Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Semi-Supervised Dictionary Learning via Structural Sparse Preserving

Di Wang, Xiaoqin Zhang*, Mingyu Fan and **Xiuzi Ye**
College of Mathematics & Information Science, Wenzhou University
Zhejiang, China
wangdi@amss.ac.cn, zhangxiaoqinnan@gmail.com, {fanmingyu, yexiuzi}@wzu.edu.cn

Abstract

While recent techniques for discriminative dictionary learning have attained promising results on the classification tasks, their performance is highly dependent on the number of labeled samples available for training. However, labeling samples is expensive and time consuming due to the significant human effort involved. In this paper, we present a novel semi-supervised dictionary learning method which utilizes the structural sparse relationships between the labeled and unlabeled samples. Specifically, by connecting the sparse reconstruction coefficients on both the original samples and dictionary, the unlabeled samples can be automatically grouped to the different labeled samples, and the grouped samples share a small number of atoms in the dictionary via mixed $\ell_{2,p}$-norm regularization. This makes the learned dictionary more representative and discriminative since the shared atoms are learned by using the labeled and unlabeled samples potentially from the same class. Minimizing the derived objective function is a challenging task because it is non-convex and highly non-smooth. We propose an efficient optimization algorithm to solve the problem based on the block coordinate descent method. Moreover, we have a rigorous proof of the convergence of the algorithm. Extensive experiments are presented to show the superior performance of our method in classification applications.

Introduction

In the recent decade, Compressed Sensing (CS) and its related works have led to state-of-the-art results in image analysis, such as image denoising (Protter and Elad 2009; Zhang et al. 2014), image restoration (Mairal, Elad, and Sapiro 2008; Mairal et al. 2009a), image alignment (Zhang et al. 2013) and image classification (Wright et al. 2009; Harandi et al. 2013). The success is partly owes to the fact that many natural images are sparse or compressible in the sense that they can be coded by a few of atoms in some dictionaries. Learning the dictionary is critical for the performance of sparse coding. Wright et al. (Wright et al. 2009) directly use the entire set of training samples as the dictionary for sparse coding, and achieve impressive performance on face recognition. However, due to the uncertain and noisy information in the original training images, it may not be effective enough to fully exploit the discriminative information hidden in the training samples.

To exploit the discriminative ability of the dictionary, Supervised Dictionary Learning (SDL) for classification tasks has gained a lot of attention. The dictionaries are learned by optimizing a unified objective function combining reconstructive and discriminative terms(Mairal et al. 2008; 2009b; Zhang and Li 2010; Yang et al. 2011; Jiang, Lin, and Davis 2011; 2013; Mairal, Bach, and Ponce 2012; Shen et al. 2013; 2015). Different from above works, some researchers make use of the structural sparsity on the coefficient matrix via mix regularization. This exploits the fact that once an atom of a dictionary has been selected to represent a samples, it may as well be used to represent other samples of the same class. Mixed regularization can promotes the use of a small subset of atoms for each class. Thus, sharing of dictionary atoms for data in the same class could increase the discriminative power of the dictionary (Bengio et al. 2009; Chi et al. 2013b; 2013a). Bengio et al. (Bengio et al. 2009) propose the method of group sparse coding by using the same dictionary words for all the images in a class, which provides a discriminative signal in the construction of image representations. Chi et al. (Chi et al. 2013b) propose an intra-block coherence suppression dictionary learning algorithm by employing the block and group regularized sparse modeling. They also present a novel affine-constrained group sparse coding framework (Chi et al. 2013a) to extend the current sparse representation-based classification (SRC) framework for classification problems with multiple inputs.

The performance of the SDL methods is highly dependent on the number of labeled training samples. Insufficient labeled training samples yield a dictionary with potentially bad generalization power. However, labeling samples is expensive and time consuming due to the significant human effort involved. On the other hand, one can easily obtain large amounts of unlabeled samples from public datasets. This has motivated researchers to develop semi-supervised algorithms for learning a better dictionary. Shrivastava et al. (Shrivastava et al. 2012) propose a Semi-Supervised Discriminative Dictionary (S2D2) learning algorithm for classification tasks, which iteratively estimates the confidence matrix of unlabeled samples and uses it to refine the learned dictionaries. In (Zhang, Jiang, and Davis 2012), it proposes

*Corresponding author

an online semi-supervised dictionary learning (OSSDL) algorithm, which integrates the reconstruction error of labeled and unlabeled data, the discriminative sparse-code error, and the classification error into an objective function to enhance the dictionary's representative and discriminative power. One major drawback of the above approaches is that the training samples in one class are used for computing the atoms in the dictionary, irrespective the training samples from other classes. More importantly, these approaches may lead to a large dictionary, as the size of the composed dictionary grows linearly with the number of classes. Babagholami-Mohamadabadi et al. (Babagholami-Mohamadabadi et al. 2013) propose a probabilistic semi-supervised dictionary learning method by introducing a discrimination term based on Local Fisher Discriminant Analysis(LFDA) and Locally Linear Embedding (LLE)(Roweis and Saul 2000). Wang et al. (Wang et al. 2013) propose a novel Semi-Supervised Robust Dictionary (SSR-D) learning method by exploiting the structural sparse regularization of labeled and unlabeled samples. In the training process, the algorithm can automatically select prominent dictionary atoms, such that the optimal dictionary size is learned from input data. However, it has the following two problems: (1) They impose $\ell_{2,0+}$ regularization to the sparse coefficient matrix corresponding to unlabeled samples which enforces all unlabeled samples to share a small subset of dictionary. This is unreasonable because the unlabeled samples can be potentially from different classes. (2) The underlying structural relationships between the labeled and unlabeled data are not exploited which are useful for classification.

