to enhance clustering performance. It has been shown that integrating TKM and LDA into a joint framework is beneficial. In [11] and [12], TKM and LDA were employed to obtain cluster labels and learn the most discriminative subspace in an alternating way. Ye *et al.* [14] proposed a joint framework, i.e., discriminative k-means (DKM) algorithm to formalize the clustering as a trace maximization problem.

Spectral clustering (SC) [2], [16], [17] has gradually become one of the most important clustering techniques and it shows more capability in partitioning data with more complicated structures compared to traditional clustering approaches. The underlying reason is that spectral clustering puts more efforts on mining the intrinsic data geometric structures [18]–[22]. SC has been widely applied and shown their effectiveness in various real-world applications, such as image segmentation [2], [16]. The fundamental idea of spectral clustering is that it predicts cluster labels by exploiting the different similarity graphs of data points. Besides NCut and k-way NCut, a new SC algorithm, i.e., local learning based clustering (LLC) [22], was developed according to the assumption that the cluster label of a data point can be determined by its neighbors, and a kernel regression model was used for label prediction.

Driven by the extensive availability of massive storage, fast networks and media sharing sites, we have witnessed an explosive growth of various related data sources on the Web in recent years. An interesting observation is that such different data embodies high relevance and similarity. For instance, in the application of human motion estimation, although different subjects varies significantly in the appearance due to the discrepancy of gender, style, clothing, etc., they are essentially similar to each other in some sense because their actions and/or behaviors follow similar patterns. Confronted

Manuscript received October 31, 2013; revised March 19, 2014 and June 24, 2014; accepted July 8, 2014. This work was supported by the ARC Discovery under Project DP130103252. The work of Y. Yang was supported by the Tianjin Key Laboratory of Cognitive Computing and Application. This paper was recommended by Associate Editor H. Qiao.

Y. Yang is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: dlyyang@gmail.com).

Z. Ma is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: kevinma@cs.cmu.edu).

Y. Yang and H. T. Shen are with the School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, QLD 4072, Australia (e-mail: yi.yang@uq.edu.au).

F. Nie is with the Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: feipingnie@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2344015

with these multiple large-scale data sources with high level of relevancy, it is imperative to develop more effective and efficient clustering scheme for simultaneously performing multiple clustering tasks on different data. However, it is nontrivial to achieve the above goal with the existing clustering approaches because they have been suffering from several new emerging challenges.

On the one hand, the intrinsic correlations among multiple clustering tasks on different yet related data, namely intertask correlation, are inevitably overlooked in traditional single-task clustering approaches, such as *k*-means and fuzzy c-means. In literature, it has been shown that combining multiple clustering tasks together may provide more potential reinforcing performance improvement [23]. The key issue is how to precisely model the intertask correlations and effectively apply them for reinforcing the performance of all the clustering tasks. To this end, several previous attempts have been made. Gu and Zhou [24] proposed to map data of different tasks into a common subspace or a common reproducing kernel Hilbert space (RKHS) [25], where standard clustering can be performed. Zhang and Zhou [26] tried to discover an optimal RKHS, where the mismatch of data distributions of pairwise tasks is minimized. Then, *k*-means clustering is performed in the discovered embedded space. Xie *et al.* [27] proposed a multitask co-clustering by integrating mutual information into the regularization term for adding intertask relations. Kong and Wang [28] used dictionary learning to simultaneously learn the "commonality" as a complemental representation for all the clusters, as well as a cluster-specific representation for each individual cluster. Dai *et al.* [29] proposed to co-cluster a target dataset as well as an "auxiliary" dataset by assuming that these two datasets share the same feature clusters.

On the other hand, although to some extent the intertask correlations are explored in previous works to enable multitask clustering, they have been consistently confronted with another serious challenge, i.e., no effective mechanism is afforded to deal with the out-of-sample data, which is especially significant confronted with the current evolution of Web data. A promising way is to learn an explicit mapping function for predicting cluster labels for the out-of-sample data outside the training data. Several approaches have been proposed to provide an additional step to patch the problem, such as Nyström method [30] and the work in [31]. However, such approaches normally separate learning cluster labels and learning mapping functions into two individual steps, thereby ignoring the relations between them. Moreover, while most of the previous works focus more on exploring the intertask relations, the within-task properties are not well considered. Although in traditional spectral clustering, by exploiting data local structures we may improve clustering performance to some extent, the task-specific property of the cluster indicator matrix requires more investigation. For instance, more discriminative information should be embedded in the cluster indicator matrix to make the clustering algorithms more effective and solid. Also, under some circumstances, the overfitting problems may occur and degrade the clustering performance due to the lack of appropriate processing.

Inspired by the aforementioned observations and analysis, in this paper, we propose a novel multitask spectral clustering (MTSC) model, which takes two types of relations into consideration: 1) intertask clustering correlation, which refers the relations among different clustering tasks and 2) intratask learning correlation, which enables the processes of learning cluster labels and learning mapping function to reinforce each other. Specifically, for each individual clustering task, we first propose to co-learn the cluster labels via traditional spectral clustering as well as the mapping function via an additional learning component (e.g., a regression model). It is notable that we do not only learn the mapping function but also integrate more discriminative information to improve the clustering performance. Then, in order to achieve the goal of performing multiple clustering tasks and making them reinforce each other, we propose to incorporate an $\ell_{2,1}$-norm regularization term over the regression coefficients of all the clustering tasks. An assumption behind the proposed regularizer is that the clustering tasks are similar or correlated so that they should share a common low-dimensional representation, which may benefit all the tasks. The benefits of the incorporated regularizer are twofold: 1) it explores the intertask correlations by identifying a common low-dimensional space for all the tasks and 2) together with the mapping function learning component, the regularization term can provide additional discriminative information as well as help avoid the overfitting problem. Further, we extend our formulation by replacing the $\ell_{2,1}$-norm with a more generalized $\ell_{2,p}$-norm, which enables more flexible control of the coherence of the intertask correlation as well as the expansion of the applicable range of our approach.

The contributions of this paper are summarized as follows.

1) We propose a novel joint MTSC framework to simultaneously perform clustering on multiple related tasks by exploring their intertask correlations as well as learn a task-specific mapping function for predicting the cluster labels for the out-of-sample data.
2) We propose to incorporate an $\ell_{2,1}$-norm regularization term over the regression coefficients of all the clustering tasks in order to identify a low-dimensional space for exploring the intertask correlations across all the clustering tasks.
3) An extension of our formulation, which replaces the $\ell_{2,1}$-norm with a more generalized $\ell_{2,p}$-norm, is proposed to enhance the proposed model with more flexible control of the coherence of the intertask correlation as well as more applicability.
4) Extensive experiments on multiple real-world datasets illustrate that our proposal outperforms the state-of-the-art clustering algorithms.

The rest of this paper is organized as follows. We first review the traditional spectral clustering in Section II. Section III elaborates details of the proposed MTSC model and explores the relationships between our proposal and the existing spectral clustering algorithms. Then, extensive experimental results are reported and analyzed in Section IV. Finally, we conclude our work in Section V.

II. Revisit of Spectral Clustering

Let us first review details of spectral clustering. Given a set of data points $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$ where d is the dimensionality of the data vector and n is the total number of data points. The original objective of clustering is to partition X into c groups $\{C_j\}_{j=1}^{c}$ such that data points within the same group are close while those in different groups are far from each other. Let us define $Y = [y_1, y_2, \ldots, y_n]^T \in \{0, 1\}^{n\times c}$, where $y_i \in \{0, 1\}^c$ $(i = 1, 2, \ldots, n)$ is the x_i's cluster indicator vector with the jth entry $y_{ij} = 1$ if $x_i \in C_j$ and $y_{ij} = 0$ otherwise. By following [14], we further define the scaled cluster indicator matrix F as below:

$$F = [F_1, F_2, \ldots, F_n]^T = Y(Y^TY)^{-1/2}$$

where F_i is the scaled cluster indicator of x_i. Note that the jth column of F indicates which data points belong to the jth cluster C_j, and it is in the following form:

$$f_j = [\underbrace{0, \ldots, 0}_{\sum_{k=1}^{j-1} n_j}, \underbrace{1/\sqrt{n_j}, \ldots, 1/\sqrt{n_j}}_{n_k}, \underbrace{0, \ldots, 0}_{\sum_{k=j+1}^{c} n_j}]$$

where n_j is the number of data points in the jth cluster.

Below is a general objective function of spectral clustering

$$\begin{aligned} &\min_{F} \ tr(F^TLF) \\ &\text{s.t. } F = Y(Y^TY)^{-1/2} \end{aligned} \tag{1}$$

where $tr(\cdot)$ is the trace operator and L denotes a graph Laplacian matrix computed according to the data local structure using different strategies. Given the dataset $X = [x_1, x_2, \ldots, x_n]$, we can construct an undirected graph which can be represented by the weighted adjacency matrix $S = (s_{ij})_{i,j=1,2,\ldots,n}$. Here $s_{ij} > 0$ indicates that x_i and x_j are connected, and $s_{ij} = 0$ means they are not connected.

