The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Dictionary Learning in Optimal Metric Space

Jiexi Yan,[1] **Cheng Deng,**[1*] **Xianglong Liu**[2]
[1]School of Electronic Engineering, Xidian University, Xian 710071, China
[2]Beihang University, Beijing 100191, China
jxyan@stu.xidian.edu.cn, chdeng.xd@gmail.com, xlliu@nlsde.buaa.edu.cn

Abstract

Dictionary learning has been widely used in machine learning field to address many real-world applications, such as classification and denoising. In recent years, many new dictionary learning methods have been proposed. Most of them are designed to solve unsupervised problem without any prior information or supervised problem with the label information. But in real world, as usual, we can only obtain limited side information as prior information rather than label information. The existing methods don't take into account the side information, let alone learning a good dictionary through using the side information. To tackle it, we propose a new unified unsupervised model which naturally integrates metric learning to enhance dictionary learning model with fully utilizing the side information. The proposed method updates metric space and dictionary adaptively and alternatively, which ensures learning optimal metric space and dictionary simultaneously. Besides, our method can also deal well with high-dimensional data. Extensive experiments show the efficiency of our proposed method, and a better performance can be derived in real-world image clustering applications.

Introduction

Dictionary Learning (Toi and Frossard 2011) and sparse representation (Olshausen and Field 1997) are crucial tools in machine learning field. Dictionary is able to learn an adaptive set of basis elements (dictionary) from data instead of predefined ones (Mallat 1999), so that every data sample can be represented by sparse linear combination of these basis vectors. Since dictionary learning has proven its effectiveness on numerous machine learning tasks, such as classification (Zhang and Li 2010), denoising (Aharon, Elad, and Bruckstein 2006) and self-taught learning (Wang, Nie, and Huang 2013), many researchers worked on this topic and a large amount of algorithms and methods were proposed to solve dictionary learning problem. (Olshausen and Field 1997; Lewicki and Sejnowski 2006; Aharon, Elad, and Bruckstein 2006; Mairal et al. 2009a)

In the past two decades, many researchers have proposed different kinds of dictionary learning approaches to adapt and solve different application problem, such as unsupervised problem (Aharon, Elad, and Bruckstein 2006; Mairal et al. 2009a), supervised problem (Mairal et al. 2009b; Zhang and Li 2010; Jiang, Lin, and Davis 2011) and semi-supervised problems (Wang et al. 2013). Among these approaches, some learn dictionary without any prior information, while the others require label information. However, when dealing with real-world problems, we often only get the side information i.e. pairwise constrains which indicate whether two objects in a pair belong to the same class.

Without the use of the side information, data representation obtained from traditional unsupervised dictionary learning method such as (Aharon, Elad, and Bruckstein 2006) is not good enough. To take full advantage of the side information, we use metric learning method as a preprocessing step and process dictionary learning after this step. But in this way, the obtained metric space is only suboptimal and the learned dictionary on this suboptimal metric space is also not optimal. So this simple two-step method is heuristic but not optimal.

To handle the above problem, we naturally integrate metric learning to dictionary learning to enhance dictionary learning model with fully utilizing the side information and derive a novel unified model. As shown in Figure 1, the proposed method updates metric space and dictionary adaptively and iteratively, so we can learn both optimal metric space and optimal dictionary. In this unified model, we explicitly incorporate a sparse coding reconstruction error criterion and a trace ratio of pairwise constraints criterion into a unified objective function. In addition, because of using metric learning to reduce the dimension of the original data, our method can deal well with high-dimensional data, in which redundant features may confuse dictionary learning. And then we introduce an optimization algorithm to update dictionary D and metric space W alternatively and iteratively. We perform image clustering experiments on several real-world image data sets, including human face recognition benchmark data sets and objective recognition benchmark data sets. Results show that our proposed dictionary learning in optimal projection subspace model works effectively and outperforms other compared methods.

Related Work

In this section, we give a brief review of the development of dictionary learning and sparse coding.

As we know, dictionary learning is widely used in the field

*Corresponding author.