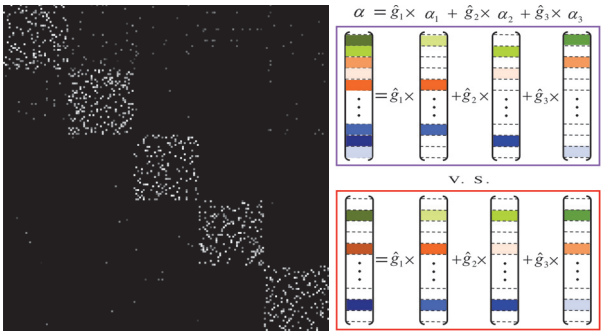
Inspired by the foregoing discussion, we propose a novel semi-supervised Dictionary Learning algorithm based on the Structural Sparse Preserving (SSP-DL). In detail, by connecting the sparse reconstruction coefficients on both the original samples and dictionary, the unlabeled samples can be automatically grouped to the different labeled samples, and the grouped samples are enforced to share a small number of atoms in the dictionary learning process. The main features of our work are as follows: (1) the dictionary are learned from both the labeled and unlabeled samples potentially from the same class, which is more representative and discriminative; (2) a compact and size-free dictionary is obtained because of the share mechanism of the grouped samples; (3) an efficient and non-trial optimization algorithm is presented to learn the dictionary with the convergence guaranteed.

Motivation

The key point of semi-supervised dictionary learning is to effectively use the label information of the labeled samples and the underlying structural relationships between the labeled and unlabeled samples. Consider the following sparse representation problem

$$\min_{G} \|G\|_{p,p}^{p} \quad \text{s.t. } X = XG,\ g_{ii} = 0,\ i = 1, \cdots, u+l \quad (1)$$

where X is the column-wised training data matrix, and G is the representing coefficient matrix, and u and l are the numbers of unlabeled and labeled samples respectively.



(a) (b)

Figure 1: Illustrations. (a) Visualization of a sub-matrix of $\hat{G}$ computed by the ℓ_p-optimization problem (1) on the dataset COIL-20. (b) The coefficients relationships (the coefficients corresponding to the color locations are non-zero).

$\|G\|_{p,p}$ is the $\ell_{p,p}$-norm ($0 < p < 1$). Here, we prefer $\ell_{p,p}$-norm to ℓ_1-norm since $\ell_{p,p}$-norm serve as a better alternative to ℓ_1-norm (Chen, Xu, and Ye 2010; Lyu et al. 2013; Wang et al. 2013). The optimal solution of problem (1) is denoted by $\hat{G} = [\hat{\boldsymbol{g}}_1, \cdots, \hat{\boldsymbol{g}}_{u+l}]$, where $\hat{\boldsymbol{g}}_i$ is the sparse representation of x_i under other training samples. Fig. 1(a) is the visualization of a sub-matrix of $\hat{G}$ on dataset COIL-20, which shows that $\hat{G}$ is very close to be block-diagonal. That means, if an unlabeled sample $\boldsymbol{x}_i$ is sparsely represented by the other samples as $\boldsymbol{x}_i = \sum_{k=1}^{\kappa} \hat{g}_{j_k i} \boldsymbol{x}_{j_k}$, where κ is a very small number and $\hat{g}_{j_k i} (k = 1, \cdots, \kappa)$ are the non-zero entries of the sparse coefficient vector $\hat{\boldsymbol{g}}_i$, then the samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}, \cdots, \boldsymbol{x}_{j_\kappa}$ are likely from the same class. The motivation of our work is how to effectively utilize this discriminative information contained in $\hat{G}$ for the semi-supervised dictionary learning.

Suppose that $\boldsymbol{\alpha}_{j_1}, \boldsymbol{\alpha}_{j_2}, \cdots, \boldsymbol{\alpha}_{j_\kappa}$ are the sparse codings of $\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}, \cdots, \boldsymbol{x}_{j_\kappa}$ under a given dictionary $D \in R^{d \times m}$ ($\boldsymbol{x}_{j_k} = D\boldsymbol{\alpha}_{j_k}$), where d is the dimension of samples and m is the number of atoms, thus we have

$$\boldsymbol{x}_i = \sum_{k=1}^{\kappa} \hat{g}_{j_k i} \boldsymbol{x}_{j_k} = \sum_{k=1}^{\kappa} D\left(\hat{g}_{j_k i} \boldsymbol{\alpha}_{j_k}\right) = D\left(\sum_{k=1}^{\kappa} \hat{g}_{j_k} \boldsymbol{\alpha}_{j_k}\right). \quad (2)$$

Let $\boldsymbol{\alpha}_i = \sum_{k=1}^{\kappa} \hat{g}_{j_k i} \boldsymbol{\alpha}_{j_k}$, then $\boldsymbol{\alpha}_i$ is a coefficient of $\boldsymbol{x}_i$ under the dictionary D. As shown in the top row of Fig. 1(b), generally, the coefficient $\boldsymbol{\alpha}_i$ can not be guaranteed to be sparse although all the parameters $\boldsymbol{\alpha}_{j_1}, \boldsymbol{\alpha}_{j_2}, \cdots, \boldsymbol{\alpha}_{j_\kappa}$ are sparse. But if we add the sparse constraint to $\boldsymbol{\alpha}_i$, then the equation $\boldsymbol{\alpha}_i = \sum_{k=1}^{\kappa} (\hat{g}_{j_k i} \boldsymbol{\alpha}_{j_k})$ will enforce the nonzero elements appear in the same location of $\boldsymbol{\alpha}_{j_1}, \boldsymbol{\alpha}_{j_2}, \cdots, \boldsymbol{\alpha}_{j_\kappa}$, which is illustrated by the bottom row of Fig. 1(b). In other words, the unlabeled and labeled samples potentially from the same class should share a small subset of D's atoms, since the samples from the same class most likely have the linear relationships $\boldsymbol{\alpha}_i = \sum_{k=1}^{\kappa} (\hat{g}_{j_k i} \boldsymbol{\alpha}_{j_k})$. The above relationship between the labeled and unlabeled samples is called *structural sparsity* in the following sections.