A common way to compute the edge weight is defined as follows:

$$s_{ij} = \begin{cases} \exp\left(-\dfrac{\|x_i - x_j\|^2}{\sigma^2}\right), & \text{if } x_i \in \mathcal{N}_k(x_j) \text{ or } x_j \in \mathcal{N}_k(x_i) \\ 0, & \text{otherwise} \end{cases}$$

where $\mathcal{N}_k(\cdot)$ is the function for searching for the k nearest neighbors and σ is the bandwidth parameter. Denote D as a diagonal matrix with its diagonal $d_{ii} = \sum_j D_{ij}$, then the graph Laplacian can be calculated as

$$L = D - S.$$

If we instead use the normalized graph Laplacian in (1)

$$L_n = D^{-1/2}LD^{-1/2} = I_n - D^{-1/2}SD^{-1/2}$$

then the objective function turns out to be the well-known spectral clustering algorithm, namely normalized cut [2]. Similarly, if we replace L with L_l which is the graph Laplacian matrix obtained by the local learning [22], then the objective function in (1) becomes the local learning clustering (LLC).

Note that the discretization constraint on F makes (1) difficult to solve. A practical way to handle this problem is to make a relaxation to allow F to be of continuous values, and then use eigenvalue decomposition on the corresponding graph Laplacian matrix.

III. MTSC

In this section, we present a novel joint MTSC framework which effectively reinforces the performance of a series of related clustering tasks by exploring two types of correlations: 1) intertask clustering correlation, which refers the relations among different clustering tasks and 2) intratask learning correlation, which enables the processes of learning cluster labels and learning mapping function to reinforce each other. Moreover, for each individual clustering task, we incorporate more discriminative information and enhance the proposed MTSC model with the ability to cluster any unseen data samples.

A. Problem Formulation

We are faced with a set of M unsupervised clustering tasks $\{\mathcal{T}^{(1)}, \mathcal{T}^{(2)}, \ldots, \mathcal{T}^{(M)}\}$. For each individual task $\mathcal{T}^{(t)}(1 \le t \le M)$, we have a set of n_t training data samples $X^{(t)} = [x_1^{(t)}, x_2^{(t)}, \ldots, x_{n_t}^{(t)}]$, where $x_i^{(t)} \in \mathbb{R}^{d\times 1}(1 \le i \le n_t)$ and d is the dimensionality of the feature space.

Our main objective is to design an algorithm which is able to: 1) partition the training data in each task into $c_t(1 \le t \le M)$ clusters and 2) for each task identify an explicit mapping function $f_t : \mathbb{R}^{d\times 1} \to \{0, 1\}^{c_t\times 1}$ that is able to cluster out-of-sample data. To achieve the above goals, we prefer the proposed algorithm to effectively explore the correlations among all the clustering tasks.

B. Proposed Approach

Most existing clustering and manifold learning algorithms have been suffering from the out-of-sample issue [32]. In other words, they cannot assign cluster labels to the data outside the training data. This apparently limits these algorithms to handle large-scale clustering task. Saul and Roweis [33] proposed a nonparametric out-of-sample extrapolation method for manifold learning. The basic idea is to reconstruct a test data point with its k nearest neighbors. This algorithm can be generalized to spectral clustering algorithms for grouping out-of-sample data. However, a recognized problem is that the performance is still unstable since it heavily relies on the local structure. If the locally linear condition is not satisfied, the performance may degrade. In this case, a strong motivation is derived to simultaneously learn an additional mapping function $f : R^d \to \{0, 1\}^c$ in the clustering process.

Let us consider the problem of how to learn an explicit mapping function $f_t : \mathbb{R}^{d\times 1} \to \{0, 1\}^{c_t\times 1}$ for each task first. A straightforward way is to firstly employ a traditional spectral clustering approach to obtain the cluster indicator matrix $F^{(t)}$ for the tth task and then a linear regression model is learnt by solving the following optimization problem:

$$\min_{W^{(t)}} \left\| F^{(t)} - \left(X^{(t)}\right)^T W^{(t)} \right\|_F^2 + \beta\Omega\left(W^{(t)}\right) \tag{2}$$

where $W^{(t)} \in \mathbb{R}^{d\times c_t}$ represents the regression coefficients for the mapping function $f_t(\cdot)$, $\Omega(\cdot)$ denotes a certain regularization function over $W^{(t)}$ and β is a balance parameter. Indeed, the above solution is able to somehow achieve the goals, however, the intrinsic consistence of simultaneously learning the

cluster indicator matrix and the mapping function is overlooked. To overcome this drawback, we propose to combine the two learning processes to make them reinforce each other

$$\min_{F^{(t)},W^{(t)}} tr\left(\left(F^{(t)}\right)^T L^{(t)} F^{(t)}\right) + \alpha \left\| F^{(t)} - \left(X^{(t)}\right)^T W^{(t)} \right\|_F^2 + \beta\Omega\left(W^{(t)}\right)$$
$$\text{s.t. } (F^{(t)})^T F^{(t)} = I_t \tag{3}$$

where $L^{(t)}$ is the Laplacian matrix for the tth task, I_t is a $c_t \times c_t$ identity matrix, and α is a new balance parameter. As seen, the model in (3) is able to simultaneously learn the cluster label matrix together with an explicit mapping function for grouping any out-of-sample data. Besides, with a proper choice of the regularization term, (3) has been proved to be able to incorporate additional discriminative capacity into the traditional spectral clustering model [32], [34].

Actually, we may directly extend (3) by summing over the corresponding objective functions of all M tasks

$$\min_{\{F^{(t)},W^{(t)}\}_{t=1}^M} \sum_{t=1}^{M} \left(tr\left(\left(F^{(t)}\right)^T L^{(t)} F^{(t)}\right) + \alpha \left\| F^{(t)} - \left(X^{(t)}\right)^T W^{(t)} \right\|_F^2 + \beta \left\| W^{(t)} \right\|_F^2 \right)$$
$$\text{s.t. } (F^{(t)})^T F^{(t)} = I_t,\ t = 1, 2, \ldots, M. \tag{4}$$

Note that we choose to use ℓ_2-norm regularization term here. Apparently, the model in (4) is not a "real" multitask learning model because there is no specific component designed for exploring the correlations across all the tasks. We cannot use such model to achieve the goal of performing multiple clustering tasks.

Therefore, we propose a novel MTSC model, which incorporates a generalized $\ell_{2,1}$-norm regularizer to control the number of common features selected for all the tasks. More specifically, we uncover the correlations across all the clustering tasks based on the assumption that a small set of features should be shared by these tasks to learn the explicit mapping functions. Inspired by [35], we propose to incorporate the following $\ell_{2,1}$-norm regularization term over the regression coefficients of all the tasks:

$$\|W\|_{2,1} = \sum_{j=1}^{d} \left\| W_j \right\|_2 \tag{5}$$

where $W = [W^{(1)}, W^{(2)}, \ldots, W^{(M)}]$ and W_j is the jth row of W. The above $\ell_{2,1}$-norm regularizer is a generalization of the well-known ℓ_1-norm regularization which forces most of the elements to be zeros. In fact, It can be regarded as the ℓ_1-norm of the vector $[\|W_1\|_2, \|W_2\|_2, \ldots, \|W_d\|_2]$, each element of which denotes the importance of the corresponding feature. Hence, an intuitive explanation of the $\ell_{2,1}$-norm regularizer is that it will enforce most of the rows of W to be zero. Only those nonzero rows are consistently used through all the clustering tasks. By replacing the regularization term in (4) with (5), we arrive at

$$\min_{\{F^{(t)},W^{(t)}\}_{t=1}^M} \sum_{t=1}^{M} \left(tr\left(\left(F^{(t)}\right)^T L^{(t)} F^{(t)}\right) + \alpha \left\| F^{(t)} - \left(X^{(t)}\right)^T W^{(t)} \right\|_F^2 \right) + \beta \|W\|_{2,1}$$
$$\text{s.t. } (F^{(t)})^T F^{(t)} = I_t,\ t = 1, 2, \ldots, M. \tag{6}$$

The above model learns a limited set of features instead of the whole feature set, which would benefit the scalability of our approach. Meanwhile, the particular correlation among all the tasks is explicitly explored in this way by discovering a common low-dimensional space which is comprised of those selected nonzero rows.

The $\ell_{2,1}$-norm regularizer $\|W\|_{2,1}$ is able to help identify some common features shared across all the tasks, nonetheless, such solution is not flexible enough to control the coherence of these tasks. Inspired by [36], in order to handle this problem, we propose to further generalize the $\ell_{2,1}$-norm regularizer to an $\ell_{2,p}$-norm regularizer

$$\|W\|_{2,p} = \left(\sum_{j=1}^{d} \left\| W_j \right\|_2^p \right)^{1/p} \tag{7}$$

where $p \in (0, 2)$. By replacing the $\ell_{2,1}$-norm regularization term in (6) with the $\ell_{2,p}$-norm regularizer, we obtain the following updated formulation:

$$\min_{\{F^{(t)},W^{(t)}\}_{t=1}^M} \sum_{t=1}^{M} \left(tr\left(\left(F^{(t)}\right)^T L^{(t)} F^{(t)}\right) + \alpha \left\| F^{(t)} - \left(X^{(t)}\right)^T W^{(t)} \right\|_F^2 \right) + \beta \sum_{j=1}^{d} \left\| W_j \right\|_2^p$$
$$\text{s.t. } (F^{(t)})^T F^{(t)} = I_t,\ t = 1, 2, \ldots, M. \tag{8}$$

Note that when p is set to 1, (8) is identically equivalent to (6). When p is larger (i.e., $p \to 2$), the $\ell_{2,p}$-norm tends to be closer to ℓ_2-norm, which implies that the correlation among all the tasks becomes weaker. On the other hand, when p gets smaller, all the tasks are more correlated. However, it is nontrivial to determine the relationship among tasks in reality. As we know, if we have enough prior knowledge about the relationships among tasks, it may guide us to choose a more proper magnitude of correlation measure, namely p. For example, if we are aware that two tasks are not very related, we may tend to choose a larger p ($p \to 2$) to force not too much intertask correlations are explored; otherwise, different tasks may exert negative influence on each other. Thus, a proper p is important to the whole learning process. Although determining the task correlations is out of the scope of this paper, we may provide some possible ways to achieve this goal. A straightforward approach is to use "grid-search" to discover an optimal p. Also, we may also utilize certain statistical approach to predetermine the correlation between two sets of data, such as Pearson's product-moment coefficient and canonical correlation analysis. Then, we normalize the correlation to the range $(0, 2)$ to obtain a prior knowledge of p.