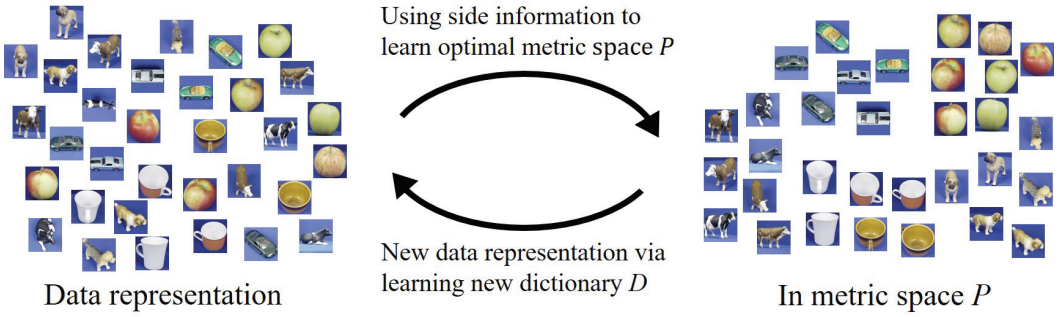


Figure 1: A brief illustration of the proposed unified model on ETH-80 data set

of machine learning, and its effectiveness has been proven in several applications such as classification and denoising. Given a data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \in \mathbb{R}^{d\times n}$ where d is the dimension of features and n is the number of data points, dictionary learning problem can be formulated as follows,

$$\min_{Y,D} \sum_{i=1}^{n} \frac{1}{2}\|\mathbf{x}_i - D\mathbf{y}_i\|_2^2 + \lambda\varphi(\mathbf{y}_i) \tag{1}$$

where the learned dictionary is denoted as D and $Y = \{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$ represents sparse representation. In this objective, the first term is the reconstruction error and the second term $\varphi(\mathbf{y}_i)$ represents a regularization term, besides λ is a parameter used to balance the two terms in problem (1).

When using ℓ_0 norm, namely $\varphi(\mathbf{y}_i) = \|\mathbf{y}\|_0$, this problem becomes an NP-hard problem and is hard to optimize. Aharon, Elad, and Bruckstein (2006) transformed problem (1) to be

$$\begin{aligned} &\min_{D,Y} \quad \|X - DY\|_F^2 \\ &s.t. \quad \forall i \ \|\mathbf{y}_i\|_0 \le T_0, \forall j \ \|\mathbf{d}_j\|_2 = 1 \end{aligned} \tag{2}$$

where T_0 is the sparsity constraint on each $\mathbf{y}_i$. They additionally proposed K-SVD algorithm to solve this problem. K-SVD is an iterative method which contains two steps. The first step is sparse coding step which represents the samples based on the current dictionary and the second step dictionary update step updates the overcomplete dictionary atoms to better fit the samples. The K-SVD algorithm is flexible to work with any pursuit method such as OMP approach. This algorithm is effective in many real-world applications. After that, K-SVD method was extended by imposing extra information in the model as a task driven method (Zhang and Li 2010; Jiang, Lin, and Davis 2011). The improved methods are supervised dictionary.

When using ℓ_1 norm, namely $\varphi(\mathbf{y}_i) = \|\mathbf{y}_i\|_1$, the problem (1) can be transformed to,

$$\min_{Y,D} \ \|X - DY\|_F^2 + \lambda\|Y\|_1 \tag{3}$$

This new problem is convex separately with respect to D and Y. To tackle it, Lee et al. (2007) introduced efficient sparse coding algorithm, in which a local optimum can be learned.

In high-dimensional data, distance can not work and many features provide little useful information, which makes learning an overcomplete dictionary difficult. To tackle it, many researchers usually adopt Principle Component Analysis (PCA) projection or random projection as a preprocessing step, after which they process dictionary learning on the obtained projection subspace.

Motivation and Proposed Model

Although there exists many dictionary learning methods which are widely used in real-world applications, when encountering the real-world data set with the side information, no existing dictionary method is suitable for it. In order to adapt the real-world problems better, we naturally integrate metric learning to enhance dictionary learning model with fully utilizing the side information and construct a novel unified model to update metric space and dictionary adaptively and alternatively, hence optimal metric space and optimal dictionary can be learned simultaneously. Because of the learned optimal metric space, the proposed method can also deal well with high-dimensional data.

Learning a Projection Matrix via Metric Learning

Distance metric plays a critical role in real-world application. Good distance metrics are crucial to many computer vision tasks, such as image classification and content-based retrieval(Liu and Tsang 2015; Wang, Nie, and Huang 2013; Liu and Tsang 2017). Especially in high-dimensional data, distance usually does not work, therefore we can adopt metric learning method and project the original data to an appropriate metric space.

The goal of metric learning is to learn an adaptive distance, such as Mahalanobis distance $d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T M(\mathbf{x}_i - \mathbf{x}_j}$ for the problem of interest using the information brought by training examples. Most of metric learning methods use weakly-supervised constrains such as pairwise constrains. The pairwise constrains contain the information whether two objects in a pair come from the same class. Pairwise constraints can be represented by $\mathcal{S}$ and $\mathcal{D}$ as

$$\begin{cases} \mathcal{S} &= \{(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are in the same class}\} \ , \\ \mathcal{D} &= \{(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are not in the same class}\} \ , \end{cases} \tag{4}$$

where we call $\mathcal{S}$ as must-links and $\mathcal{D}$ as cannot-links (Xing et al. 2003).

