Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Twin Learning for Similarity and Clustering: A Unified Kernel Approach

Zhao Kang, Chong Peng, Qiang Cheng
Department of Computer Science, Southern Illinois University, Carbondale, IL 62901, USA
{Zhao.Kang, pchong, qcheng}@siu.edu

Abstract

Many similarity-based clustering methods work in two separate steps including similarity matrix computation and subsequent spectral clustering. However, similarity measurement is challenging because it is usually impacted by many factors, e.g., the choice of similarity metric, neighborhood size, scale of data, noise and outliers. Thus the learned similarity matrix is often not suitable, let alone optimal, for the subsequent clustering. In addition, nonlinear similarity often exists in many real world data which, however, has not been effectively considered by most existing methods. To tackle these two challenges, we propose a model to simultaneously learn cluster indicator matrix and similarity information in kernel spaces in a principled way. We show theoretical relationships to kernel k-means, k-means, and spectral clustering methods. Then, to address the practical issue of how to select the most suitable kernel for a particular clustering task, we further extend our model with a multiple kernel learning ability. With this joint model, we can automatically accomplish three subtasks of finding the best cluster indicator matrix, the most accurate similarity relations and the optimal combination of multiple kernels. By leveraging the interactions between these three subtasks in a joint framework, each subtask can be iteratively boosted by using the results of the others towards an overall optimal solution. Extensive experiments are performed to demonstrate the effectiveness of our method.

Introduction

Clustering is a fundamental topic in data mining and machine learning (Peng et al. 2016). It partitions data points into different groups, such that the objects within a group are similar to one another and different from those in other groups. Various methods have been proposed over the past decades. Some well-known algorithms include k-means clustering (MacQueen 1967), spectral clustering (Ng et al. 2002), and hierarchical clustering (Johnson 1967).

Thanks to the simplicity and the effectiveness, the k-means algorithm is widely used. However, it fails to identify arbitrarily shaped clusters. Kernel k-means (Schölkopf, Smola, and Müller 1998) has been developed to capture nonlinear structure information hidden in data sets. Kernel-based learning methods requires one to specify a kernel, which means one assumes a certain shape of the underlying data space. Thus the performance of kernel-based methods are largely affected by the choice of kernel.

Spectral clustering does a low-dimensional embedding of the similarity matrix of the data before performing k-means clustering (Ng et al. 2002). The similarity between every pair of points, as an input, leverages the manifold information in this clustering model. Thus similarity-based clustering methods usually show better performance than k-means algorithm. However, the performance of this kind of methods is largely determined by the similarity matrix (Huang, Nie, and Huang 2015). Any variations during the similarity measurement, such as metric, neighborhood size, and data scale, may lead to suboptimal performance.

Recently, self-expression has been successfully utilized in subspace recovery (Elhamifar and Vidal 2009; Luo et al. 2011), low rank representation (Kang, Peng, and Cheng 2015b; 2015a), and recommender systems (Kang and Cheng 2016). It represents each data point in terms of the other points. By solving an optimization problem, the similarity information is automatically learned from the data. This approach can not only reveal low-dimensional structure, but also be robust to noise and data scale (Huang, Nie, and Huang 2015).

Contributions

In this paper, we perform clustering built upon the idea of using samples from the data to "express itself". Rather than local structure learning (Nie, Wang, and Huang 2014), this approach extracts the global structure of data and can be extended to kernel spaces. Unlike existing clustering algorithms that work in two separate steps, we simultaneously learn similarity matrix and cluster indicators by imposing a rank constraint on the Laplacian matrix of the learned similarity matrix. By leveraging the intrinsic interactions between learning similarity and cluster indicators, our proposed model seamlessly integrates them into a joint framework, where the result of one task is used to improve the other one. To capture the nonlinear structure information inherent in many real world data sets, we directly develop our method in a kernel space, which is well known for its ability to explore the nonlinear relation. We design an efficient algorithm to find an optimal solution to our model, and show the theoretical analysis on the connections to kernel k-means, k-means, and spectral clustering methods.

While effective, the kernel in use often has enormous influence on the performance of any kernel method. Unfortunately, the most suitable kernel for a specific task is usually unknown in advance. Exhaustive search on a user-defined pool of kernels is time-consuming and impractical when the sizes of the pool and data become large (Zeng and Cheung 2011). Thus we further propose a multiple kernel algorithm for our model. Another benefit of applying multiple kernels is that we can fully utilize information from different sources equipped with heterogeneous features (Yu et al. 2012). To alleviate the effort for kernel construction and integrating complementary information, we learn an appropriate consensus kernel from a linear combination of multiple input kernels. As a result, our joint model can simultaneously learn the similarity information, cluster indicator matrix, and the optimal combination of multiple kernels. Extensive empirical results on real-world benchmark data sets show that our method consistently outperforms other state-of-the-art methods.