In the next part, we will develop an effective algorithm to solve the problem in (8).

C. Optimization

It is noted that solving the problem of (8) is difficult since: 1) it is nonconvex w.r.t. all the variables at the same time; 2) the nonsmooth property of the incorporated $\ell_{2,p}$-norm regularizer makes it difficult to optimize the problem as a whole; and 3) the problem is not separable w.r.t. each individual task, which makes it harder to design an iterative algorithm over all the tasks. In order to overcome the above challenges, we incorporate an additional updating variable, which encodes not only the importance of each feature but also the correlation information among all the tasks. In this subsection, we present an alternating algorithm to solve the problem and provide a rigorous mathematical analysis on the convergence of the algorithm.

Firstly, we present an alternative optimization problem in (8)

$$\begin{aligned}\min_{\{F^{(t)},W^{(t)}\}|_{t=1}^{M}} &\sum_{t=1}^{M}\left(tr\left(\left(F^{(t)}\right)^{T}L^{(t)}F^{(t)}\right)\right.\\ &\left.+\alpha\left\|F^{(t)}-\left(X^{(t)}\right)^{T}W^{(t)}\right\|_{F}^{2}\right)+\beta tr(W^{T}UW)\\ \text{s.t. } &(F^{(t)})^{T}F^{(t)}=I_t,\ t=1,2,\ldots,M \end{aligned} \tag{9}$$

where U is a $d\times d$ diagonal matrix with its jth diagonal element U_{jj} defined as

$$U_{jj}=\frac{1}{\frac{2}{p}\left\|W_j\right\|_2^{2-p}} \tag{10}$$

where W_j is the jth row of W. By setting the derivative of the objective function in (9) w.r.t. $W^{(t)}$ to 0, we obtain the regression coefficient of each individual task

$$\begin{aligned}&2\alpha\left(X^{(t)}(X^{(t)})^{T}W^{(t)}-X^{(t)}F^{(t)}\right)+2\beta UW^{(t)}=0\\ &\Rightarrow\ W^{(t)}=\alpha\left(\alpha X^{(t)}(X^{(t)})^{T}+\beta U\right)^{-1}X^{(t)}F^{(t)}\end{aligned} \tag{11}$$

where $t=1,2,\ldots,M$. By denoting $A^{(t)}=\alpha\left(\alpha X^{(t)}(X^{(t)})^{T}+\beta U\right)^{-1}$ and substituting (11) into (9) we obtain

$$\begin{aligned}\min_{F^{(t)},W^{(t)}}\ & tr\left((F^{(t)})^{T}L^{(t)}F^{(t)}\right)\\ &+\alpha\left\|F^{(t)}-(X^{(t)})^{T}A^{(t)}X^{(t)}F^{(t)}\right\|_{F}^{2}\\ &+\beta tr\left((F^{(t)})^{T}(X^{(t)})^{T}(A^{(t)})^{T}UA^{(t)}X^{(t)}F^{(t)}\right)\\ \text{s.t. }&(F^{(t)})^{T}F^{(t)}=I_t\end{aligned} \tag{12}$$

which can be further rewritten as

$$\begin{aligned}\min_{F^{(t)}}\ & tr\left((F^{(t)})^{T}H^{(t)}F^{(t)}\right)\\ \text{s.t. }&(F^{(t)})^{T}F^{(t)}=I_t\end{aligned} \tag{13}$$

where $H^{(t)}=\left(L^{(t)}-\alpha(X^{(t)})^{T}A^{(t)}X^{(t)}+\alpha I_t\right)$. It is well-known that the problem in (13) can be effectively solved by performing eigen-decomposition over $H^{(t)}$.

Note that U is actually dependant on W; nonetheless, if U is constant then we can easily solve the problem in (9) by (13) and (11) over all the tasks $t=1,2,\ldots,M$. Based on this analysis, we devise an efficient iterative algorithm (as shown in Algorithm 1) to solve (8). In each iteration, we firstly calculate U based on W, and then update $W^{(t)}$ and $F^{(t)}$ for each task in an alternating way. Next, we prove Algorithm 1 will converge when $p\in(0,2)$. To this end, we present the following lemmas.

Algorithm 1 Algorithm for Optimizing the Proposed MTSC Model

Input: M datasets $\{X^{(t)}\}|_{t=1}^{M}$ for M clustering tasks;
Output: Cluster indicator matrix $F^{(t)}$ and regression coefficients $W^{(t)}$ for each task ($1\le t\le M$);
1: **for** $t=1$ **to** M **do**
2: Construct Laplacian matrix $L^{(t)}$ based on $X^{(t)}$;
3: **end for**
4: Initialize $\{W^{(t)}\}|_{t=1}^{M}$;
5: **repeat**
6: $W=[W^{(1)},W^{(2)},\ldots,W^{(M)}]$;
7: Compute the diagonal matrix U based on W according to (10);
8: **for** $t=1$ **to** M **do**
9: Compute $A^{(t)}=\alpha\left(\alpha X^{(t)}(X^{(t)})^{T}+\beta U\right)^{-1}$ and $H^{(t)}=(L^{(t)}-\alpha(X^{(t)})^{T}A^{(t)}X^{(t)}+\alpha I_t)$;
10: Update $F^{(t)}$ by performing eigen-decomposition over $H^{(t)}$;
11: Update $W^{(t)}=A^{(t)}X^{(t)}F^{(t)}$;
12: **end for**
13: **until** convergence
14: **return** $\{F^{(t)}\}|_{t=1}^{M}$ and $\{W^{(t)}\}|_{t=1}^{M}$;

Lemma 1: Let W_j be the jth row of the updated W in previous iteration and $\tilde{W}_j$ be the jth row of the variable $\tilde{W}$ in current iteration, then the following inequality holds:

$$\|\tilde{W}_j\|_2^p-\frac{p\|\tilde{W}_j\|_2^2}{2\|W_j\|_2^{2-p}}\le\|W_j\|_2^p-\frac{p\|W_j\|_2^2}{2\|W_j\|_2^{2-p}}. \tag{14}$$

Proof: Please refer to Appendix A for more details. ■

Lemma 2: Given $W=[W_1^T,W_2^T\ldots,W_d^T]^T$, where W_j is the jth row of W, then we have the following conclusion:

$$\sum_{j=1}^{d}\|\tilde{W}_j\|_2^p-\sum_{j=1}^{d}\frac{p\|\tilde{W}_j\|_2^2}{2\left\|W_j\right\|_2^{2-p}}\le\sum_{j=1}^{d}\|W_j\|_2^p-\sum_{j=1}^{d}\frac{p\left\|W_j\right\|_2^p}{2\left\|W_j\right\|_2^{2-p}}. \tag{15}$$

Proof: According to Lemma 1, we can easily obtain the conclusion of Lemma 2 by summing up the inequalities over all the rows in W. ■

Theorem 1: At each iteration (line 6-11) of Algorithm 1, the value of the objective function in (8) monotonically decreases.

Proof: Please refer to Appendix B for more details. ■

It is clear that Theorem 1 guarantees that Algorithm 1 will converge an optima.

D. Complexity Analysis

In this part, we analyze the time complexity of our algorithm for both the training clustering and the testing clustering. Without loss of generality, suppose we have M tasks and each

task have N training samples. The key operation in each iteration is the update of $F^{(t)}$ (line 10), which is $O(N^3)$. Thus, the cost for updating $F^{(t)}$ of all the tasks is $O(M \times N^3)$. Suppose our algorithm converges to a local optima within l iterations, which indicates the overall cost of the training process is $O(l \times M \times N^3)$. In reality, we usually have $N \gg l$ and $N \gg l$, therefore, the cost of our algorithm is $O(N^3)$. As to the testing cost, the prediction of the cluster label for each individual data sample is $O(d \times c)$ where d is the feature dimension and c is the number of clusters. The overall testing cost for N_t out-of-sample data points is $O(N_t \times d \times c)$. Note that $N_t \gg d$ and $N_t \gg c$, so the testing complexity is approximately linear to N_t.

IV. Discussion

In this section, we analyze and discuss the relationship between the proposed MTSC model and the most popular clustering algorithms, such as the family of spectral clustering, k-means, DKM. Specifically, we illustrate our conclusions in the following propositions.

Proposition 1: The MTSC model leads to spectral clustering for each individual task when $\alpha = 0$ and $\beta = 0$.

Proof: Please refer to Appendix C for detailed proof. ■

It is clear that Proposition 1 implies the proposed MTSC model can naturally lead to any member of the spectral clustering family. For instance, if we set $\alpha = 0$ and $\beta = 0$, and replace each $L^{(t)}$ with the Laplacian matrix L_l proposed in local learning clustering (LLC) [22], then we prove that LLC is a special case of our model. Likewise, any other spectral clustering approach can be proved to be a special case of our model.

Proposition 2: The MTSC model leads to k-means clustering when $\frac{\alpha}{\beta} \to 0$, $\alpha \to \infty$ and $p = 2$.