Xing et al. (2003) first studied how to learn a distance metric from must-links and cannot-links. Relevance Component Analysis (RCA) (Bar-Hillel et al. 2003) was then proposed and was improved later by Discriminative Component Analysis (DCA) and Kernel DCA (Hoi et al. 2006). Despite their effectiveness, when dealing with high-dimensional data, Xing's approach is computationally inefficient, and both RCA and DCA face the singular problem when computing the covariance matrix for the data point pairs in the must-links. To tackle this, Xiang, Nie, and Zhang (2008) proposed a new framework which formulated the distance metric learning problem as a trace ratio minimization problem as follows.

Considering that Mahalanobis distance metric $M \in \mathbb{R}^{d\times d}$ is a positive semi-definite matrix, we can reasonably write $M = PP^T$ by eigen-decomposition, where $P \in \mathbb{R}^{d\times k}$ with $k \leq d$ and k is the projection dimensionality. Thus the Mahalanobis metric can be rewritten as $d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T PP^T(\mathbf{x}_i - \mathbf{x}_j} = \|P^T(\mathbf{x}_i - \mathbf{x}_j)\|_2$, which indeed defines a transformation of $\mathbf{z} = P^T\mathbf{x}$ under the projection matrix P. Then we calculate the covariance matrix of the data pairs in the must-links and cannot-links respectively as follow:

$$\begin{aligned} S_w &= \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T, \\ S_b &= \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T. \end{aligned} \tag{5}$$

Xiang, Nie, and Zhang (2008) proposed to learn the projection matrix W by solving the following objective:

$$\begin{aligned} \min_P \quad & \frac{\mathrm{Tr}(P^T S_w P)}{\mathrm{Tr}(P^T S_b P)} \\ =& \frac{\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \|(\mathbf{x}_i - \mathbf{x}_j)^T P\|_2^2}{\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \|(\mathbf{x}_i - \mathbf{x}_j)^T P\|_2^2} \\ =& \frac{\|AP\|_F^2}{\|BP\|_F^2} \\ s.t. \quad & P^T P = I \end{aligned} \tag{6}$$

where each row of A is one $(\mathbf{x}_i - \mathbf{x}_j)^T$ that satisfies $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}$, and similarly each row of B is one $(\mathbf{x}_i - \mathbf{x}_j)^T$ that satisfies $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}$.

Dictionary Learning in Optimal Metric Space

In traditional dictionary, when facing high-dimensional data, many previous studies adopt a two-step method i.e. using a preprocessing step to project the original data into a low-dimensional subspace via Principle Component Analysis (PCA) projection or random projection and then do dictionary learning on it. We can also follow this idea to use metric learning as a preprocessing step. But metric learning on the original data only obtain a suboptimal metric space and can not help learn a good dictionary on this suboptimal metric space, which means this two-step method is heuristic but not optimal. A bad projection can lead to a failure in the following dictionary learning process, in which case the sparse representation of a sample based on the degenerated dictionary can be misleading and useless. Thus, it is necessary to pay attention to finding an optimal metric space that works effectively in the following dictionary learning step.

In this paper, we propose to learn dictionary and metric space simultaneously, which confirms both of them optimal. To do this, we combine objective (6) with the traditional dictionary learning model. In this way, a metric space which is the most suitable for dictionary learning can be learned automatically. It not only takes full advantage of the side information to learn dictionary in optimal metric space but also deals well with high-dimensional data.

Given a data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \in \mathbb{R}^{d\times n}$ and relevant side information, our proposed unified dictionary learning model, which learns dictionary in optimal metric space, can be formulated as follow:

$$\begin{aligned} \min_{D,Y,P} \quad & \frac{\|P^T X - DY\|_F^2 \|AP\|_F^2}{\|BP\|_F^2} \\ s.t. \quad & \forall i \quad \|\mathbf{y}_i\|_0 \leq T_0 \\ & \forall j \quad \|\mathbf{d}_j\|_2 = 1 \\ & P^T P = I \end{aligned} \tag{7}$$

where $P \in \mathbb{R}^{d\times k}$ is the metric space matrix, $\mathbf{y}_i$ is the ith column of sparse representation matrix $Y \in \mathbb{R}^{p\times n}$ and $\mathbf{d}_j$ is the jth column of dictionary matrix $D \in \mathbb{R}^{k\times p}$. T_0 is the sparsity constraint on each $\mathbf{y}_i$.