Therefore, as shown in Fig. 2, by using the structural

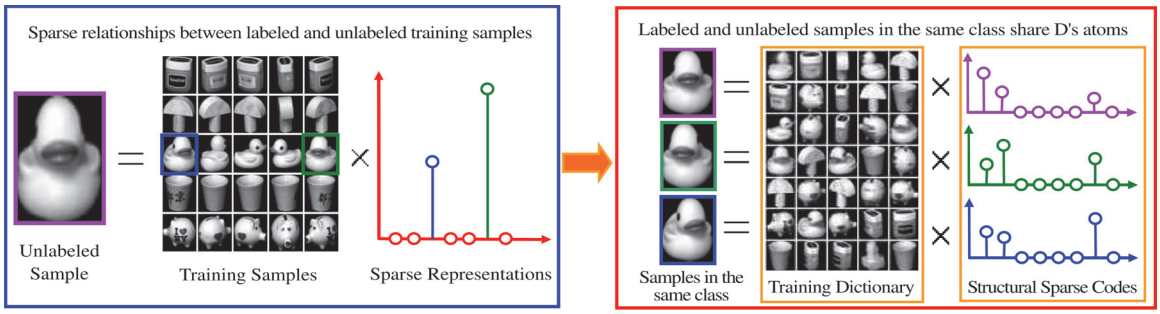


Figure 2: Dictionary learning based on the structural sparse relationships between labeled and unlabeled samples.

sparse relationships, the unlabeled samples can be automatically grouped to the different labeled samples, and the grouped samples are enforced to share a small number of atoms in the dictionary learning process. The advantages of this procedure are two-fold: (1) the dictionary is learned from both the labeled and unlabeled samples potentially from the same class, which is more representative and discriminative; (2) a compact and size-free dictionary is obtained because of the share mechanism of the grouped samples.

Proposed Semi-Supervised Dictionary Learning Method

Based on the above motivation, we propose a novel method for dictionary learning which utilizes the underlying structural sparse relationships between the labeled and unlabeled samples. In this section, we first present the formulation of the proposed semi-supervised dictionary method, and then show how the optimization problem is solved with convergence guaranteed. Finally, we present how the learned dictionary is applied for classification.

Problem Formulation

We first introduce the notations used in the remaining parts of the paper. Matrices are written as uppercase letters while vectors are written as boldface lowercase letters. Given a real $m \times n$ matrix $A = (\alpha_{ij})_{m\times n}$, $\boldsymbol{\alpha}^i \in R^n (i = 1, \cdots, m)$ and $\boldsymbol{\alpha}_j \in R^m (j = 1, \cdots, n)$ are respectively the i-th row and j-th column vectors of A. The Frobenius norm of the matrix A is denoted as $\|A\|_F$. The $\ell_{p,p}$-norm and the mixed $\ell_{2,p}$-norm of A are defined as $\|A\|_{p,p} = \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^p\right)^{\frac{1}{p}}$ and $\|A\|_{2,p} = \left(\sum_{i=1}^m \|\boldsymbol{\alpha}^i\|_2^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^m \left(\sqrt{\sum_{j=1}^n a_{ij}^2}\right)^p\right)^{\frac{1}{p}}$ respectively, where $0 < p < 1$.

Given a classification task with K class, denote by $X = [X_0, X_1, \cdots, X_K] \in R^{d\times(u+l)}$ the matrix of all training samples. $X_c \in R^{d\times n_c}$ is the matrix of the c-th class consisting of n_c training samples such that $\sum_{c=1}^K n_c = l$, and $X_0 \in R^{d\times u}$ is the unlabeled data matrix. Denote by A the sparse coding matrix of X over the dictionary D. Note that A can be rewritten as $A = [A_0, A_1, \cdots, A_K]$, where $A_c (c = 1, \cdots, K)$ is the sparse coding matrix for the samples belonging to the c-th class and A_0 is that for the unlabeled samples. Thus, the objective function for our dictionary learning is defined as

$$\begin{aligned}\langle \hat{D}, \hat{A}\rangle = \ & \arg\min_{D\in\mathcal{C}, A} \frac{1}{2}\|X - DA\|_F^2 + \lambda_1 \sum_{c=1}^K \|A_c\|_{2,p}^p \\ & + \lambda_2 \|A_0\|_{p,p}^p + \frac{\lambda_3}{2}\|A - A\hat{G}^\top\|_F^2 \end{aligned} \quad (3)$$

where $\mathcal{C} = \{D \in R^{d\times m}, \text{s.t. } \boldsymbol{d}_j^\top \boldsymbol{d}_j \le 1,\ \forall j = 1, \cdots, m\}$. In problem (3), $\|X - DA\|_F^2$ represents the reconstruction error. $\|A - A\hat{G}^\top\|_F^2$ measures the error in preserving the sparse relationships between labeled and unlabeled samples, which is followed from the equation $\boldsymbol{\alpha}_i = \sum_{k=1}^{\kappa} (\hat{g}_{j_k i} \boldsymbol{\alpha}_{j_k})$. As illustrated in Fig. 1(b), the minimization of $\|A_0\|_{p,p}^p$ and $\|A - A\hat{G}^\top\|_F^2$ enforces the unlabeled and labeled samples from the same class share a small number of D's atoms in the learning process. $\|A_c\|_{2,p}^p$ is the mixed $\ell_{2,p}$-norm regularization for class c. It is the supervised term and indicates which atoms in D should be shared for class c.