Notations. In this paper, matrices are written as upper case letters and vectors are represented by boldface lowercase letters. The i-th column and the (i, j)-th element of matrix X are denoted by X_i and x_{ij}, respectively. The ℓ_2-norm of a vector $\mathbf{x}$ is defined as $\|\mathbf{x}\|^2 = \mathbf{x}^T \cdot \mathbf{x}$, where T means transpose. I denotes the identity matrix and $\mathbf{1}$ denotes a column vector with all the elements as one. Tr$(\cdot)$ is the trace operator. $0 \leq Z \leq 1$ means all elements of Z are in the range $[0, 1]$.

Clustering with Single Kernel

According to the self-expressive property (Elhamifar and Vidal 2009),

$$X_i \approx \sum_j X_j z_{ij}, \quad s.t. \quad Z_i^T \mathbf{1} = 1, 0 \leq z_{ij} \leq 1, \qquad (1)$$

where z_{ji} is the weight for j-th sample. More similar data points should receive bigger weights and the weights should be smaller for less similar points. Thus Z is also called similarity matrix, which represents the global structure of data. Note that (1) is in a similar spirit of Locally Linear Embedding (LLE) (Roweis and Saul 2000), which assumes that the data points lie on a manifold and each data point can be expressed as a linear combination of its nearest neighbors. The difference from LLE lies in the fact that we specify no neighborhood, which is automatically determined by our method.

To obtain Z, we solve the following problem:

$$\begin{aligned} &\min_Z \|X - XZ\|_F^2 + \alpha \|Z\|_F^2, \\ &s.t. \quad Z^T \mathbf{1} = \mathbf{1}, \quad 0 \leq Z \leq 1, \end{aligned} \qquad (2)$$

where the first term is to measure reconstruction error, the second term is imposed to avoid the trivial solution $Z = I$, and α is a trade-off parameter.

One drawback of (2) is that it assumes linear relations between samples. To recover the nonlinear relations between the data points, we extend (2) to kernel spaces by deploying a general kernelization framework (Zhang, Nie, and Xiang 2010). Define $\phi : \mathcal{R}^D \to \mathcal{H}$ to be a kernel mapping the data samples from the input space to a reproducing kernel Hilbert space $\mathcal{H}$. For $X = [X_1, \cdots, X_n]$ containing n samples, the transformation is $\phi(X) = [\phi(X_1), \cdots, \phi(X_n)]$. The kernel similarity between data samples X_i and X_j is defined through a predefined kernel as $K_{X_i, X_j} =< \phi(X_i), \phi(X_j) >$. It is easy to observe that all similarities can be computed exclusively using the kernel function and one does not need to know the transformation ϕ. This is known as the kernel trick and it greatly simplifies the computations in the kernel space when the kernels are precomputed. Then (2) becomes

$$\begin{aligned} &\min_Z Tr(K - 2KZ + Z^T K Z) + \alpha \|Z\|_F^2 \\ &s.t. \quad Z^T \mathbf{1} = \mathbf{1}, \quad 0 \leq Z \leq 1. \end{aligned} \qquad (3)$$

By solving above problem, we learn the linear sparse relations of $\phi(X)$, and thus the nonlinear relations among X. Note that (3) goes back to (2) if a linear kernel is adopted.

Ideally, we expect that the number of connected components in Z are exactly c if the given data set X consists of c clusters (that is, Z is block diagonal with proper permutations). However, the solution Z from (3) might not satisfy this desired property. Therefore, we will add another constraint based on the following theorem (Mohar et al. 1991).

Theorem 1. *The multiplicity c of the eigenvalue 0 of the Laplacian matrix L of Z is equal to the number of connected components in the graph with the similarity matrix Z.*

Theorem 1 means that $rank(L) = n - c$ if the similarity matrix Z contains exactly c connected components. Thus our new clustering model is to solve:

$$\begin{aligned} &\min_Z Tr(K - 2KZ + Z^T K Z) + \alpha \|Z\|_F^2 \\ &s.t. \quad Z^T \mathbf{1} = \mathbf{1}, \quad 0 \leq Z \leq 1, \quad rank(L) = n - c. \end{aligned} \qquad (4)$$

Problem (4) is not easy to tackle, since $L := D - \frac{Z^T + Z}{2}$, where $D \in \mathcal{R}^{n \times n}$ is a diagonal matrix with the i-th diagonal element $\sum_j \frac{z_{ij} + z_{ji}}{2}$, also depends on similarity matrix Z.