Proof: Please refer to Appendix D for detailed proof. ■

Recently, some research progresses have been focusing on incorporating dimensionality reduction and clustering together into a joint framework to cope with high-dimensional data. Denote W as the projection matrix for dimensionality reduction, this family of algorithms, referred to as DKM, try to solve the following maximization problem:

$$\max_{W,F,F^TF=I} tr\left((W^T(XX^T + \gamma I)W)^{-1}W^TXFF^TX^TW\right). \quad (16)$$

As can be seen, in (16) two variables, i.e., W and F are jointly optimized. The original solution to this problem is to iteratively update W and F [12]. More recently, Ye *et al.* [14] proves that the optimization of the above problem in (16) can be simplified by only solving F in the following optimization function:

$$\max_{F,F^TF=I} tr(F^T(I - (I + \frac{1}{\gamma}X^TX)^{-1})F). \quad (17)$$

In the following, we explore the connection of the DKM and the proposed MTSC model.

Proposition 3: The MTSC model leads to Discriminative k-means clustering when $\beta \to \infty$ and $p = 2$.

Proof: Please refer to Appendix E for detailed proof. ■

TABLE I
Dataset Description

Dataset	Task	Size	Dimension	#Class
WebKB4	Task 1	226	2000	4
	Task 2	252	2000	4
	Task 3	255	2000	4
	Task 4	308	2000	4
HumanEva	Task 1	5000	168	5
	Task 2	5000	168	5
Comp.vs.Sci	Task 1	1875	2000	2
	Task 2	1827	2000	2
Rec.vs.Talk	Task 1	1844	2000	2
	Task 2	1545	2000	2

V. Experiments

In this section, we evaluate the effectiveness of the proposed MTSC model by comparing it to the state-of-the-art approaches, including both single-task and multitask algorithms. For single-task algorithms, TKM clustering, DKM clustering [14], kernel k-means clustering [37], fuzzy c-means clustering [38], k-way normalized cuts (Ncuts) [16], and CLGR [39] are used for comparison. Also, we compare to existing multitask algorithms, including LSSMTC [24], LSKMTC [25], and LNKMTC [25]. We also compare to the spectral embedding clustering (SEC) [21], [32], which can be regarded as the single task version of our model.

A. Data and Experiment Settings

In our experiments, four real-world datasets are used for evaluation. The first dataset is WebKB4, which is a subset of WebKB.[1] Four frequently-used classes (i.e., student, project, faculty and course) are selected, and data from four universities (i.e., Cornell, Texas, Washington and Wisconsin) forms four tasks. The second one is the HumanEva 3-D Motion Data [40]. Data of 3-D joints of two subjects are included. Each subject has five classes and corresponds to a clustering task. The other two cross-domain datasets, i.e., Comp.vs.Sci and Rec.vs.Talk [24], [25], are generated from 20Newsgroup.[2] Table I summarizes the detailed statistics of the above four datasets.

We refer training clustering to the process of grouping training data into several clusters, while testing clustering is the process of predicting cluster labels for out-of-sample test data. In training clustering, we perform all comparison algorithms to cluster training data and then compare their clustering performance. For WebKB4, Comp.vs.Sci and Rec.vs.Talk, we randomly divide each task into a training part (50%) and a test part (50%), while for HumanEva 3-D Motion, 1000 data points were selected from each task for training and the remaining data were used for test. The relaxed solution needs to be discretized. Following [14], in DKM we utilized TKM for discretization. For all the comparison clustering algorithms in the family of spectral clustering, i.e., NCut, CLGR, SEC, LNKMTC, LSKMTC, and MTSC, we consistently employ spectral relaxation and spectral rotation to

[1] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

[2] http://people.csail.mit.edu/jrennie/20Newsgroups/

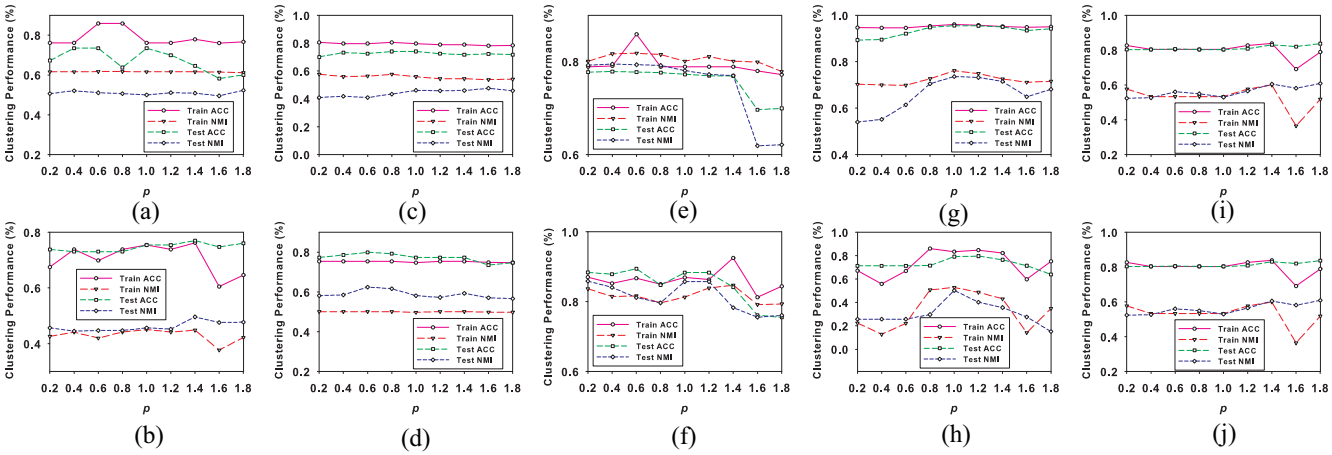


Fig. 1. Effects of p on WebKB4, HumanEva and Comp.vs.Sci datasets. p is set in the range of $\{0.2, 0.4, \ldots, 1.8\}$. (a) WebKB4 task 1. (b) WebKB4 task 2. (c) WebKB4 task 3. (d) WebKB4 task 4. (e) HumanEva task 1. (f) HumanEva task 2. (g) Comp.vs.Sci task 1. (h) Comp.vs.Sci task 2. (i) Rec.vs.talk task 2. (j) Rec.vs.talk task 2.

calculate the discrete cluster indicator matrix. We use Gaussian kernel for KKM clustering. When evaluating testing clustering, for TKM, DKM, FCM, and LSSMTC, a conventional way to predict a test data point's cluster label is to find the closest group center learnt from the training data. For SEC and MTSC, we can directly use the learnt mapping function to compute the cluster label indicator. As to all the other spectral clustering algorithms (i.e., NCut, CLGR), they are not able to handle out-of-sample data in the first place. Nonetheless, we have shown that these methods have strong connections with our model in certain specific circumstance. As we can see, when we set $\alpha \to 0$ and $\beta \to 0$, and use the specific graph Laplacian in each spectral clustering algorithm, we may easily get their extension to handle out-of-sample data. Note that we did not evaluate the testing performance for LNKMTC and LSKMTC since they do not have extensions for handling out-of-sample data.

For spectral clustering algorithms which need to specify the number of neighbors, we always set it to $k = 5$. We perform the self-tuning algorithm [41] to determine an adaptive bandwidth. For fair comparison, the parameters in all the comparison algorithms are consistently tuned from the range of $\{10^{-6}, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}, 10^{6}\}$ and the best results are reported. To reduce statistical variation, each clustering algorithm is run repeatedly for 20 times and the average results are reported.

B. Evaluation Metrics

By following the convention of clustering study, we use accuracy (ACC) and normalized mutual information (NMI) as our evaluation metrics in the following experiments.

Denote q_i as the clustering label result from a clustering algorithm and p_i as the corresponding ground truth label of any data sample x_i, then ACC is defined as follows:

$$\text{ACC} = \frac{\sum_{i=1}^{n} \delta(p_i, map(q_i))}{n} \tag{18}$$

where $\delta(x, y) = 1$ if $x = y$; $\delta(x, y) = 0$ otherwise, and $map(q_i)$ is the best mapping function that permutes clustering labels to match the ground truth labels using the Kuhn–Munkres algorithms. A larger ACC indicates a better clustering performance.

For any two arbitrary variable P and Q, NMI is defined as follows [42]:

$$\text{NMI} = \frac{I(P, Q)}{\sqrt{H(P)H(Q)}} \tag{19}$$

where $I(P, Q)$ computes the mutual information between P and Q, and $H(P)$ and $H(Q)$ are the entropies of P and Q. Denote t_l as the number of data in the cluster $\mathcal{C}_l (1 \le l \le c)$ generated by a clustering algorithm and $\tilde{t}_h (1 \le l \le c)$ as the number of data points from the hth ground truth class. NMI metric is then defined as below [42]

$$\text{NMI} = \frac{\sum_{l=1}^{c}\sum_{h=1}^{c} t_{l,h} \log\left(\frac{n \times t_{l,h}}{t_l \tilde{t}_h}\right)}{\sqrt{\left(\sum_{l=1}^{c} t_l \log\frac{t_l}{n}\right)\left(\sum_{h=1}^{c} \tilde{t}_h \log\frac{\tilde{t}_h}{n}\right)}} \tag{20}$$

where $t_{l,h}$ is the number of data samples that lie in the intersection between $\mathcal{C}_l$ and hth ground truth class. Likewise, a larger NMI indicates a better clustering performance.