Optimization Procedure

Solving (3) is a challenging task because the objective function is non-convex and highly non-smooth. In the following, we will iteratively optimize A_0, A_c and D based on the block coordinate descent (BCD) method.

Let $W = I - \hat{G}$, and denote by $W = [W_0, W_1, \cdots, W_K]$ where W_c are the sub-matrices of W corresponding to $X_c (c = 0, 1, \cdots, K)$. Thus, (3) can be written as

$$\begin{aligned}\min_{D\in\mathcal{C}, A} \ & \frac{1}{2}\|X - DA\|_F^2 + \lambda_1 \sum_{c=1}^K \|A_c\|_{2,p}^p + \lambda_2 \|A_0\|_{p,p}^p \\ & + \frac{\lambda_3}{2}\left\|A_0 W_0^\top + \sum_{c=1}^K A_c W_c^\top\right\|_F^2 . \end{aligned} \quad (4)$$

Updating Sparse Codes A_0 We fixed D and $A_c (c = 1, \cdots, K)$ and rewrite the symbols A_0, X_0 and W_0 as A, X and W respectively for convenience, then the optimization problem for A_0 can be formulated as

$$\min_A I_1(A) = \frac{1}{2}\|X - DA\|_F^2 + \lambda_2 \|A\|_{p,p}^p + \frac{\lambda_3}{2}\left\|AW^\top + Q\right\|_F^2 \quad (5)$$

where $Q = \sum_{c=1}^{K} A_c W_c^\top$. Based on the majorization-minimization (MM) technique (Toh and S.Yun 2010; Oliveira, Bioucas-Dias, and Figueiredo 2009), we use the optimization results in Theorem 1 in (Marjanovic and Solo 2012) to derive a MM-based algorithm for iteratively reducing I_1. For any given $A^{(k)}$, we introduce the following candidate majorizer of I_1.

$$
\begin{aligned}
& J_1(A, A^{(k)}) \\
\triangleq\ & \frac{1}{2}\|X - DA^{(k)}\|_F^2 + \langle D^\top(DA^{(k)} - X), A - A^{(k)}\rangle \\
& + \frac{\eta_1}{2}\|A - A^{(k)}\|_F^2 + \lambda_2\|A\|_{p,p}^p + \lambda_3\left(\frac{1}{2}\|A^{(k)}W^\top + Q\|_F^2\right. \\
& \left. + \langle (A^{(k)}W^\top + Q)W, A - A^{(k)}\rangle + \frac{\eta_3}{2}\|A - A^{(k)}\|_F^2\right) \quad (6) \\
=\ & \frac{\eta_1 + \eta_3\lambda_3}{2}\left\|A - Z^{(k)}\right\|_F^2 + \lambda_2\|A\|_{p,p}^p + \text{Constant} \quad (7)
\end{aligned}
$$

where

$$
Z^{(k)} = A^{(k)} - \frac{1}{\eta_1 + \eta_3\lambda_3}\left(D^\top(DA^{(k)} - X) + \lambda_3(A^{(k)}W^\top + Q)W\right). \quad (8)
$$

Thus, we have

$$
\begin{aligned}
& \arg\min_A J_1(A, A^{(k)}) = \arg\min_A \frac{1}{2}\|A - Z^{(k)}\|_F^2 + \lambda\|A\|_{p,p}^p \\
=\ & \arg\min_A \sum_{i=1}^{m}\sum_{j=1}^{u}\left\{\frac{1}{2}\left(a_{ij} - z_{ij}^{(k)}\right)^2 + \lambda|a_{ij}|^p\right\} \quad (9)
\end{aligned}
$$

where $\lambda = \frac{\lambda_2}{\eta_1 + \eta_3\lambda_3}$. Using the optimization result in Theorem 1 in (Marjanovic and Solo 2012), the solution for the problem (9) is

$$
\hat{A} = [\hat{a}_{ij}]_{m\times u} = \left[\mathcal{T}_\lambda\left(z_{ij}^{(k)}\right)\right]_{m\times u} \quad (10)
$$

where

$$
\mathcal{T}_\lambda(z) = \begin{cases} 0 & \text{if } |z| < \tau \\ \{0, \text{sgn}(z)\hat{a}\} & \text{if } |z| = \tau \\ \text{sgn}(z)\hat{a}^* & \text{if } |z| > \tau \end{cases}. \quad (11)
$$

In (11), $\hat{a} = [2\lambda(1-p)]^{\frac{1}{2-p}}$, $\tau = \hat{a} + \lambda p\hat{a}^{p-1}$. $\hat{a}^* \in (\hat{a}, |h_{ij}|)$ is the lager solution of

$$
a + \lambda p a^{p-1} = |z|, \text{ where } a > 0 \quad (12)
$$

which can be obtained from the following iteration

$$
a_{(t+1)} = |z| - \lambda p a_{(t)}^{p-1} \quad (13)
$$

with the initial value $a_{(0)} \in (\hat{a}, |h_{ij}|)$.

Propostion 1. *If $\eta_1 \geq \lambda_{max}(D^\top D)$ and $\eta_3 \geq \lambda_{max}(W^\top W)$, then we have*

$$
I_1(A) \leq J_1(A, A^{(k)}), \ \textit{for}\ \forall A, A^{(k)}\ . \quad (14)
$$

Let

$$
A^{(k+1)} = \arg\min_A J_1(A, A^{(k)}) \quad (15)
$$

and combine Proposition 1 and the fact $I_1(A^{(k)}) = J_1(A^{(k)}, A^{(k)})$, we have

$$
I_1(A^{(k+1)}) \leq J_1(A^{(k+1)}, A^{(k)}) \leq I_1(A^{(k)})\ . \quad (16)
$$

Motivated by Eq. (16), we propose an iterative scheme to obtain the solution of the objective function I_1, which is described in Algorithm 1.