Here L is positive semi-definite, thus $\sigma_i(L) \geq 0$, where $\sigma_i(L)$ is the i-th smallest eigenvalue of L. $rank(L) = n - c$ is equivalent to $\sum_{i=1}^c \sigma_i(L) = 0$. It is not easy to enforce this constraint because the optimization problem with a rank constraint is known to be of combinatorial complexity. To mitigate the difficulty, (Wang et al. 2015; Nie et al. 2016) incorporates the rank constraint into the objective function as a regularizer. Motivated by this consideration, we relax the constraint and reformulate our model as

$$\begin{aligned} &\min_Z Tr(K - 2KZ + Z^T K Z) + \alpha \|Z\|_F^2 + \beta \sum_{i=1}^c \sigma_i(L) \\ &s.t. \quad Z^T \mathbf{1} = \mathbf{1}, \quad 0 \leq Z \leq 1. \end{aligned} \qquad (5)$$

The minimization will make the regularizer $\sum_{i=1}^c \sigma_i(L) \to 0$ if β is large enough. Then the constraint $rank(L) = n - c$ will be satisfied.

Problem (5) is still a challenging problem because of the last term. Fortunately, it can be solved by using Ky Fan's

Theorem (Fan 1949), i.e.,

$$\sum_{i=1}^{c}\sigma_i(L) = \min_{P^TP=I} Tr(P^TLP), \tag{6}$$

where $P \in \mathcal{R}^{n\times c}$ is the indicator matrix. The c elements of i-th row $P_{i,:} \in \mathcal{R}^{c\times 1}$ are the measure of the membership of data point X_i belonging to the c clusters. Finally, our model of twin learning for similarity and clustering with a single kernel (SCSK) is formulated as

$$\begin{aligned} &\min_{Z,P} Tr(K - 2KZ + Z^TKZ) + \alpha\|Z\|_F^2 + \beta Tr(P^TLP), \\ &s.t. \quad Z^T\mathbf{1} = \mathbf{1}, \quad 0 \le Z \le 1, \quad P^TP = I. \end{aligned} \tag{7}$$

By solving (7), we directly obtain the indicator matrix P; therefore, we do not need to perform spectral clustering any more. By alternatively updating Z and P, they can improve each other and optimize (7).

Optimization Algorithm

We use an alternating optimization strategy for (7). When Z is fixed, (7) becomes

$$\min_{P^TP=I} Tr(P^TLP). \tag{8}$$

The optimal solution P is obtained by the c eigenvectors of L corresponding to the c smallest eigenvalues.

When P is fixed, (7) can be reformulated column-wisely as:

$$\begin{aligned} &\min_{Z_i} K_{ii} - 2K_{i,:}Z_i + Z_i^TKZ_i + \alpha Z_i^TZ_i + \frac{\beta}{2}d_i^TZ_i, \\ &s.t. \quad Z_i^T\mathbf{1} = 1, \quad 0 \le z_{ij} \le 1, \end{aligned} \tag{9}$$

where $d_i \in \mathcal{R}^{n\times 1}$ is a vector with the j-th element d_{ij} being $d_{ij} = \|P_{i,:} - P_{j,:}\|^2$. To obtain (9), the important equation in spectral analysis

$$\sum_{i,j}\frac{1}{2}\|P_{i,:} - P_{j,:}\|^2 z_{ij} = Tr(P^TLP) \tag{10}$$

is used. (9) can be further simplified as

$$\begin{aligned} &\min_{Z_i} Z_i^T(\alpha I + K)Z_i + (\frac{\beta d_i^T}{2} - 2K_{i,:})Z_i, \\ &s.t. \quad Z_i^T\mathbf{1} = 1, \quad 0 \le z_{ij} \le 1. \end{aligned} \tag{11}$$

This problem can be solved by many existing quadratic programing packages. The complete algorithm is outlined in Algorithm 1.

Algorithm 1 The algorithm of SCSK

Input: Kernel matrix K, parameters $\alpha > 0, \beta > 0$.

Initialize: Random matrix Z.

REPEAT

1: Update P, which is formed by the c eigenvectors of $L = D - \frac{Z^T+Z}{2}$ corresponding to the c smallest eigenvalues.

2: For each i, update the i-th column of Z by solving problem (11).

UNTIL stopping criterion is met.

Theoretical Analysis of SCSK Model

In this section, we present a theoretical analysis of SCSK and its connections to kernel k-means, k-means, and SC.

Connection to Kernel K-means and K-means

Here we introduce a theorem which states the equivalence of SCSK and kernel k-means, k-means under some condition.