In the objective function (7), the first term represents dictionary learning and sparse learning procedure when we obtain a projection matrix P, and the second term represents the metric learning step which can obtain a projection matrix P. It is obvious that when we perform dictionary learning and sparse representation method to minimize the reconstruction error, we will also take the projection procedure into account. It is easy to know that the update of projection matrix P is affected by the reconstruction error. After the optimization, a metric space matrix P which is suitable for dictionary learning can be learned automatically.

By constructing a unified model which integrates metric learning into dictionary learning, we obtain an optimal metric space, in which the manifold structures are preserved and the distance metric is meaningful. Therefore, comparing with traditional dictionary learning methods which utilize PCA projection or random projection as a preprocessing to reduce dimensionality, the proposed method is more effective in high-dimensional data.

Optimization Algorithm

So far, we have constructed our objective function (7), and it is a non-convex function, which is hard to optimize directly. In this paper, we use Alternative Direction Method (ADM) to simplify the problem (7). In this way, we can convert this complicated multivariate non-convex problem to two subproblems. In the optimization procedure, K-SVD method (Aharon, Elad, and Bruckstein 2006) and projected gradient descent method will be utilized to solve each subproblem.

Algorithm 1 The K-SVD algorithm to solve problem (8)

Input: $Z = P^T X \in \mathbb{R}^{k\times n}$
Output: $D \in \mathbb{R}^{k\times p}, \quad Y \in \mathbb{R}^{p\times n}$
Initialization: Set the dictionary matrix $D^{(0)} \in \mathbb{R}^{k\times p}$ with ℓ_2 normalized columns. Set $t = 1$.
while *not converge* **do**

Sparse Coding Step: Use orthogonal matching pursuit method to compute the sparse representation vectors $\mathbf{y}_i^t$, by approximating the solution of $i = 1, 2, ..., n$, $\min_{\mathbf{y}_i} \|\mathbf{z}_i - D\mathbf{y}_i\|_F^2 \ s.t. \|\mathbf{y}_i\|_0 \le T_0$.
Dictionary Update Step: For each column $j = 1, 2, ..., p$ in D^{t-1}, update it by

- Define the group of examples that use this atom,
$$w_j = \left\{i | 1 \le i \le n, \mathbf{y}_j^t(i) \ne 0\right\}$$
- Compute the overall representation error matrix,
$$E_j = Z - \sum_{h\ne j} \mathbf{d}_h \mathbf{y}_h^t$$
- Restrict E_j by choosing only the columns in w_j and obtain E_j^R.
- Apply SVD decomposition $E_j^R = U\Delta V$. Choose the updated dictionary column $\hat{\mathbf{d}}_j^t$ to be the first column in U. Update the coefficient vector $\mathbf{y}_j^t$ to be the first column of V multiplied by $\Delta(1,1)$.

$t = t + 1$
end while

Update dictionary D and sparse representation Y: When fixing the projection matrix P, the problem (7) becomes:

$$\begin{aligned} \min_{D,Y} \quad & \|P^T X - DY\|_F^2 \\ s.t. \quad & \forall i \quad \|\mathbf{y}_i\|_0 \le T_0 \\ & \forall j \quad \|\mathbf{d}_j\|_2 = 1 \end{aligned} \tag{8}$$

It is a traditional dictionary learning problem, and there are many effective methods to handle it. In this paper, we select K-SVD algorithm to solve it. The procedure of this algorithm is described briefly in Algorithm 1.

Update projection matrix P: When fixing D and Y, the problem (7) becomes a problem with matrix P as:

$$\begin{aligned} \min_{P} \quad & \frac{\|P^T X - DY\|_F^2 \|AP\|_F^2}{\|BP\|_F^2} \\ s.t. \quad & P^T P = I. \end{aligned} \tag{9}$$

where P is the orthogonal matrix. The denominator of this function makes this problem computationally complicated to solve. We adopt the idea from trace ratio problem (Jiang and Chung 2014) to simply this problem, and problem (9) can be transformed to

$$\begin{aligned} \min_{P} \quad & \|P^T X - DY\|_F^2 + \gamma\left(\|AP\|_F^2 - \lambda\|BP\|_F^2\right) \\ s.t. \quad & P^T P = I \end{aligned} \tag{10}$$

Algorithm 2 Algorithm to solve problem (9)

Input: $X \in \mathbb{R}^{d\times n}$, $D \in \mathbb{R}^{k\times p}$, $Y \in \mathbb{R}^{p\times n}$
Output: $W \in \mathbb{R}^{d\times k}$
Initialization: Initialize $D^0 \in \mathbb{R}^{d\times k}$ and set $t = 1$.
while *not converge* **do**