Algorithm 1: Updating Sparse Codes A_0

Input: Given an integer ι and a real number ϵ.

1: Initialize $A^{(0)}$, $dif = \inf$, and let $k = 0$.
2: **while** $k < \iota$ & $dif > \epsilon$ **do**
3: Compute $Z^{(k)}$ via (8).
4: Compute $A^{(k+1)} = \arg\min_A J_1(A, A^{(k)})$ via (10).
5: Let $dif = |J_1(A^{(k+1)}, A^{(k)}) - J_1(A^{(k)}, A^{(k)})|$, and $k = k + 1$.
6: **end while**

Output: The sparse codes $A_0 = A^{(k)}$.

Updating Sparse Codes A_c Without loss of generality, we fixed D, A_0 and $A_1, \cdots, A_{c-1}, A_{c+1}, \cdots, A_K$ and rewritten A_c, X_c and W_c as A, X and W respectively for convenience, then the optimization problem for A_c can be formulated as

$$
\min_A I_2(A) = \frac{1}{2}\|X - DA\|_F^2 + \lambda_1\|A\|_{2,p}^p + \frac{\lambda_3}{2}\left\|AW^\top + Q\right\|_F^2 \quad (17)
$$

where $Q = \sum_{i\neq c} A_i W_i^\top$. Similar as J_1, we also introduce the candidate majorizer of I_2 as follows.

$$
\begin{aligned}
& J_2(A, A^{(k)}) \\
\triangleq\ & \frac{1}{2}\|X - DA^{(k)}\|_F^2 + \langle D^\top(DA^{(k)} - X), A - A^{(k)}\rangle \\
& + \frac{\eta_1}{2}\|A - A^{(k)}\|_F^2 + \lambda_1\|A\|_{2,p}^p + \lambda_3\left(\frac{1}{2}\|A^{(k)}W^\top + Q\|_F^2\right. \\
& \left. + \langle (A^{(k)}W^\top + Q)W, A - A^{(k)}\rangle + \frac{\eta_3}{2}\|A - A^{(k)}\|_F^2\right) \quad (18) \\
=\ & \frac{\eta_1 + \eta_3\lambda_3}{2}\left\|A - Z^{(k)}\right\|_F^2 + \lambda_1\|A\|_{2,p}^p + \text{Constant} \quad (19)
\end{aligned}
$$

where $Z^{(k)}$ is defined by (8). Then

$$
\arg\min_A J_2(A, A^{(k)}) = \arg\min_A \frac{1}{2}\|A - Z^{(k)}\|_F^2 + \lambda\|A\|_{2,p}^p \quad (20)
$$

where $\lambda = \frac{\lambda_1}{\eta_1 + \eta_3\lambda_3}$. Before solving problem (20), we give the following non-trivial $\ell_{2,p}$ penalized least square problem which is closely related to (20)

$$
\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{z}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^p \quad (21)
$$

where $\boldsymbol{z}$ is a vector constant. The solution of problem (21) can be obtained from the following proposition.

Propostion 2. *Let $\delta = \frac{\lambda}{\|\boldsymbol{z}\|_2^{2-p}}$, then the solution $\hat{\boldsymbol{\alpha}}$ to the problem (21) is*

$$
\hat{\boldsymbol{\alpha}} = \mathcal{T}_\delta(1) * \boldsymbol{z}\ . \quad (22)
$$

According to Proposition 2, problem (20) can be solved by

$$
\hat{A} = \begin{pmatrix} \mathcal{T}_{\delta_1}(1) & 0 & \cdots & 0 \\ 0 & \mathcal{T}_{\delta_2}(1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathcal{T}_{\delta_m}(1) \end{pmatrix} * Z^{(k)} \quad (23)
$$

Algorithm 2: The optimization procedure for the objective function (3).

Input: The data matrix of training samples $X = [X_0, X_1, \cdots, X_K] \in R^{d\times(u+l)}$.

1: Compute the sparse representation matrix $\hat{G}$ via problem (1), and let $W = I - \hat{G}$.
2: Initialize the dictionary $D^{(0)}$, and sparse codings $A^{(0)} = [A_0^{(0)}, A_1^{(0)}, \cdots, A_K^{(0)}]$. Let $t = 0$.
3: Repeat for fixed number of iterations (or until convergence):
4: **loop**
5: Let $Q = \sum_{c=1}^{K} A_c^{(t)} W_c^\top$, and update coefficients $A_0^{(t+1)} = \arg\min_A I_1(A)$ via Algorithm 1.
6: **for** $c = 1, \ldots, K$ **do**
7: Let $Q = \sum_{i=0}^{c-1} A_i^{(t+1)} W_i^\top + \sum_{i=c+1}^{K} A_i^{(t)} W_i^\top$.
8: Update coefficients $A_c^{(t+1)} = \arg\min_A I_2(A)$ via the similar procedure of Algorithm 1.
9: **end for**
10: Update the dictionary $D^{(t+1)}$ by using the optimization process in Eq. (26).
11: Let $t = t + 1$.
12: **end loop**

Output: The dictionary $\hat{D} = D^{(T)}$ and sparse codings $\hat{A} = [\hat{A}_0, \hat{A}_1, \cdots, \hat{A}_K] = [A_0^{(T)}, A_1^{(T)}, \cdots, A_K^{(T)}]$, where T is the iteration number at which the learning algorithm converges.

where $\delta_r = \frac{\lambda}{\|\boldsymbol{z}_{(k)}^r\|_2^{2-p}}$, and $\boldsymbol{z}_{(k)}^r$ is the r-th row of $Z^{(k)}$ ($r = 1, \cdots, m$). Similar as the relationship between I_1 and J_1, we also have

$$I_2(A^{(k+1)}) \leq J_2(A^{(k+1)}, A^{(k)}) \leq I_2(A^{(k)}) \,. \tag{24}$$

The algorithm for solving problem (17) is similar with Algorithm 1 and is skipped due to space limit.