C. Parameter Analysis

1) Effects of p: In this part, we evaluate the effect of the parameter p, which controls of the coherence among clustering tasks. As aforementioned, when p is larger (i.e., $p \to 2$), the $\ell_{2,p}$-norm tends to be closer to ℓ_2-norm, which implies that the correlation among all the tasks becomes weaker. On the other hand, when p gets smaller, all the tasks are more correlated. For each task in a dataset, p is chosen in $\{0.2, 0.4, \ldots, 1.8\}$. α and β are chosen according to the previous grid search results. As shown, Fig. 1(a)–(d) illustrates the results of four tasks of dataset "WebKB4," while Fig. 1(e) and (f) shows the results of two tasks of the "HumanEva" dataset. The results of dataset "Comp.vs.Sci" are listed in Fig. 1(g) and (h). Fig. 1(i) and (j) shows the effects of p for "Rec.vs.Talk" dataset. As we see, p is difficult to determine since it is data-dependent. We consistently fix it as $p = 1$.

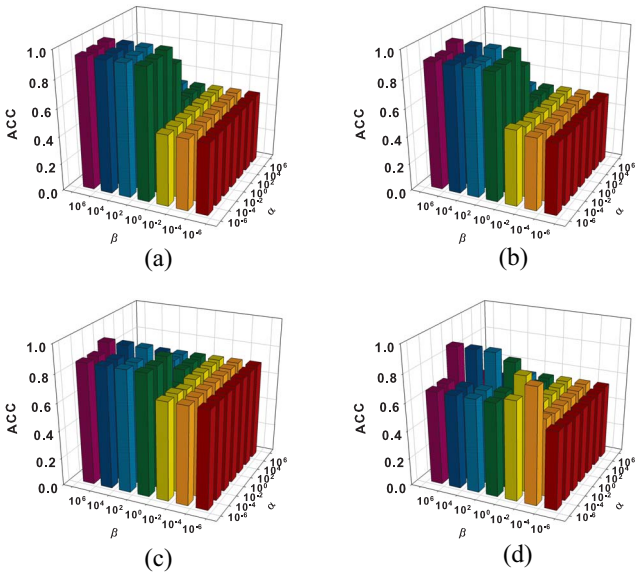


Fig. 2. Sensitivity study of α and β on dataset "Comp.vs.Sci." (a) Training ACC of task 1. (b) Testing ACC of task 1. (c) Training ACC of task 2. (d) Testing ACC of task 2.

As we can see, when p is small (from 0.2 to 1.0), all the four performance either keep increasing [see Fig. 1(g) and (h)] or keep steady [Fig. 1(c) and (d)]. This phenomenon shows that even though we force different tasks to be highly correlated, the clustering performance cannot be expectedly improved. We conclude that too small p cannot necessarily help our model to reveal an effective low-dimensional space for different tasks to be genuinely connected. A possible explanation is that if we force different tasks to achieve high similarity or correlation, then unique yet valuable properties embedded within each individual task will be inevitably lost, which will lead to the performance degrade. On the other hand, when p becomes larger (from 1.0 to 1.8), the performance either exhibits decreasing trends [Fig. 1(e) and (f)] or steady trends [Fig. 1(c) and (d)]. This observation demonstrate that if less correlation is uncovered, the performance gain becomes naturally less.

2) Effects of α and β: In this part, we evaluate the parameter sensitivity of our proposed model. We choose to use dataset "Comp.vs.Sci" and the performance metric ACC is used for demonstration. Both α and β are chosen in the range of $\{10^{-6}, 10^{-4}, \ldots, 10^{4}, 10^{6}\}$. p is fixed as 1. The results of clustering training and testing data are illustrated in Fig. 2(a)–(d).

As can be seen, in all the four figures, in most cases, the better performance of our model for clustering both training and testing data is achieved within a focused region, i.e., large β and small α; besides, our performance is not sensitive to α and β in other settings. For both task 1 and task 2, when β is large, we get relatively better performance. This implies that the effects of the $\ell_{2,p}$-norm regularization term indeed helps explore the intertask coherence across tasks to boost clustering performance. As to the effect of α, for example in Fig. 2(c) and (d), either too large or too small α helps achieves better results. This indicates that a proper α should be chosen to get balance between learning the mapping function and learning cluster labels.

D. Comparisons

In this subsection, we first compare the proposed MTSC model to several baselines and state-of-the-art multitask clustering algorithms in terms of training performance and testing performance. We fix $p = 1$ for our proposed MTSC model. Tables II–V demonstrate all the comparison results, from which we have the following observations and analysis.

1) DKM performs better than TKM in almost all the cases since it is a joint framework incorporating discriminative dimension reduction and clustering together [14], which might make each cluster more identifiable and thus improve the clustering performance. Such observation implies it is beneficial to take discriminative information into clustering.
2) SEC is able to achieve comparable results compared to other single-task methods. For DKM, TKM and NCut, SEC performs better in most cases. We believe that it owes to the integration of discriminative information. CLGR, which is one of the state-of-the-art spectral clustering, surpasses SEC. This is mainly because CLGR possesses more tunable parameters than our method, which in turn makes it more sensitive.
3) MTSC consistently outperforms SEC over all the datasets and tasks. This outcome illustrates that the intertask relationships are well captured and indeed help to improve the clustering performance.
4) All the multitask clustering algorithms are able to achieve comparable performance, which is consistent with the existing outcomes [25]. Especially, for both Rec.vs.Talk and Comp.vs.Sci, LSSMTC, LSKMTC and LNKMTC achieve similar or even better results as in their original work [25].
5) MTSC always achieves the best results among all the comparison clustering algorithms, including both single-task and multitask methods. This observation clearly reveals that exploring the intertask correlations and integrating discriminative information play significant roles in boosting clustering performance.

We also illustrate the clustering results of our algorithm and one of the representative single-task algorithm, i.e., k-means in Fig. 3. As seen, the results of our method on both the tasks of Rec.vs.Talk dataset are closer to the original ground truth as compared to k-means.

E. Convergence Study

As analyzed before, Algorithm 1 is able to converge to an optimal solution. In this part, we conduct empirical study on the convergence, as demonstrated in Fig. 4. As we can see, our algorithm can quickly converge within less than 10 iterations, which clearly shows its efficiency.

F. Running Time

In this part, we compare the running time of our method to that of two representative algorithms, i.e., CLGR

TABLE II
TRAINING CLUSTERING PERFORMANCE COMPARISON (ACC % +/- STANDARD DEVIATION) ON FOUR DATASETS

Dataset	Task	TKM	DKM	KKM	FCM	NCut	CLGR	SEC	LSSMTC	LNKMTC	LSKMTC	MTSC
WebKB4	1	54.3±9.3	63.4±4.8	52.2±5.5	64.6±3.9	62.1±8.2	68.7±8.1	70.3±4.9	71.9±4.5	63.0±9.2	55.4±5.2	$\mathbf{77.3 \pm 5.9}$
	2	53.5±5.7	66.2±6.5	44.4±2.7	61.9±4.4	56.7±4.6	71.4±4.4	71.1±7.3	73.5±4.8	60.2±10.5	53.7±6.0	$\mathbf{77.3 \pm 1.6}$
	3	68.3±4.8	72.3±3.9	64.8±1.8	66.4±1.6	67.8±9.7	75.5±3.6	78.8±2.6	69.8±3.2	70.9±7.9	55.9±7.5	$\mathbf{82.7 \pm 1.5}$
	4	59.1±4.6	65.1±5.6	68.2±2.0	68.8±4.0	61.9±8.8	70.3±3.9	70.3±4.7	70.9±2.6	65.1±4.8	50.6±1.9	$\mathbf{74.7 \pm 1.8}$
HumanEva	1	66.8±1.0	70.1±1.5	76.2±3.1	67.3±2.1	63.5±2.7	68.6±2.3	65.5±0.7	60.3±1.1	70.1±6.2	76.6±4.1	$\mathbf{82.2 \pm 3.7}$
	2	73.6±2.2	73.6±2.1	72.8±2.0	76.6±1.7	69.7±5.0	81.8±5.9	80.0±4.0	64.4±3.6	74.4±0.6	90.3±4.5	$\mathbf{96.0 \pm 1.1}$
Comp vs. Sci	1	91.4±0.5	92.6±0.3	92.0±0.5	93.5±0.5	91.3±0.5	88.6±4.8	93.0±0.4	93.6±0.4	81.1±1.0	94.5±0.1	$\mathbf{95.6 \pm 0.5}$
	2	69.1±1.5	69.5±1.4	63.7±2.2	85.6±2.0	78.1±2.9	83.9±2.1	84.0±1.4	83.8±1.3	82.1±1.4	77.5±1.8	$\mathbf{90.0 \pm 1.3}$
Rec vs. Talk	1	70.3±1.4	72.9±1.0	83.4±1.3	81.1±8.4	87.2±5.8	91.3±0.1	87.4±5.9	92.1±0.8	75.4±3.7	92.0±0.3	$\mathbf{94.7 \pm 0.6}$
	2	76.1±0.3	78.4±4.1	83.3±2.3	89.7±0.9	85.3±0.8	87.5±0.8	86.9±0.3	86.2±0.9	83.7±5.3	70.1±8.3	$\mathbf{89.9 \pm 1.9}$