1. Calculate the gradient of problem (10) with respect to P.
$$\begin{aligned} \frac{\partial l(P^{t-1})}{\partial W^{t-1}} &= XX^T P^{t-1} - X\left(DY\right)^T \\ &\quad + \gamma\left(A^T AP^{t-1} - \lambda B^T BP^{t-1}\right) \end{aligned}$$
2. Compute alternative matrix $\hat{P}$:
$$\hat{P} = P^{t-1} - \eta\frac{\partial l(P^{t-1})}{\partial P^{t-1}}$$
3. Compute projection matrix P by solving problem (14):
$$P = UI_{d,k}V$$
where U and V is obtained by SVD of $\hat{P}$, namely $\hat{P} = U\Delta V$.
4. $t = t + 1$

end while

Then this formulation is easy to use projected gradient descent algorithm to solve. At first, we ignore the constraint $P^T P = I$ and adopt gradient descent approach to compute matrix P. Defining the objective function of the problem (10) to be $l(P)$, we have:

$$\begin{aligned} \frac{\partial l(P^t)}{\partial P^t} &= XX^T P^{t-1} - X\left(DY\right)^T \\ &\quad + \gamma\left(A^T AP^t - \lambda B^T BP^t\right) \end{aligned} \tag{11}$$

$$\hat{P} = P^t - \eta\frac{\partial l(P^t)}{\partial P^t} \tag{12}$$

where $\hat{P}$ is the alternative matrix. After that, we project $\hat{P}$ to the domain of matrix P according to the constraint in problem (10) and get:

$$P^{t+1} = \pi(\hat{P}), \tag{13}$$

where

$$\pi(\hat{P}) = arg\min_{P^T P = I} \|W - \hat{P}\|_F^2. \tag{14}$$

Defining the Singular Value Decomposition (SVD) of $\hat{P}$ to be $\hat{P} = U\Delta V$, according to (Manton 2002), the solution to problem (14) is $P = UI_{d,k}V$. Algorithm 2 shows the procedures to solve the problem (9).

Using the algorithms above, we separate the whole problem (7) to two subproblems. The first one is fixing projection matrix to update W and Y, while the other one is fixing matrix D and Y to update P. We compute these two subproblems alternatively and iteratively. In the end, the objective value of problem (7) will converge. To sum up, Algorithm 3 presents the brief structure of our algorithm.

Algorithm 3 Algorithm to solve problem (7)

Input: $X \in \mathbb{R}^{d\times n}$
Output: $D \in \mathbb{R}^{k\times p}$, $Y \in \mathbb{R}^{p\times n}$, $P \in \mathbb{R}^{d\times k}$
Initialization: Initialize $P^0 \in \mathbb{R}^{d\times k}$ and set $t = 1$.
while *not converge* **do**

1. Update matrix D, Y by using K-SVD method in algorithm (1).

$$\mathbf{d}_j = \mathbf{u}_1.$$
$$\mathbf{y}_j = \Delta(1,1)\mathbf{v}_1$$

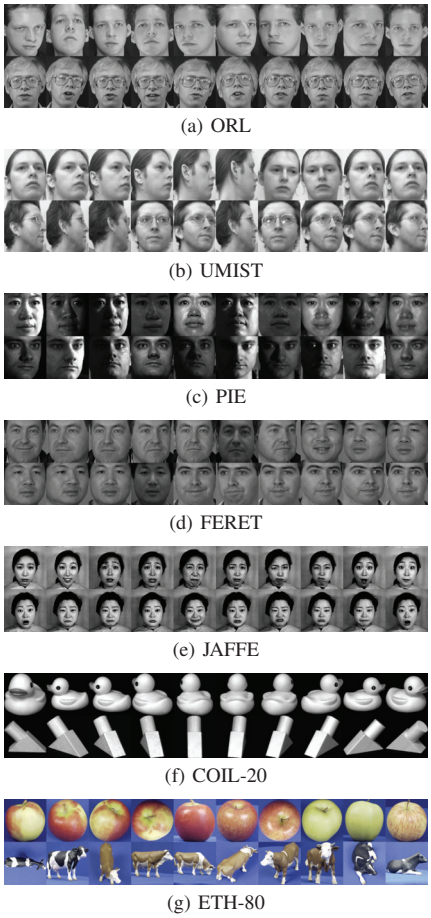
2. Update W by solving problem (9) as Algorithm (2).

$$\frac{\partial l(P^{t-1})}{\partial P^{t-1}} = XX^T P^{t-1} - X\left(DY\right)^T + \gamma\left(A^T AP^{t-1} - \lambda B^T BP^{t-1}\right)$$

$$P^t = \pi(P^{t-1} - \eta\frac{\partial l(P^{t-1})}{\partial P^{t-1}})$$

3. $t = t + 1$

end while



(a) ORL

(b) UMIST

(c) PIE

(d) FERET

(e) JAFFE

(f) COIL-20

(g) ETH-80

Figure 2: The image samples of benchmark data sets

Experimental Results

In this section, we evaluate the proposed method in the task of data clustering, where our goal is to examine the effectiveness of our new method when dealing with the side information.