Updating Dictionary D With fixed A (including A_0 and A_c), the optimization problem for D can be formulated as

$$\min_{D\in\mathcal{C}} I_3(D) = \frac{1}{2}\|X - DA\|_F^2 \,. \tag{25}$$

Denote by $I_{\mathcal{C}}$ the indicator of $\mathcal{C}$, and introduce the auxiliary variable H, then the problem (25) changes to

$$\begin{aligned} \min_{D,H} \quad & \frac{1}{2}\|X - DA\|_F^2 + I_{\mathcal{C}}(H) \\ \text{s.t.} \quad & D = H \end{aligned}$$

A typical iterative process based on the alternating direction method of multipliers(ADMM) (Ren and Lin 2013; Lin, Liu, and Li 2015) for computing D can be written explicitly as

$$\begin{cases} D^{(k+1)} := (XA^\top - Y^{(k)} + \mu * H^{(k)})(AA^\top + \mu I)^{-1} \\ H^{(k+1)} := \Pi_{\mathcal{C}}\left(D^{(k+1)} + Y^{(k)}/\mu\right) \\ Y^{(k+1)} := Y^{(k)} + \mu(D^{(k+1)} - H^{(k+1)}) \end{cases} \tag{26}$$

where Y is the Lagrange multiplier, μ is a positive scalar and $\Pi_{\mathcal{C}}$ is the projection operator on $\mathcal{C}$.

The optimization procedure of problem (3) is described in Algorithm 2. Note that in the training process, if the entries in some rows of A are all zeros or very small values, then the corresponding atoms in D are useless and can be deleted. Hence, we reconstruct the dictionary D by using all $\boldsymbol{d}_j$ in the following set as its columns:

$$\mathcal{D} = \{\boldsymbol{d}_j |\ \|\boldsymbol{\alpha}^j\|_2 > \varepsilon\} \tag{27}$$

where ε is a very small value. Therefore, the dictionary size can be automatically learned in the training process. The convergence of Algorithm 2 is guaranteed by the following theorem.

Theorem 1. *If the parameters satisfy $\eta_1 \geq \lambda_{max}(D^\top D)$ and $\eta_3 \geq \lambda_{max}(W^\top W)$ in the training process, Algorithm 2 decreases the objective value in (3) in each iteration.*

The detail proofs of **Proposition 1**, **Proposition 2**, and **Theorem 1** are skipped due to space limit and will be provided in the extended version of the paper.

Class Label Prediction

Once we obtain the discriminative dictionary $\hat{D}$ from Algorithm 2, we define the atoms set of the c-th class as

$$\mathcal{D}_c = \{\hat{\boldsymbol{d}}_j |\ \|\hat{\boldsymbol{\alpha}}_c^j\|_2 > 0\}, \quad c = 1, \cdots, K \tag{28}$$

where $\hat{\boldsymbol{\alpha}}_c^j$ is the j-th row of the c-th class sparse codings $\hat{A}_c$ and $\hat{\boldsymbol{d}}_j$ is the j-th atom of $\hat{D}$. Thus, the c-th class specific dictionary $\hat{D}_c \in R^{d\times|\mathcal{D}_c|}$ is constructed by using all $\hat{\boldsymbol{d}}_j \in \mathcal{D}_c$ as its columns.

After obtaining the specific dictionaries $\hat{D}_c (c = 1, \cdots, K)$, classifying an unlabeled sample $\boldsymbol{x}$ is performed by the following three steps:

Step 1: Compute the sparse codings of $\boldsymbol{x}$ over the c-th class specific dictionary $\hat{D}_c$, denoted by $\hat{\boldsymbol{\alpha}}_c (c = 1, \cdots, K)$, via

$$\hat{\boldsymbol{\alpha}}_c = \arg\min_{\boldsymbol{\alpha}_c} \|\boldsymbol{\alpha}_c\|_1, \ \text{ s.t. } \boldsymbol{x} = \hat{D}_c\boldsymbol{\alpha}_c \,. \tag{29}$$

Step 2: Compute the reconstruction error of $\boldsymbol{x}$ with respect to $D_c (c = 1, \cdots, K)$

$$e_c = \|\boldsymbol{x} - \hat{D}_c\hat{\boldsymbol{\alpha}}_c\|_2 \,. \tag{30}$$

Step 3: The predicted class label of the sample $\boldsymbol{x}$ is the one that minimizes the reconstruction error

$$y_{\boldsymbol{x}} = \arg\min_{c=1,\cdots,K} e_c \,. \tag{31}$$

Experimental Results

In this section, We first perform handwritten digit recognition on the two widely used datasets: MNIST (LeCun et al. 1998) and USPS (Hull 1994). And then, we apply the proposed algorithm to Face Recognition on the UMIST (Wechsler et al. 1998) face dataset. At last, we evaluate our approach on two public object datasets: SBData (Li and Allinson 2009) and COIL-20 (Nene, Nayar, and Murase 1996). An overall description of the data sets is presented

Table 1: Overall description of the datasets.