TABLE III
TRAINING CLUSTERING PERFORMANCE COMPARISON (NMI +/- STANDARD DEVIATION) ON FOUR DATASETS

Dataset	Task	TKM	DKM	KKM	FCM	NCut	CLGR	SEC	LSSMTC	LNKMTC	LSKMTC	MTSC
WebKB4	1	0.290±0.103	0.339±0.048	0.305±0.040	0.428±0.015	0.478±0.091	0.499±0.071	0.530±0.040	0.492±0.032	0.473±0.077	0.294±0.030	$\mathbf{0.596 \pm 0.034}$
	2	0.343±0.077	0.370±0.077	0.426±0.009	0.341±0.011	0.417±0.033	0.453±0.025	0.483±0.031	0.499±0.025	0.388±0.069	0.277±0.043	$\mathbf{0.523 \pm 0.018}$
	3	0.443±0.070	0.452±0.047	0.410±0.067	0.407±0.002	0.484±0.073	0.541±0.059	0.568±0.026	0.473±0.033	0.484±0.087	0.317±0.076	$\mathbf{0.612 \pm 0.026}$
	4	0.405±0.066	0.452±0.081	0.468±0.011	0.484±0.020	0.429±0.059	0.465±0.036	0.494±0.036	0.490±0.038	0.457±0.071	0.333±0.044	$\mathbf{0.540 \pm 0.038}$
HumanEva	1	0.586±0.004	0.629±0.006	0.627±0.013	0.563±0.013	0.628±0.017	0.664±0.020	0.641±0.006	0.475±0.014	0.674±0.005	0.718±0.010	$\mathbf{0.771 \pm 0.040}$
	2	0.609±0.019	0.670±0.040	0.638±0.017	0.623±0.015	0.780±0.017	0.792±0.028	0.773±0.015	0.551±0.032	0.637±0.001	0.855±0.031	$\mathbf{0.908 \pm 0.023}$
Comp vs. Sci	1	0.580±0.021	0.629±0.015	0.599±0.020	0.695±0.026	0.588±0.038	0.593±0.004	0.639±0.018	0.662±0.011	0.622±0.027	0.698±0.002	$\mathbf{0.744 \pm 0.023}$
	2	0.232±0.025	0.225±0.004	0.357±0.012	0.524±0.051	0.321±0.032	0.421±0.032	0.398±0.022	0.417±0.031	0.329±0.033	0.318±0.022	$\mathbf{0.557 \pm 0.046}$
Rec vs. Talk	1	0.237±0.010	0.254±0.012	0.451±0.012	0.429±0.175	0.566±0.005	0.586±0.005	0.589±0.002	0.626±0.025	0.632±0.051	0.621±0.008	$\mathbf{0.720 \pm 0.013}$
	2	0.309±0.009	0.358±0.073	0.362±0.012	0.518±0.034	0.465±0.032	0.512±0.004	0.492±0.022	0.505±0.013	0.432±0.094	0.398±0.014	$\mathbf{0.624 \pm 0.002}$

TABLE IV
TESTING CLUSTERING PERFORMANCE COMPARISON (ACC % +/- STANDARD DEVIATION) ON FOUR DATASETS

Dataset	Task	TKM	DKM	KKM	FCM	NCut	CLGR	SEC	LSSMTC	MTSC
WebKB4	1	60.7±12.6	67.1±4.5	57.3±2.2	69.0±8.5	65.5±7.0	69.6±6.5	66.7±3.5	71.7±1.7	$\mathbf{73.5 \pm 2.3}$
	2	64.1±10.5	70.6±6.8	56.8±2.0	68.3±4.5	62.1±8.4	75.5±1.6	63.5±2.9	71.1±1.8	$\mathbf{75.6 \pm 3.5}$
	3	68.7±8.6	74.8±2.3	66.1±5.6	67.7±2.7	70.7±4.5	75.3±2.9	75.4±3.1	71.0±4.0	$\mathbf{77.2 \pm 2.9}$
	4	65.1±7.5	72.5±6.6	68.4±4.8	68.1±5.5	72.1±4.0	74.9±2.0	74.0±4.2	71.8±2.4	$\mathbf{75.3 \pm 1.7}$
HumanEva	1	65.5±0.4	68.5±2.0	74.8±1.5	67.3±1.1	66.5±1.9	71.2±0.8	63.0±0.8	59.2±2.3	$\mathbf{77.1 \pm 2.5}$
	2	70.5±0.3	71.5±0.2	73.5±0.4	75.7±2.5	68.8±3.9	84.0±2.5	71.1±1.3	64.4±4.9	$\mathbf{93.7 \pm 1.4}$
Comp vs. Sci	1	91.0±0.3	91.5±0.2	89.9±0.5	94.1±0.9	92.7±2.0	88.6±7.4	91.6±0.5	93.9±0.5	$\mathbf{95.1 \pm 0.4}$
	2	67.3±1.6	67.8±1.0	65.0±5.2	85.7±2.0	79.9±0.8	86.0±4.0	77.7±1.1	83.1±0.6	$\mathbf{90.0 \pm 0.1}$
Rec vs. Talk	1	69.4±1.0	71.8±1.3	75.2±6.8	82.3±5.7	93.3±0.2	94.1±0.2	91.7±0.1	91.3±1.0	$\mathbf{94.3 \pm 0.4}$
	2	75.7±2.5	76.2±0.8	81.4±0.9	83.4±1.0	85.7±0.8	87.8±1.6	88.0±0.6	85.9±0.6	$\mathbf{88.4 \pm 3.7}$

TABLE V
TESTING CLUSTERING PERFORMANCE COMPARISON (NMI +/- STANDARD DEVIATION) ON FOUR DATASETS

Dataset	Task	TKM	DKM	KKM	FCM	NCut	CLGR	SEC	LSSMTC	MTSC
WebKB4	1	0.403±0.112	0.477±0.063	0.376±0.033	0.376±0.050	0.500±0.052	0.557±0.067	0.507±0.045	0.520±0.038	$\mathbf{0.564 \pm 0.026}$
	2	0.392±0.090	0.415±0.044	0.421±0.011	0.383±0.044	0.389±0.057	0.440±0.011	0.408±0.042	0.413±0.015	$\mathbf{0.466 \pm 0.026}$
	3	0.442±0.051	0.462±0.072	0.377±0.062	0.330±0.026	0.484±0.029	0.495±0.033	0.500±0.026	0.479±0.050	$\mathbf{0.537 \pm 0.017}$
	4	0.483±0.050	0.502±0.037	0.474±0.011	0.493±0.086	0.529±0.054	0.555±0.052	0.560±0.050	0.503±0.034	$\mathbf{0.572 \pm 0.018}$
HumanEva	1	0.572±0.001	0.591±0.001	0.616±0.014	0.531±0.004	0.618±0.011	0.631±0.010	0.496±0.011	0.459±0.021	$\mathbf{0.713 \pm 0.046}$
	2	0.581±0.003	0.622±0.027	0.644±0.027	0.608±0.031	0.586±0.039	0.785±0.012	0.647±0.006	0.535±0.032	$\mathbf{0.879 \pm 0.020}$
Comp vs. Sci	1	0.563±0.009	0.582±0.009	0.529±0.015	0.710±0.037	0.630±0.072	0.660±0.054	0.585±0.018	0.672±0.015	$\mathbf{0.719 \pm 0.019}$
	2	0.193±0.017	0.231±0.038	0.397±0.053	0.482±0.048	0.339±0.019	0.452±0.099	0.249±0.012	0.428±0.011	$\mathbf{0.509 \pm 0.006}$
Rec vs. Talk	1	0.239±0.018	0.255±0.025	0.441±0.041	0.451±0.132	0.649±0.007	0.678±0.006	0.595±0.005	0.601±0.015	$\mathbf{0.687 \pm 0.013}$
	2	0.331±0.041	0.341±0.015	0.343±0.005	0.547±0.046	0.419±0.015	0.530±0.077	0.492±0.027	0.508±0.012	$\mathbf{0.562 \pm 0.049}$

and LSSMTC, which achieve relatively better clustering performance among all the comparing algorithms. We set the maximum number of iterations in our method to 10 since our method normally converges very quickly as discussed in Section V-E. As shown in Table VI, the proposed MTSC is obviously more efficient than LSSMTC, which further indicates that our method is not only an effective but also efficient multitask clustering algorithm. Note that our method costs more time than CLGR due to the exploration of intertask correlation. We argue that such efficiency sacrifice is still worthwhile because more performance gain can be obtained.

TABLE VI
RUNNING TIME OF CLGR, LSSMTC, AND MTSC ON WEBKB4 DATASET

Method	CLGR	LSSMTC	MSTC
Time (s)	0.9467	14.1449	7.1220

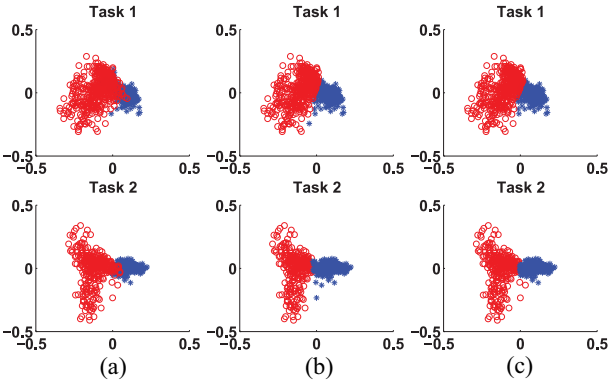


Fig. 3. Illustration of clustering results for our approach and k-means on Rec.vs.Talk dataset. (a) Original ground truth. (b) TKM. (c) MTSC.

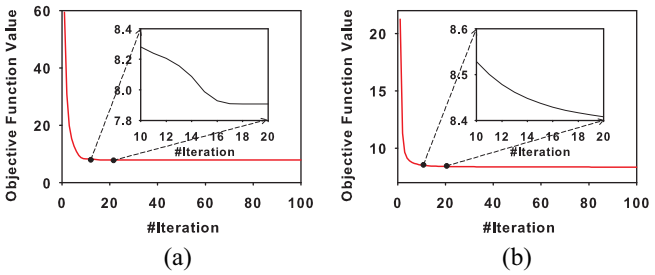


Fig. 4. Convergence study on HumanEva dataset. We set $\alpha = 1$ and $\beta = 1$. Two settings are chosen for p. (a) $p = 0.5$. (b) $p = 1$.