Theorem 2. *When $\alpha \to \infty$, the problem (4) is equivalent to kernel k-means problem.*

Proof. The constraint $rank(L) = n - c$ in (4) makes the solution Z block diagonal. Let $Z^i \in \mathcal{R}^{n_i\times n_i}$ denote the similarity matrix of the i-th component of Z, where n_i is the number of data samples in the component. Problem (4) is equivalent to solving the following problem for each i:

$$\begin{aligned} &\min_{Z^i} \|\phi(X^i) - \phi(X^i)Z^i\|_F^2 + \alpha\|Z^i\|_F^2 \\ &s.t. \quad (Z^i)^T\mathbf{1} = \mathbf{1}, \quad 0 \le Z^i \le 1, \end{aligned} \tag{12}$$

where X^i consists of the samples corresponding to the i-th component of Z. When $\alpha \to \infty$, the above problem becomes

$$\begin{aligned} &\min_{Z^i} \alpha\|Z^i\|_F^2 \\ &s.t. \ (Z^i)^T\mathbf{1} = \mathbf{1}, \ 0 \le Z^i \le 1. \end{aligned} \tag{13}$$

The optimal solution is that all elements of Z^i are equal to $\frac{1}{n_i}$.

Thus when $\alpha \to \infty$, the optimal solution Z to problem (4) is

$$z_{ij} = \begin{cases} \frac{1}{n_k}, & \text{if } X_i \text{ and } X_j \text{ are in the same } k\text{-th component} \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

Denote the solution set of this form by $\mathcal{C}$. It is easy to observe that $\|Z\|_F^2 = c$. Thus (4) becomes

$$\min_{Z_i\in\mathcal{C}} \sum_i \|\phi(X_i) - \phi(X)Z_i\|^2 \tag{15}$$

It is easy to deduce that $\phi(X)Z_i$ is the mean of cluster c_i in the kernel space. Therefore, (15) is exactly the kernel k-means. Thus our proposed algorithm is to solve the kernel k-means problem when $\alpha \to \infty$. □

Corollary 1. *When $\alpha \to \infty$ and a linear kernel is adopted, the problem (4) is equivalent to k-means problem.*

Proof. It is obvious when one does not use any transformations on X in (15). □

Connection to Spectral Clustering

With a predefined similarity Z, spectral clustering is to solve the following problem:

$$\min_{P^TP=I} Tr(P^TLP). \tag{16}$$

The optimal solution P is obtained by the c eigenvectors of L corresponding to the c smallest eigenvalues. Generally, P can not be directly used for clustering since Z does not have exactly c connected components. To obtain the final clustering results, k-means or some other discretization procedures must be performed on P (Huang, Nie, and Huang 2013).

In our proposed algorithm, the similarity matrix Z is not predefined as the existing spectral clustering methods in the literature. Also, the similarity matrix Z is learned by taking account of the clustering task at hand, as opposed to the existing subspace clustering methods in the literature which only focus on learning the similarity matrix Z without considering the effect of clustering on Z (Peng et al. 2015). In our method, the graph with the learned Z will be partitioned into c connected components by using P. The optimal solution P is formed by the c eigenvectors of L, which is defined by Z, corresponding to the c smallest eigenvalues. Therefore, the proposed algorithm learns the similarity matrix Z and the cluster indicator matrix P simultaneously in a coupled way, which leads to a better result in real applications than existing spectral methods as shown in our experiments, since it learns an adaptive graph for clustering.

Clustering with Multiple Kernels

Although model (7) can automatically learn the similarity matrix and cluster indicator matrix, its performance will largely be determined by the choice of kernel. It is often impractical to exhaustively search for the most suitable kernel. Moreover, real world data sets are often generated from different sources along with heterogeneous features. Single kernel method may not be able to fully utilize such information. Multiple kernel learning is capable of integrating complementary information and identifying a suitable kernel for a given task. Here we present a way to learn an appropriate consensus kernel from a convex combination of several predefined kernel matrices.

Suppose there are a total number of r different kernel functions $\{K^i\}_{i=1}^r$. Correspondingly, there would be r different kernel spaces denoted as $\{\mathcal{H}^i\}_{i=1}^r$. An augmented Hilbert space, $\tilde{\mathcal{H}} = \bigoplus_{i=1}^r \mathcal{H}^i$, can be constructed by concatenating all kernel spaces and by using the mapping of $\tilde{\phi}(\mathbf{x}) = [\sqrt{w_1}\phi_1(\mathbf{x}), \sqrt{w_2}\phi_2(\mathbf{x}), ..., \sqrt{w_r}\phi_r(\mathbf{x})]^T$ with different weights $\sqrt{w_i}(w_i \geq 0)$. Then the combined kernel K_w can be represented as (Zeng and Cheung 2011)

$$K_w(\mathbf{x}, \mathbf{y}) =< \phi_w(\mathbf{x}), \phi_w(\mathbf{y}) >= \sum_{i=1}^r w_i K^i(\mathbf{x}, \mathbf{y}). \tag{17}$$