Datasets	DIM	Data#	Class#	τ	ν
MNIST	784	2000	10	20	80
USPS	256	1100	10	20	40
UMIST	750	564	20	5	10
SBData	638	3192	40	5	20
COIL-20	1521	1440	20	2∼10	33∼25

in Table 1. In order to clearly illustrate the advantage of the proposed method, we compare our method with SRC (Wright et al. 2009), three well-known SDL methods including Discriminative K-SVD (DKSVD) (Zhang and Li 2010), Fisher Discrimination Dictionary Learning (FDDL) (Yang et al. 2011) and Label Consistent K-SVD (LCKSVD) (Jiang, Lin, and Davis 2011), as well as three state of the art Semi-Supervised Dictionary Learning (SSDL) methods including OSSDL (Zhang et al. 2013), S2D2 (Shrivastava et al. 2012) and SSR-D (Wang et al. 2013). In the experiments, we use the whole image as the feature vector, and normalize the vector to have unit ℓ_2-norm. The parameters of all methods are obtained by using 5-fold cross validation. For each dataset X, we first rearrange the order of data samples randomly. Then, in each class of X, we randomly select τ samples as labeled samples, ν samples as unlabeled samples, and the rest are left for testing samples. Following the common evaluation procedure, we repeat the experiments 10 times with different random spits of the datasets to report the average classification accuracy together with standard deviation, and the best classification results are in boldface.

Experimental results

We present the recognition accuracies together with the standard deviation in Table 2, from which it is clear that the proposed method performs better than the other methods. SRC has the lowest accuracy because of the limited number of labeled samples. The SSDL methods always performs significantly better than the SDL methods (except for FDDL method). This is because the SDL methods cannot utilize the unlabeled samples for dictionary learning, and may overfit to the labeled samples when the number of labeled samples is small. Our method not only considers the unlabeled samples, but also consider the structural sparse relationships between labeled and unlabeled samples, which makes the unlabeled samples be automatically grouped to the different labeled samples in the training process. Hence, both SDL and SSDL methods are less accurate than our method.

To further demonstrate the effect of the number of labeled samples on our performance in comparison with others, we conduct the last experiment on the dataset COIL-20. For each object, we randomly select 35 samples as training samples, and the rest 37 are for testing. Out of the 35 training samples, we randomly use $2, 3, \cdots, 10$ to form the labeled samples respectively, and the rest as the unlabeled samples. The average accuracies together with the standard deviation are presented in Table 3. A first glance at the results in Table 3 show that, the improvement of SSDL methods compared to SDL methods is not too obvious when there

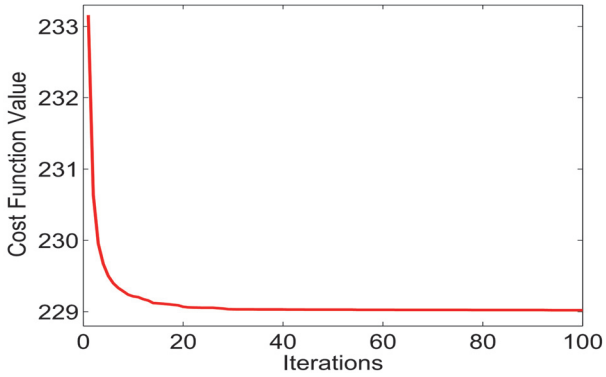


Figure 3: Relationship between the cost function values and the iterations on the USPS dataset.

are more labeled samples. However, the benefit of SSDL method can be significant when the labeled samples are few. This is the main focus of the SSDL methods. From Table 3, SSR-D and our method always have significant better results than the other two SSDL methods. This is because they use the structural sparsity on the coefficient matrix which promotes the share of the same dictionary atoms for the samples in the same class and yields some form of discrimination. Moreover, since we consider the structural sparse relationships of the training samples in the dictionary learning process, which makes the unlabeled samples be automatically grouped to the different labeled samples, our method performs better than SSR-D method.

Analysis of Optimization Process

To show the effectiveness of the proposed optimization method, we investigate and analyze the value of cost function in Eq. (3) in the optimization process. Here, we take the USPS dataset as an example, as shown in Fig. 3, we can see that the curve of cost function drops very quickly in the iteration process, and can achieve a satisfactory performance when the iteration number is 20. This curve greatly validates the claimed advantage of the optimization problem and verifies the Theorem 1.

Conclusion

By utilizing the structural sparse relationships between the labeled and unlabeled samples, we propose a novel semi-supervised method for learning a discriminative dictionary. Specifically, the sparse reconstruction coefficients on the original samples is preserved in the dictionary learning process. This makes the unlabeled samples be automatically grouped to the different labeled samples, and the grouped samples share the same dictionary via the mixed $\ell_{2,p}$-norm regularization. In this way, we enhance the representative and discriminative power of the dictionary since the shared atoms are learned from the labeled and unlabeled samples from the same class. Moreover, an efficient and non-trial optimization algorithm based on the block coordinate descent method is proposed to solve the highly non-smooth and

Table 2: Classification accuracy of different methods.

Datasets	SRC	DKSVD	FDDL	LCKSVD1	LCKSVD2	OSSDL	S2D2	SSR-D	SSP-DL
MNIST	17.30±0.8	71.44±1.7	82.52±1.3	72.95±1.3	72.95±1.3	73.18±1.8	77.58±0.8	83.83±1.2	**85.75±1.2**
USPS	30.00±0.9	67.50±1.8	85.22±1.2	76.91±1.5	76.92±1.3	80.78±2.8	86.57±1.6	87.20±0.5	**87.84±1.1**
UMIST	70.91±4.4	80.78±3.7	85.00±2.3	86.44±2.7	86.76±2.6	85.02±2.9	85.18±3.2	87.25±2.7	**88.73±2.5**
SBData	36.60±1.4	40.50±2.8	58.63±1.9	52.23±2.2	55.28±1.5	53.86±2.2	56.97±1.9	63.98±1.8	**64.33±1.4**

Table 3: Classification accuracy of different methods with changing the number of labeled samples on the COIL-20 dataset.