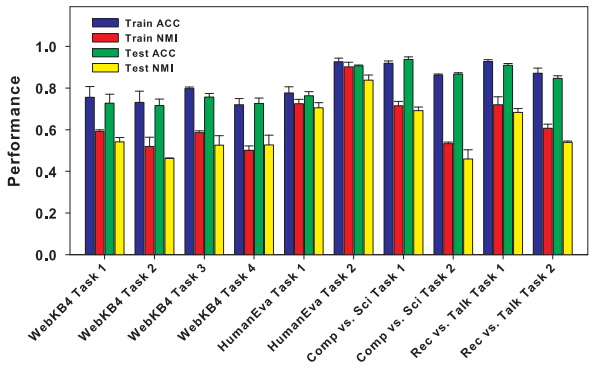


Fig. 5. Effects of the initialization on the performance of our approach. We repeated the random initialization for 20 times, and the average results and standard deviation are reported.

G. Stability of Initialization

In this part, we further evaluate the effects of the initialization to the proposed algorithm. For each dataset, we choose to use the parameters that leads to the best results in our approach, and we repeated the random initialization for 20 times for each setting, and the average results and standard deviation are reported as in Fig. 5. We can see that the results is not very sensitive to the initialization, which indicates the stability of our approach.

VI. Conclusion

In this paper, we proposed a novel clustering model, namely MTSC, to cope with the emerging challenges faced by traditional clustering approaches. We proposed to incorporate a novel $\ell_{2,p}$-norm regularizer to control the coherence of all the tasks based on a wild assumption that related tasks should share a common low-dimensional representation. Moreover, for each individual task, an explicit mapping function was learnt by mapping features to the cluster label matrix. We discussed the connections between our proposed model and several representative clustering techniques, including spectral clustering, k-means and DKM. Extensive experiments on various real-world datasets illustrated the advantage of the proposed MTSC model over the state-of-the-art clustering approaches. In the future, we intend to explore more properties of the cluster label matrix, i.e., non-negative nature, to enhance the performance of the current proposal.

Appendix A
Proof of Lemma 1

Proof: We first consider a function

$$f(\sigma) = p\sigma^2 - 2\sigma^p + (2-p) \tag{21}$$

where $p \in (0, 2)$. We expect to show that when $\sigma > 0$, $f(\sigma) \geq 0$. The first and second order derivatives of the function in (21) are $f'(\sigma) = 2p\sigma - 2p\sigma^{p-1}$ and $f''(\sigma) = 2p - 2p(p-1)\sigma^{p-2}$, respectively. We can see that $\sigma = 1$ is the only point that satisfies $f'(\sigma) = 0$. Also, when $0 < \sigma < 1, f'(\sigma) < 0$ and when $\sigma > 1, f'(\sigma) > 0$. This means that $f(\sigma)$ is monotonically decreasing when $0 < \sigma < 1$ and monotonically increasing when $\sigma > 1$. Moreover, we have $f''(1) = 2p(2-p) > 0$. Therefore, for $\forall \sigma > 0, f(\sigma) \geq f(1) = 0$.

Then, by substituting $\sigma = \frac{\|\tilde{W}_j\|_2}{\|W_j\|_2}$ into (21), we obtain the conclusion

$$\begin{aligned}
&p\frac{\|\tilde{W}_j\|_2^2}{\|W_j\|_2^2} - 2\frac{\|\tilde{W}_j\|_2^p}{\|W_j\|_2^p} + (2-p) \geq 0 \\
&\Leftrightarrow p\|\tilde{W}_j\|_2^2 - 2\|\tilde{W}_j\|_2^p\|W_j\|_2^{2-p} + (2-p)\|W_j\|_2^2 \geq 0 \\
&\Leftrightarrow p\|\tilde{W}_j\|_2^2\|W_j\|_2^{p-2} - 2\|\tilde{W}_j\|_2^p + (2-p)\|W_j\|_2^p \geq 0 \\
&\Leftrightarrow 2\|\tilde{W}_j\|_2^p - p\|\tilde{W}_j\|_2^2\|W_j\|_2^{p-2} \leq (2-p)\|W_j\|_2^p \\
&\Leftrightarrow \|\tilde{W}_j\|_2^p - \frac{p\|\tilde{W}_j\|_2^2}{2\|W_j\|_2^{2-p}} \leq \|W_j\|_2^p - \frac{p\|W_j\|_2^2}{2\|W_j\|_2^{2-p}}.
\end{aligned}$$

■

Appendix B
Proof of Theorem 1

Proof: For brevity, we denote $\mathcal{R}(\mathcal{F}, W)$ as the loss function corresponding to the first two terms of the objective function in (8) and $\mathcal{F} = \{F^{(t)}\}_{t=1}^M$. Suppose $\{\tilde{W}, \tilde{\mathcal{F}}\}$ is the solution to the optimization problem in (9), then we have

$$\begin{aligned}
&\mathcal{R}(\tilde{\mathcal{F}}, \tilde{W}) + \beta Tr(\tilde{W}^T U \tilde{W}) \leq \mathcal{R}(\mathcal{F}, W) + \beta Tr(W^T D W) \\
&\Rightarrow \mathcal{R}(\tilde{\mathcal{F}}, \tilde{W}) + \beta \sum_{j=1}^d \frac{p\|\tilde{W}_j\|_2^2}{2\|W_j\|_2^{2-p}} \leq \mathcal{R}(\mathcal{F}, W) + \beta \sum_{j=1}^d \frac{p\|W_j\|_2^2}{2\|W_j\|_2^{2-p}} \\
&\Rightarrow \mathcal{R}(\tilde{\mathcal{F}}, \tilde{W}) + \beta \sum_{j=1}^d \|\tilde{W}_j\|_2^2 - \beta\left(\sum_{j=1}^d \|\tilde{W}_j\|_p^2 - \sum_{j=1}^d \frac{p\|\tilde{W}_j\|_2^2}{2\|W_j\|_2^{2-p}}\right) \\
&\leq \mathcal{R}(\mathcal{F}, W) + \beta \sum_{j=1}^d \|W_j\|_p^2 - \beta\left(\sum_{j=1}^d \|W_j\|_p^2 - \sum_{j=1}^d \frac{p\|W_j\|_2^2}{2\|W_j\|_2^{2-p}}\right).
\end{aligned}$$

According to Lemma 2, we obtain

$$\mathcal{R}(\tilde{\mathcal{F}}, \tilde{W}) + \beta \sum_{j=1}^{d} \left\| \tilde{W}_j \right\|_2^p \leq \mathcal{R}(\mathcal{F}, W) + \beta \sum_{j=1}^{d} \left\| W_j \right\|_2^p.$$

Thus, we have proved that at each iteration, the value of the objective function in (8) monotonically decreases. ■

APPENDIX C
PROOF OF PROPOSITION 1

Proof: As shown in (8), when $\alpha = 0$ and $\beta = 0$, the formulation is reduced to the following form:

$$\begin{aligned} &\min_{\{F^{(t)}\}_{t=1}^{M}} \sum_{t=1}^{M} \left(tr\left(\left(F^{(t)} \right)^T L^{(t)} F^{(t)} \right) \right) \\ &\text{s.t. } (F^{(t)})^T F^{(t)} = I_t, \ t = 1, 2, \ldots, M. \end{aligned} \tag{22}$$

The above formulation is apparently separable, thus it equals to performing spectral clustering on each task individually. ■

APPENDIX D
PROOF OF PROPOSITION 2

Proof: When $p = 2$, (8) can be written as

$$\begin{aligned} &\min_{\mathcal{F}, W} \ \mathcal{R}(\mathcal{F}, W) + \beta \|W\|_F^2 \\ &\text{s.t. } (F^{(t)})^T F^{(t)} = I_t, \ t = 1, 2, \ldots, M \end{aligned} \tag{23}$$

where $\mathcal{R}(\mathcal{F}, W)$ is the loss function corresponding to the first two terms of the objective function in (8) and $\mathcal{F} = \{F^{(t)}\}_{t=1}^{M}$. Since (23) is separable, we only focus on one task t, which leads to

$$\begin{aligned} &\min_{F, W} \ tr(F^T L F) + \alpha \left\| F - X^T W \right\|_F^2 + \beta \|W\|_F^2 \\ &\text{s.t. } F^T F = I. \end{aligned} \tag{24}$$

Note that we omit all the subscripts and superscripts for brevity. By setting the derivative of (24) w.r.t. W to 0, we obtain

$$W = \alpha \left(\alpha X X^T + \beta I \right)^{-1} X F. \tag{25}$$

Then, by substituting (25) into (24) we arrive at

$$\min_{F, F^T F = I} \ tr\left(F^T \left(L - \alpha X^T \left(\alpha X X^T + \beta I \right)^{-1} X \right) F \right). \tag{26}$$

As $\alpha \to \infty$, (26) is reduce to

$$\max_{F, F^T F = I} \ tr\left(F^T X^T \left(\alpha X X^T + I \right)^{-1} X F \right). \tag{27}$$

Given that $\frac{\alpha}{\beta} \to 0$, the problem in (27) is further reduced to

$$\max_{F, F^T F = I} \ tr\left(F^T X^T X F \right) \tag{28}$$

which is exactly the same as the k-means algorithm shown in [43]. Therefore, when $\frac{\alpha}{\beta} \to 0$, $\alpha \to \infty$ and $p = 2$, The MTSC model leads to k-means clustering. ■

APPENDIX E
PROOF OF PROPOSITION 3

Proof: Recall that the proposed model is equivalent to (27) when $\beta \to \infty$ and α is set to be a constant. Further, we have the following conclusion:

$$\begin{aligned} &\max_{F, F^T F = I} \ tr\left(F^T \left(I - \left(I + \frac{1}{\gamma} X^T X \right)^{-1} \right) F \right) \\ &\Leftrightarrow \max_{F, F^T F = I} \ tr\left(F^T \left(\left(I + \frac{1}{\gamma} X^T X \right)^{-1} \left(I + \frac{1}{\gamma} X^T X - I \right) \right) F \right) \\ &\Leftrightarrow \max_{F, F^T F = I} \ tr\left(F^T X^T \left(I + \frac{1}{\gamma} X X^T \right)^{-1} X F \right) \end{aligned} \tag{29}$$

which is exactly the same as the problem in (27) when $\gamma = 1/\alpha$. ■

REFERENCES

[1] A. Jain, M. Murty, and P. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[2] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[3] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.