Note that the convex combination of the positive semi-definite kernel matrices $\{K^i\}_{i=1}^r$ is still a positive semi-definite kernel matrix. Thus the combined kernel still satisfies Mercer's condition. Then we propose our joint similarity learning and clustering with multiple kernel (SCMK) model which can be written as

$$\begin{aligned} &\min_{Z,P,\mathbf{w}} Tr(K_w - 2K_wZ + Z^TK_wZ) + \alpha\|Z\|_F^2 + \beta Tr(P^TLP), \\ &s.t. \quad Z^T\mathbf{1} = \mathbf{1}, \quad 0 \leq Z \leq 1, \quad P^TP = I, \\ &\qquad K_w = \sum_{i=1}^r w_i K^i, \sum_{i=1}^r \sqrt{w_i} = 1, w_i \geq 0. \end{aligned} \tag{18}$$

By iteratively updating $Z, P, \mathbf{w}$, each of them will be adaptively refined according to the results of the other two.

Optimization

Problem (18) can be solved by alternatively updating Z, P, and $\mathbf{w}$, while holding the other variables as constant.

1) Optimizing with respect to Z and P when $\mathbf{w}$ is fixed: We can directly calculate K_w, and the optimization problem is exactly (7). Thus we just need to use Algorithm 1 with K_w as the input kernel matrix.

2) Optimizing with respect to $\mathbf{w}$ when Z and P are fixed: Solving (18) with respect to $\mathbf{w}$ can be rewritten as (Cai et al. 2013)

$$\min_{\mathbf{w}} \sum_{i=1}^r w_i h_i \quad s.t. \quad \sum_{i=1}^r \sqrt{w_i} = 1, \quad w_i \geq 0, \tag{19}$$

where

$$h_i = Tr(K^i - 2K^iZ + Z^TK^iZ). \tag{20}$$

The Lagrange function of (19) is

$$\mathcal{J}(\mathbf{w}) = \mathbf{w}^T\mathbf{h} + \gamma(1 - \sum_{i=1}^r \sqrt{w_i}). \tag{21}$$

By utilizing the Karush-Kuhn-Tucker (KKT) condition with $\frac{\partial \mathcal{J}(\mathbf{w})}{\partial w_i} = 0$ and the constraint $\sum_{i=1}^r \sqrt{w_i} = 1$, we obtain the solution of $\mathbf{w}$ as follows:

$$w_i = (h_i \sum_{j=1}^r \frac{1}{h_j})^{-2}. \tag{22}$$

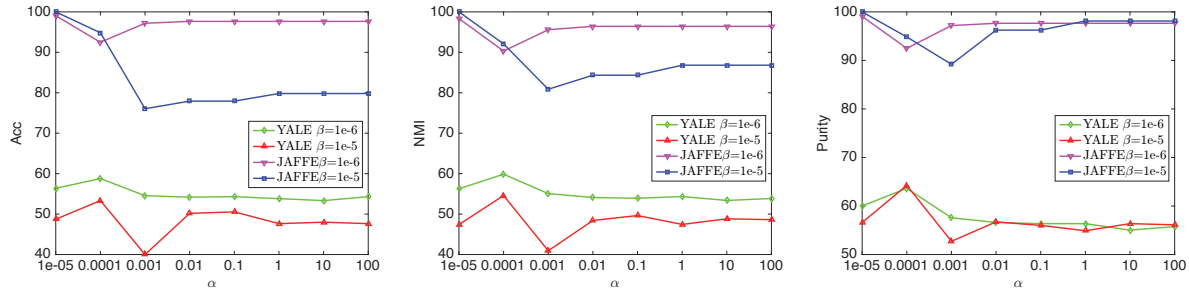
In Algorithm 2 we provide a complete algorithm for solving (18).

Algorithm 2 The algorithm of SCMK

Input: A set of kernel matrices $\{K^i\}_{i=1}^r$, parameters $\alpha > 0$, $\beta > 0$.
Initialize: Random matrix Z, $w_i = 1/r$.
REPEAT
1: Calculate $K_{\mathbf{w}}$ by (17).
2: Update P with the c smallest eigenvectors of $L = D - \frac{Z^T+Z}{2}$.
3: For each i, update the i-th column of Z by (11).
4: Calculate $\mathbf{h}$ by (20).
5: Update $\mathbf{w}$ by (22).
UNTIL stopping criterion is met.

Experiments

In this section, we demonstrate the effectiveness of the proposed method on real world benchmark data sets.



(a) Acc (b) NMI (c) Purity

Figure 1: The effect of parameters α and β on the YALE and JAFFE data sets.