Data Preparation

We experiment with seven benchmark data sets including five face recognition benchmark data sets **ORL** (Samaria and Harter 1994), **UMIST** (Phillips, Bruce, and Soulie 1998), **PIE** (Sim, Baker, and Bsat 2002), **FERET** (Phillips et al. 1998), **JAFFE** (Lyons et al. 1998) and two objective recognition benchmark data sets **COIL-20** (Nene et al. 1996), **ETH-80** (Leibe and Schiele 2003), whose details are summarized in Table 1. Figure 2 shows the sample images of these benchmark data sets.

Following relevant research, we generate the side information for each set as follows. For each constraint, we randomly pick up one pair of data points from the original data set (the labels of which are available for evaluation purpose but unavailable for clustering). If the labels of this pair of data points are the same, we generate a must-link, otherwise a cannot-link is generated.

Experiment Setup

To evaluate the performance of our proposed method, we compare it with some state-of-the-art two-step methods as follow:

- **Rand+KSVD:** Use Random Projection to obtain Projection Subspace and then use K-SVD method (Aharon, Elad, and Bruckstein 2006) to do dictionary learning on it.
- **Rand+ODL:** Use Random Projection to obtain Projection Subspace and then use Online Dictionary Learning method (Mairal et al. 2009a) to do dictionary learning on it.
- **Rand+SSC:** Use Random Projection to obtain Projection Subspace and then use Semi-Supervised Clustering method (Basu, Banerjee, and Mooney 2004) with the side information to do clustering on it.
- **Rand+Kmeans:** Use Random Projection to obtain Projection Subspace and then use K-means method to do clustering on it.
- **PCA+KSVD:** Use Principle Component Analysis (PCA) method to obtain Projection Subspace and then use K-SVD method (Aharon, Elad, and Bruckstein 2006) to do dictionary learning on it.
- **PCA+ODL:** Use Principle Component Analysis (PCA) method to obtain Projection Subspace and then use Online Dictionary Learning method (Mairal et al. 2009a) to do dictionary learning on it.
- **PCA+SSC:** Use Principle Component Analysis (PCA) method to obtain Projection Subspace and then use Semi-Supervised Clustering method (Basu, Banerjee, and Mooney 2004) with side information to do clustering on it.
- **PCA+Kmeans:** Use Principle Component Analysis (PCA) method to obtain Projection Subspace and then use K-means method to do clustering on it.

Table 1: Descriptions of the experimental data sets

Data Set	ORL	UMIST	PIE	FERET	JAFFE	COIL-20	ETH-80
# Number of Samples (n)	400	575	1632	1400	213	1440	656
# Input Dimensionality (d)	10304	10304	4096	6400	16384	16384	16384
# Number of Clusters (c)	40	20	68	200	10	20	8
# Reduced Dimensionality (k)	50	40	50	40	50	50	40
# Dimensionality of Dictionary (p)	80	60	90	50	60	60	50

Table 2: Clustering performances of the compared mehods measured by ACC (%)

	ORL	UMIST	PIE	FERET	JAFFE	COIL-20	ETH-80
Rand+KSVD	30.00	35.30	29.60	17.50	41.31	53.33	42.38
Rand+ODL	44.75	41.91	14.52	23.50	70.42	53.89	43.14
Rand+SSC	28.75	32.70	11.27	20.79	44.60	36.67	42.07
Rand+Kmeans	26.50	29.22	10.66	19.57	40.38	35.63	37.04
PCA+KSVD	63.25	48.17	54.47	25.07	58.22	60.56	46.04
PCA+ODL	58.00	43.30	22.12	24.26	82.63	60.69	44.36
PCA+SSC	70.00	47.30	17.16	28.00	84.51	66.39	47.87
PCA+Kmeans	69.75	44.00	16.73	27.21	82.16	64.10	41.62
ML+KSVD	43.75	39.13	54.96	25.21	38.03	52.08	19.21
Our Method	**71.75**	**55.48**	**57.84**	**32.14**	**90.61**	**67.01**	**49.06**

Table 3: Clustering performances of the compared mehods measured by NMI (%)