[4] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1370–1386, Nov. 2004.

[5] K. Hammouda and M. Kamel, "Efficient phrase-based document indexing for web document clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 10, pp. 1279–1296, Nov. 2004.

[6] S. Gordon, H. Greenspan, and J. Goldberger, "Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations," in *Proc. 9th IEEE Int. Conf. Comput. Vis. (CVPR)*, Nice, France, Oct. 2003, pp. 370–377.

[7] J. Jia, N. Yu, and X. Hua, "Annotating personal albums via web mining," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 459–468.

[8] J. Li and J. Wang, "Real-time computerized annotation of pictures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 985–1002, Jun. 2008.

[9] X. Wang, L. Zhang, X. Li, and W. Ma, "Annotating images by mining image search results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1919–1932, Nov. 2008.

[10] C. Li, E. Chang, H. Garcia-Molina, and G. Wiederhold, "Clustering for approximate similarity search in high-dimensional spaces," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 4, pp. 792–808, Jul./Aug. 2002.

[11] C. Ding and T. Li, "Adaptive dimension reduction using discriminant analysis and k-means clustering," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, Corvallis, OR, USA, 2007, pp. 521–528.

[12] F. De la Torre and T. Kanade, "Discriminative cluster analysis," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, Pittsburgh, PA, USA, 2006, pp. 241–248.

[13] J. Ye, Z. Zhao, and H. Liu, "Adaptive distance metric learning for clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, Jun. 2007, pp. 1–7.

[14] J. Ye, Z. Zhao, and M. Wu, "Discriminative k-means for clustering," in *Proc. Conf. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 20. 2007, pp. 1649–1656.

[15] C. Liu, W. Hsaio, C. Lee, and F. Gou, "Semi-supervised linear discriminant clustering," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 989–1000, Jul. 2014.

[16] S. Yu and J. Shi, "Multiclass spectral clustering," in *Proc. 9th IEEE Int. Conf. Comput. Vis. (ICCV)*, Nice, France, Oct. 2003, pp. 313–319.

[17] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," *Pattern Recognit.*, vol. 41, no. 1, pp. 176–190, Jan. 2008.

[18] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Jan. 2006.

[19] D. Meng, Y. Leung, and Z. Xu, "Detecting intrinsic loops underlying data manifold," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 2, pp. 337–347, Feb. 2013.

[20] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural comput.*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003.

[21] F. Nie, D. Xu, I. Tsang, and C. Zhang, "Spectral embedded clustering," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Pasadena, CA, USA, 2009, pp. 1181–1186.

[22] M. Wu and B. Scholkopf, "A local learning approach for clustering," in *Advances in Neural Information Processing Systems 19*, B. Scholkopf, J. C. Platt, and T. Hoffman, Eds., vol. 19. Cambridge, MA, USA: MIT Press, Dec. 2007 pp. 1529–1536.

[23] J. Zhou, J. Chen, and J. Ye, "Clustered multi-task learning via alternating structure optimization," in *Proc. 24th Conf. Adv. Neural Inf. Process. Syst. (NIPS)*, 2011, pp. 702–710.

[24] Q. Gu and J. Zhou, "Learning the shared subspace for multi-task clustering and transductive transfer classification," in *Proc. 9th IEEE Int. Conf. Data Min. (ICDM)*, Miami, FL, USA, 2009, pp. 159–168.

[25] Q. Gu, Z. Li, and J. Han, "Learning a kernel for multi-task clustering," in *Proc. 25th AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, Aug. 2011.

[26] Z. Zhang and J. Zhou, "Multi-task clustering via domain adaptation," *Pattern Recognit.*, vol. 45, no. 1, pp. 465–473, 2012.

[27] S. Xie, H. Lu, and Y. He, "Multi-task co-clustering via nonnegative matrix factorization," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Tsukuba, Japan, Nov. 2012, pp. 2954–2958.

[28] S. Kong and D. Wang, "A multi-task learning strategy for unsupervised clustering via explicitly separating the commonality," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Tsukuba, Japan, Nov. 2012, pp. 771–774.

[29] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Self-taught clustering," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, Helsinki, Finland, 2008, pp. 200–207.

[30] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nystrom method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 214–225, Feb. 2004.

[31] Y. Bengio *et al.*, "Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L.K. Saul, and B. Scholkopf, vol. 16. Cambridge, MA, USA: MIT Press, Dec. 2004, pp. 177–184.

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

[33] L. Saul and S. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *J. Mach. Learn. Res.*, vol. 4, pp. 119–155, Jan. 2003.

[34] Y. Yang *et al.*, "Discriminative nonnegative spectral clustering with out-of-sample extension," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1760–1771, Aug. 2013.

[35] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.

[36] Z. Ma, Y. Yang, N. Sebe, and A. Hauptmann, "Knowledge adaptation with partially shared features for event detection using few exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 9, pp. 1789–1802, Sep. 2014.

[37] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: Spectral clustering and normalized cuts," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Seattle, WA, USA, 2004, pp. 551–556.

[38] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic, 1981.

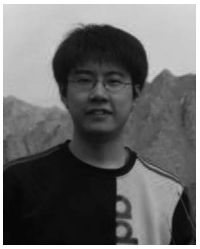
[39] F. Wang, C. Zhang, and T. Li, "Clustering with local and global regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 12, pp. 1665–1678, Dec. 2009.

[40] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vis.*, vol. 87, no. 1, pp. 4–27, 2010.

[41] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17. 2004, pp. 1601–1608.

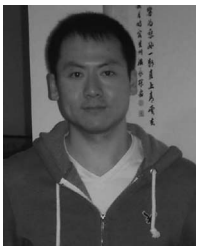
[42] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 583–617, Apr. 2003.

[43] H. Zha, C. Ding, M. Gu, X. He, and H. Simon, "Spectral relaxation for k-means clustering," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 14. 2001, pp. 1057–1064.



Yang Yang received the B.S. degree from Jilin University, Changchun, China, in 2006, the M.E. degree from Peking University, Beijing, China, in 2009, and the Ph.D. degree from the University of Queensland, Brisbane, QLD, Australia, in 2013.

He was a Research Fellow with the School of Computing, National University of Singapore, Singapore. His current research interests include multimedia information retrieval, social media analysis, and machine learning.



Zhigang Ma received the Ph.D. degree in computer science from the University of Trento, Trento, Italy, in 2013.

He is currently a Post-Doctoral Research Fellow with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. His current research interest include mainly on machine learning and its applications to multimedia analysis and computer vision.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010.

He is currently a DECRA Fellow with the University of Queensland, Brisbane, QLD, Australia. Prior to that, he was a Post-Doctoral Research Fellow with the school of computer science, Carnegie Mellon University, Pittsburgh, PA, USA. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, multimedia indexing and retrieval, surveillance video analysis, and video semantics understanding.

Feiping Nie received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2009.

He is currently a Research Assistant Professor with the University of Texas at Arlington, Arlington, TX, USA. His current research interests include machine learning and its application fields, such as pattern recognition, data mining, computer vision, image processing, and information retrieval. He has published over 100 papers in the following journals and conferences: TPAMI, IJCV, TIP, TNNLS/TNN, TKDE, TKDD, TVCG, TCSVT, TMM, TSMCB/TC, *Machine Learning*, *Pattern Recognition*, *Medical Image Analysis*, *Bioinformatics*, ICML, NIPS, KDD, IJCAI, AAAI, ICCV, CVPR, SIGIR, ACM MM, ICDE, ECML/PKDD, ICDM, MICCAI, IPMI, and RECOMB. According to the Google scholar, his papers have been cited over 2000 times.

Dr. Nie is currently serving as an Associate Editor or PC Member for several prestigious journals and conferences in the related fields.

Heng Tao Shen received the B.Sc. (1st class Hons.) and Ph.D. degrees from the Department of Computer Science, National University of Singapore, Singapore, in 2000 and 2004, respectively.

He is a Professor of Computer Science with the School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, QLD, Australia. He then joined the University of Queensland as a Lecturer and became a Professor in 2011. His current research interests include multimedia/mobile/web search and big data management.

Prof. Shen is the winner of Chris Wallace Award for Outstanding Research Contribution in 2010 from CORE Australasia. He is an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, and will serve as a PC Co-Chair for ACM Multimedia 2015.