Table 1: Description of the data sets

	# instances	# features	# classes
YALE	165	1024	15
JAFFE	213	676	10
ORL	400	1024	40
AR	840	768	120
BA	1404	320	36
TR11	414	6429	9
TR41	878	7454	10
TR45	690	8261	10

Data Sets

There are altogether eight benchmark data sets used in our experiments. Table 1 summarizes the statistics of these data sets. Among them, five are image ones, and the other three are text corpora[1]. The five image data sets consist of four commonly used face databases (ORL[2], YALE[3], AR[4] (Martinez and Benavente 2007) and JAFFE[5]), and a binary alpha digits data set BA[6].

Experiment Setup

To assess the effectiveness of multiple kernel learning, we design 12 kernels which include: seven Gaussian kernels of the form $K(\mathbf{x},\mathbf{y}) = exp(-\|\mathbf{x}-\mathbf{y}\|_2^2/(td_{max}^2))$, where d_{max} is the maximal distance between samples and t varies over the set $\{0.01, 0.05, 0.1, 1, 10, 50, 100\}$; a linear kernel $K(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y}$; four polynomial kernels $K(\mathbf{x},\mathbf{y}) = (a + \mathbf{x}^T\mathbf{y})^b$ with $a = \{0, 1\}$ and $b = \{2, 4\}$. Furthermore, all kernels are rescaled to $[0, 1]$ by dividing each element by the largest pair-wise squared distance.

For single kernel methods, we run kernel k-means (KKM) (Schölkopf, Smola, and Müller 1998), spectral clustering (SC) (Ng et al. 2002), robust kernel k-means (RKKM) (Du et al. 2015), and our proposed SCSK[7] on each kernel separately. The methods in comparison are downloaded from the their authors' websites. And we report both the best and the average results over all these kernels.

For multiple kernel methods, we implement the following algorithms on a combination of above kernels.

MKKM[8]. MKKM (Huang, Chuang, and Chen 2012b) extends k-means in a multiple-kernel setting. However, it uses a different way to learn the kernel weight.

AASC[9]. AASC (Huang, Chuang, and Chen 2012a) extends spectral clustering to the situation where multiple affinities exist.

RMKKM[10]. RMKKM (Du et al. 2015) adopts ℓ_{21} norm to measure the loss of k-means.

SCMK. Our proposed method for joint similarity learning and clustering with multiple kernels.

For spectral clustering like SC and AASC, we run k-means on spectral embedding to obtain the clustering results. To reduce the influence of initialization, we follow the strategy suggested in (Yang et al. 2010; Du et al. 2015), and we repeat clustering 20 times and present the results with the best objective values. We set the number of clusters to the true number of classes for all clustering algorithms.

To quantitatively evaluate the clustering performance, we adopt the three widely used metrics, accuracy (Acc), normalized mutual information (NMI) (Cai et al. 2009), and Purity.

Clustering Result

Table 2 shows the clustering results in terms of accuracy, NMI and Purity on all the data sets. It can be seen that the proposed SCSK and SCMK produce promising results. Especially, our method can substantially improve the performance on JAFFE, AR, BA, TR11, and TR45 data sets. The big difference between best and average results confirms the fact that the choice of kernel has a huge influence on the performance of single kernel methods. This difference motivates the development of multiple kernel learning method.

[1] http://www-users.cs.umn.edu/ han/data/tmdata.tar.gz

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

[3] http://vision.ucsd.edu/content/yale-face-database

[4] http://www2.ece.ohio-state.edu/ aleix/ARdatabase.html

[5] http://www.kasrl.org/jaffe.html

[6] http://www.cs.nyu.edu/ roweis/data.html

[7] https://github.com/sckangz/AAAI17

[8] http://imp.iis.sinica.edu.tw/IVCLab/research/Sean/mkfc/code

[9] http://imp.iis.sinica.edu.tw/IVCLab/research/Sean/aasc/code

[10] https://github.com/csliangdu/RMKKM

(a) Accuracy(%)

Data	KKM	KKM-a	SC	SC-a	RKKM	RKKM-a	SCSK	SCSK-a	MKKM	AASC	RMKKM	SCMK
YALE	47.12	38.97	49.42	40.52	48.09	39.71	**55.85**	45.35	45.70	40.64	52.18	**56.97**
JAFFE	74.39	67.09	74.88	54.03	75.61	67.98	**99.83**	86.64	74.55	30.35	87.07	**100.00**
ORL	53.53	45.93	57.96	46.65	54.96	46.88	**62.35**	50.50	47.51	27.20	55.60	**65.25**
AR	33.02	30.89	28.83	22.22	33.43	31.20	**56.79**	41.35	28.61	33.23	34.37	**62.38**
BA	41.20	33.66	31.07	26.25	42.17	34.35	**47.72**	39.50	40.52	27.07	43.42	**47.34**
TR11	51.91	44.65	50.98	43.32	53.03	45.04	**71.26**	54.79	50.13	47.15	57.71	**73.43**
TR41	55.64	46.34	63.52	44.80	56.76	46.80	**67.43**	53.13	56.10	45.90	62.65	**67.31**
TR45	58.79	45.58	57.39	45.96	58.13	45.69	**74.02**	53.38	58.46	52.64	64.00	**74.35**