	ORL	UMIST	PIE	FERET	JAFFE	COIL-20	ETH-80
Rand+KSVD	54.36	43.37	49.50	60.88	58.83	62.43	30.43
Rand+ODL	64.57	56.47	41.17	64.32	69.35	62.99	36.69
Rand+SSC	54.33	41.33	32.52	64.08	44.38	45.32	35.47
Rand+Kmeans	50.94	40.50	30.77	63.74	42.58	45.18	35.05
PCA+KSVD	77.56	59.50	66.53	62.84	55.47	70.89	37.94
PCA+ODL	77.93	60.18	47.28	63.73	83.00	68.64	39.10
PCA+SSC	85.28	67.35	44.76	68.87	86.79	74.92	44.15
PCA+Kmeans	**86.25**	65.90	44.52	68.37	88.40	**78.97**	**46.86**
ML+KSVD	67.76	57.37	70.14	64.08	47.39	63.16	16.04
Our Method	84.97	**72.40**	**70.45**	**68.98**	**88.55**	72.83	42.89

- **ML+KSVD:** Use Metric Learning method (Xiang, Nie, and Zhang 2008) to obtain Projection Subspace and then use K-SVD method (Aharon, Elad, and Bruckstein 2006) to do dictionary learning on it.

Except **Rand+SSC**, **Rand+Kmeans**, **PCA+SSC** and **PCA+Kmeans**, other methods all need K-means as the postprocessing step to get the clustering indicator. We perform 50 times K-means for each method and choose the best result. Besides, we resort to clustering accuracy (ACC) and normalized mutual information (NMI) as the metric, which are defined as follows.

- Clustering Accuracy (ACC) is defined as follow:

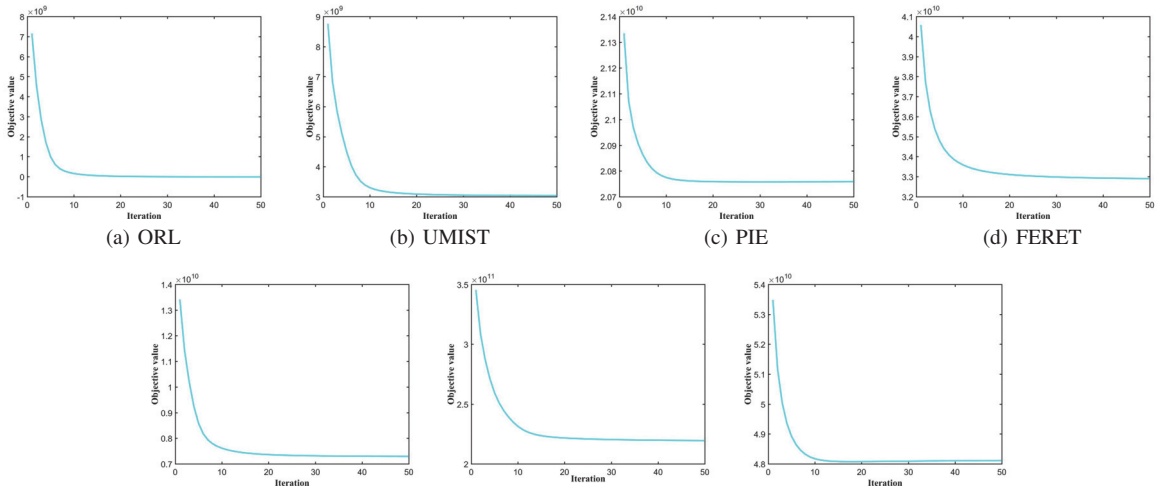
$$ACC = \frac{\sum_{i=1}^{n} \delta(pred_i, y_i)}{n} \tag{15}$$

where $pred_i$ is the clustering label and y_i is the true label. $\delta(x, y) = 1$ if $x = y$, otherwise $\delta(x, y) = 0$. Larger ACC indicates a better clustering result.

- Normalized Mutual Information (NMI) is defined as

$$NMI = \frac{I(Pred, Y)}{(H(Pred) + H(Y))/2} \tag{16}$$

where $I(Pred, Y)$ is the mutual information between the predicted clustering label $Pred$ and the true label



(a) ORL (b) UMIST (c) PIE (d) FERET

(e) JAFFE (f) COIL-20 (g) ETH-80

Figure 3: The objective value of our method running on benchmark data sets

Y. $H(\cdot)$ denotes the entropy function. Larger NMI indicates a better clustering result.

Experiment Analysis

The clustering performance of all the methods are reported in Table 2 and Table 3, from which we have the following interesting observations.

First, our method has better performance than all other compared dictionary learning methods, which demonstrates that our method is able to make full use of the side information and learn a good dictionary in optimal metric space. Without the use of side information, traditional dictionary learning method cannot obtain an optimal dictionary. Our unified model integrates metric learning into dictionary learning to learn an optimal metric space from the side information and learn a good dictionary on it. This novel model has better performance than traditional dictionary learning method without the use of the side information. And the learned metric space can improve the quality of dictionary.

Second, our method is consistently better than **ML+KSVD**, which confirms that our proposed unified model is effective and has better performance than the simple two-step method. The simple two-step method which use metric learning as a preprocessing can only learn a suboptimal metric space rather global optimal. The unified model can learn metric space W and dictionary D iteratively and alternatively to confirm that the learned metric space is optimal.

Third, our method performs better than normal clustering methods, which demonstrates that our method takes full advantage of the side information better and it can help learn a good dictionary and sparse representation for clustering.

Figure 3 shows the objective value of our method running on these benchmark data sets. As shown in Figure 3, our method can converge fast when running on these data sets, which confirms the efficiency of our proposed optimization algorithm.

Conclusion

In this paper, we designed a novel unified dictionary learning model to do dictionary learning in optimal metric space. One important advantage of our method is integrating metric learning to enhance dictionary learning with fully utilizing of the side information. And our proposed unified model updates metric space W and dictionary D alternatively and iteratively to confirm that both metric space and dictionary are optimal. Besides, the novel model can handle well high-dimensional data. The proposed method is evaluated on clustering tasks in real-world image data sets. The experimental results demonstrated that our model outperforms the other compared methods.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61572388, Grant 61402026 and Grant 61703327, and in part by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2017ZDCXL-GY-05-04-02 and Grant 2017ZDCXL-GY-05-04-02.

References

Aharon, M.; Elad, M.; and Bruckstein, A. 2006. rmk-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing* 54(11):4311–4322.

Bar-Hillel, A.; Hertz, T.; Shental, N.; and Weinshall, D. 2003. Learning distance functions using equivalence relations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 11–18.

Basu, S.; Banerjee, A.; and Mooney, R. J. 2004. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM international conference on data mining*, 333–344. SIAM.

Hoi, S. C.; Liu, W.; Lyu, M. R.; and Ma, W.-Y. 2006. Learning distance metrics with contextual constraints for image retrieval. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, 2072–2078. IEEE.

Jiang, W., and Chung, F.-l. 2014. A trace ratio maximization approach to multiple kernel-based dimensionality reduction. *Neural Networks* 49:96–106.

Jiang, Z.; Lin, Z.; and Davis, L. S. 2011. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 1697–1704. IEEE.

Leibe, B., and Schiele, B. 2003. Analyzing appearance and contour based methods for object categorization. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, II–409. IEEE.

Lewicki, M. S., and Sejnowski, T. J. 2006. Learning overcomplete representations. *Learning* 12(2).

Liu, W., and Tsang, I. W. 2015. Large margin metric learning for multi-label prediction. In *AAAI*, 2800–2806.

Liu, W., and Tsang, I. W. 2017. Making decision trees feasible in ultrahigh feature and label dimensions. *Journal of Machine Learning Research* 18(81):1–36.

Lyons, M.; Akamatsu, S.; Kamachi, M.; and Gyoba, J. 1998. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 200–205. IEEE.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2009a. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, 689–696. ACM.

Mairal, J.; Ponce, J.; Sapiro, G.; Zisserman, A.; and Bach, F. R. 2009b. Supervised dictionary learning. In *Advances in neural information processing systems*, 1033–1040.

Mallat, S. 1999. *A wavelet tour of signal processing*. Academic press.

Manton, J. H. 2002. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing* 50(3):635–650.

Nene, S. A.; Nayar, S. K.; Murase, H.; et al. 1996. Columbia object image library (coil-20).

Olshausen, B. A., and Field, D. J. 1997. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research* 37(23):3311–3325.

Phillips, P. J.; Wechsler, H.; Huang, J.; and Rauss, P. J. 1998. The feret database and evaluation procedure for face-recognition algorithms. *Image and vision computing* 16(5):295–306.

Phillips, J.; Bruce, V.; and Soulie, F. F. 1998. Face recognition: From theory to applications.

Samaria, F. S., and Harter, A. C. 1994. Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, 138–142. IEEE.

Sim, T.; Baker, S.; and Bsat, M. 2002. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, 53–58. IEEE.

Toi, I., and Frossard, P. 2011. Dictionary learning. *IEEE Signal Processing Magazine* 28(2):27–38.

Wang, H.; Nie, F.; Cai, W.; and Huang, H. 2013. Semi-supervised robust dictionary learning via efficient l-norms minimization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1145–1152.

Wang, H.; Nie, F.; and Huang, H. 2013. Robust and discriminative self-taught learning. In *International Conference on Machine Learning*, 298–306.

Xiang, S.; Nie, F.; and Zhang, C. 2008. Learning a mahalanobis distance metric for data clustering and classification. *Pattern Recognition* 41(12):3600–3612.

Xing, E. P.; Jordan, M. I.; Russell, S. J.; and Ng, A. Y. 2003. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, 521–528.

Zhang, Q., and Li, B. 2010. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2691–2698. IEEE.