(b) NMI(%)

Data	KKM	KKM-a	SC	SC-a	RKKM	RKKM-a	SCSK	SCSK-a	MKKM	AASC	RMKKM	SCMK
YALE	51.34	42.07	52.92	44.79	52.29	42.87	**56.50**	45.07	50.06	46.83	55.58	**56.52**
JAFFE	80.13	71.48	82.08	59.35	83.47	74.01	**99.35**	84.67	79.79	27.22	89.37	**100.00**
ORL	73.43	63.36	75.16	66.74	74.23	63.91	**78.96**	63.55	68.86	43.77	74.83	**80.04**
AR	65.21	60.64	58.37	56.05	65.44	60.81	**76.02**	59.70	59.17	65.06	65.49	**81.51**
BA	57.25	46.49	50.76	40.09	57.82	46.91	**63.04**	52.17	56.88	42.34	58.47	**62.94**
TR11	48.88	33.22	43.11	31.39	49.69	33.48	**58.60**	37.58	44.56	39.39	56.08	**60.15**
TR41	59.88	40.37	61.33	36.60	60.77	40.86	**65.50**	43.18	57.75	43.05	63.47	**65.11**
TR45	57.87	38.69	48.03	33.22	57.86	38.96	**74.24**	44.36	56.17	41.94	62.73	**74.97**

(c) Purity(%)

Data	KKM	KKM-a	SC	SC-a	RKKM	RKKM-a	SCSK	SCSK-a	MKKM	AASC	RMKKM	SCMK
YALE	49.15	41.12	51.61	43.06	49.79	41.74	**57.27**	55.79	47.52	42.33	53.64	**60.00**
JAFFE	77.32	70.13	76.83	56.56	79.58	71.82	**99.85**	96.53	76.83	33.08	88.90	**100.00**
ORL	58.03	50.42	61.45	51.20	59.60	51.46	**74.00**	70.37	52.85	31.56	60.23	**77.00**
AR	35.52	33.64	33.24	25.99	35.87	33.88	**63.45**	62.37	30.46	34.98	36.78	**82.62**
BA	44.20	36.06	34.50	29.07	45.28	36.86	**52.36**	49.79	43.47	30.29	46.27	**52.12**
TR11	67.57	56.32	58.79	50.23	67.93	56.40	**82.85**	80.76	65.48	54.67	72.93	**87.44**
TR41	74.46	60.00	73.68	56.45	**74.99**	60.21	73.23	71.21	72.83	62.05	**77.57**	73.69
TR45	68.49	53.64	61.25	50.02	68.18	53.75	**78.26**	77.76	69.14	57.49	75.20	**78.26**

Table 2: Clustering results measured on benchmark data sets. '-a' denotes the average performance on those 12 kernels. Both the best results for single kernel and multiple kernel methods are highlighted in boldface.

Besides, multiple kernel clustering approaches usually improve the results over single kernel clustering methods.

Parameter Selection

There are two parameters α and β in our models. We let α vary over the range of $\{$1e-5, 1e-4, 1e-3, 0.01, 0.1, 1, 10, 100$\}$, and β over $\{$1e-6, 1e-5$\}$. Figure 1 shows how the clustering results of SCMK in terms of Acc, NMI, and Purity vary with α and β on JAFFE and YALE data sets. We can observe that the performance of SCMK is very stable with respect to a large range of α values and it is more sensitive to the value of β.

Conclusion

In this paper, we first propose a clustering method to simultaneously perform similarity learning and the cluster indicator matrix construction. In our method, the similarity learning and the cluster indicator learning are integrated within one framework; the method can be easily extended to kernel spaces, so as to capture nonlinear structure information. The connections of the proposed method to kernel k-means, k-means, and spectral clustering are also established. To avoid extensive search of the best kernel, we further incorporate multiple kernel learning into our model. Similarity learning, cluster indicator construction, and kernel weight learning can be boosted by using the results of the other two. Extensive experiments have been conducted on real-world benchmark data sets to demonstrate the superior performance of our method.

Acknowledgements

This work is supported by US National Science Foundation Grants IIS 1218712. Q. Cheng is the corresponding author.

References

Cai, D.; He, X.; Wang, X.; Bao, H.; and Han, J. 2009. Locality preserving nonnegative matrix factorization. In *IJCAI*, volume 9, 1010–1015.

Cai, X.; Nie, F.; Cai, W.; and Huang, H. 2013. Heterogeneous image features integration via multi-modal semi-

supervised learning model. In *Proceedings of the IEEE International Conference on Computer Vision*, 1737–1744.

Du, L.; Zhou, P.; Shi, L.; Wang, H.; Fan, M.; Wang, W.; and Shen, Y.-D. 2015. Robust multiple kernel k-means using 2; 1-norm. In *Proceedings of the 24th International Conference on Artificial Intelligence*, 3476–3482. AAAI Press.

Elhamifar, E., and Vidal, R. 2009. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2790–2797. IEEE.

Fan, K. 1949. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences of the United States of America* 35(11):652.

Huang, H.-C.; Chuang, Y.-Y.; and Chen, C.-S. 2012a. Affinity aggregation for spectral clustering. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 773–780. IEEE.

Huang, H.-C.; Chuang, Y.-Y.; and Chen, C.-S. 2012b. Multiple kernel fuzzy clustering. *IEEE Transactions on Fuzzy Systems* 20(1):120–134.

Huang, J.; Nie, F.; and Huang, H. 2013. Spectral rotation versus k-means in spectral clustering. In *AAAI*.

Huang, J.; Nie, F.; and Huang, H. 2015. A new simplex sparse learning model to measure data similarity for clustering. In *Proceedings of the 24th International Conference on Artificial Intelligence*, 3569–3575. AAAI Press.

Johnson, S. C. 1967. Hierarchical clustering schemes. *Psychometrika* 32(3):241–254.

Kang, Z., and Cheng, Q. 2016. Top-n recommendation with novel rank approximation. In *2016 SIAM Int. Conf. on Data Mining (SDM 2016)*, 126–134.

Kang, Z.; Peng, C.; and Cheng, Q. 2015a. Robust subspace clustering via smoothed rank approximation. *IEEE Signal Processing Letters* 22(11):2088–2092.

Kang, Z.; Peng, C.; and Cheng, Q. 2015b. Robust subspace clustering via tighter rank approximation. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 393–401. ACM.

Luo, D.; Nie, F.; Ding, C.; and Huang, H. 2011. Multi-subspace representation and discovery. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 405–420. Springer.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 281–297. Oakland, CA, USA.

Martinez, A., and Benavente, R. 2007. The ar face database, 1998. *Computer Vision Center, Technical Report* 3.

Mohar, B.; Alavi, Y.; Chartrand, G.; and Oellermann, O. 1991. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications* 2(871-898):12.

Ng, A. Y.; Jordan, M. I.; Weiss, Y.; et al. 2002. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* 2:849–856.

Nie, F.; Wang, X.; Jordan, M. I.; and Huang, H. 2016. The constrained laplacian rank algorithm for graph-based clustering. In *Thirtieth AAAI Conference on Artificial Intelligence*. Citeseer.

Nie, F.; Wang, X.; and Huang, H. 2014. Clustering and projected clustering with adaptive neighbors. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 977–986. ACM.

Peng, C.; Kang, Z.; Li, H.; and Cheng, Q. 2015. Subspace clustering using log-determinant rank approximation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 925–934. ACM.

Peng, C.; Kang, Z.; Hu, Y.; Cheng, J.; and Cheng, Q. 2016. Nonnegative matrix factorization with integrated graph and feature learning. *ACM Transactions on Intelligent Systems and Technology*.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.

Schölkopf, B.; Smola, A.; and Müller, K.-R. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* 10(5):1299–1319.

Wang, X.; Liu, Y.; Nie, F.; and Huang, H. 2015. Discriminative unsupervised dimensionality reduction. In *Proceedings of the 24th International Conference on Artificial Intelligence*, 3925–3931. AAAI Press.

Yang, Y.; Xu, D.; Nie, F.; Yan, S.; and Zhuang, Y. 2010. Image clustering using local discriminant models and global integration. *IEEE Transactions on Image Processing* 19(10):2761–2773.

Yu, S.; Tranchevent, L.; Liu, X.; Glanzel, W.; Suykens, J. A.; De Moor, B.; and Moreau, Y. 2012. Optimized data fusion for kernel k-means clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(5):1031–1039.

Zeng, H., and Cheung, Y.-m. 2011. Feature selection and kernel learning for local learning-based clustering. *IEEE transactions on pattern analysis and machine intelligence* 33(8):1532–1547.

Zhang, C.; Nie, F.; and Xiang, S. 2010. A general kernelization framework for learning algorithms based on kernel pca. *Neurocomputing* 73(4):959–967.