Labeled#	SRC	DKSVD	FDDL	LCKSVD1	LCKSVD2	OSSDL	S2D2	SSR-D	SSP-DL
$\tau = 2$	44.89±0.7	53.34±8.4	70.47±1.1	67.26±1.97	67.77±1.5	65.68±3.3	67.30±3.0	71.99±1.8	**74.47±1.7**
$\tau = 3$	45.05±2.4	67.16±5.6	76.69±1.9	71.66±3.2	72.74±2.9	72.16±3.0	73.14±2.5	78.78±1.9	**80.11±1.8**
$\tau = 4$	46.62±2.8	70.07±5.71	80.14±2.3	73.04±3.6	75.64±1.9	76.45±3.4	78.92±1.7	82.64±2.4	**83.99±1.7**
$\tau = 5$	48.70±1.0	69.16±4.2	82.64±2.4	76.96±3.6	78.07±2.3	79.26±2.9	80.84±2.2	83.72±2.4	**86.03±1.4**
$\tau = 6$	49.46±2.2	74.49±1.9	85.68±1.0	75.78±4.4	80.95±1.3	81.57±3.9	83.21±0.7	86.25±1.4	**88.35±1.1**
$\tau = 7$	51.81±1.3	78.07±2.8	87.64±0.7	72.94±2.1	82.70±0.8	82.21±3.3	84.86±1.6	88.55±1.0	**89.07±1.2**
$\tau = 8$	53.76±1.2	79.76±2.4	88.49±1.4	76.72±3.9	83.61±1.4	82.97±2.9	88.18±1.0	89.66±0.9	**90.47±0.9**
$\tau = 9$	53.83±1.8	83.24±2.1	90.72±1.2	76.18±0.9	85.81±1.3	84.28±2.2	88.31±1.1	**92.10±1.2**	91.48±0.7
$\tau = 10$	55.41±1.5	83.92±2.2	90.68±1.0	74.22±3.8	86.05±0.8	85.91±3.2	88.99±0.8	91.35±0.6	**91.81±0.6**

non-convex optimization problem. Experiments using various benchmark datasets demonstrate the superiority of the proposed method over the state-of-the-art SDL and SSDL methods. Possible future work includes the robustness of dictionary against outlier samples by replacing the Frobenius norm with the $\ell_{2,p}$-norm for the reconstruction error.

Acknowledgments

This work is supported by NSFC (Grant Nos. 61472285, 61511130084, 61473212, 61203241 and 61305035), Zhejiang Provincial Natural Science Foundation (Grants Nos. LY16F020023, LY15F030011 and LQ13F030009), Project of science and technology plans of Zhejiang Province (Grants Nos. 2014C31062, 2015C31168).

References

Babagholami-Mohamadabadi, B.; Zarghami, A.; Zolfaghari, M.; and Baghshah, M. S. 2013. Pssdl: Probabilistic semi-supervised dictionary learning. In *ECML/PKDD*, 192–207.

Bengio, S.; Pereira, F.; Singer, Y.; and Strelow, D. 2009. Group sparse coding. In *Advances in Neural Information Processing Systems*, 82–89.

Chen, X.; Xu, F.; and Ye, Y. 2010. Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization. *SIAM Journal on Scientific Computing* 32:2832–2852.

Chi, Y. T.; Ali, M.; Rajwade, A.; and Ho, J. 2013a. Block and group regularized sparse modeling for dictionary learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 377–382.

Chi, Y. T.; Ali, M.; Rushdi, M.; and Ho, J. 2013b. Affine-constrained group sparse coding and its application to image-based classifications. In *Proceedings of IEEE International Conference on Computer Vision*, 681–688.

Harandi, M.; Sanderson, C.; Shen, C.; and Lovell, B. 2013. Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution. In *Proceedings of IEEE International Conference on Computer Vision*, 3120–3127.

Hull, J. 1994. A database for handwritten test recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16:550–554.

Jiang, Z.; Lin, Z.; and Davis, L. 2011. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *Proceedings of IEEE International Conference on Computer Vision*, 1697–1704.

Jiang, Z.; Lin, Z.; and Davis, L. 2013. Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(11):2651–2664.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Li, J., and Allinson, N. M. 2009. Dimensionality reduction-based building recognition. In *Proceedings of IASTED International Conference on Visualization, Imaging and Image Processing*.

Lin, Z.; Liu, R.; and Li, H. 2015. Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning. *Machine Learning* 99(2):287–325.

Lyu, Q.; Lin, Z.; She, Y.; and Zhang, C. 2013. A comparison of typical ℓ_p minimization algorithms. *Neurocomputing* 119:413–424.

Mairal, J.; Bach, F.; and Ponce, J. 2012. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(4):791–804.

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; and Zisserman, A. 2008. Discriminative learned dictionaries for local image analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; and Zisserman, A.

2009a. Non-local sparse models for image restoration. In *Proceedings of IEEE International Conference on Computer Vision*, 2272–2279.

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; and Zisserman, A. 2009b. Supervised dictionary learning. In *Advances in Neural Information Processing Systems*, 1033–1040.

Mairal, J.; Elad, M.; and Sapiro, G. 2008. Sparse representation for color image restoration. *IEEE Transactions on Image Processing* 17(1):53–69.

Marjanovic, G., and Solo, V. 2012. On l_q optimization and matrix completion. *IEEE Transactions on Signal Processing* 60(11):5714–5724.

Nene, S. A.; Nayar, S. K.; and Murase, H. 1996. Columbia object image library (coil-20). Technical Report CUCS-006-96, Department of Computer Science, Columbia University, New York, N. Y.

Oliveira, J.; Bioucas-Dias, J.; and Figueiredo, M. 2009. Adaptive total variation image deblurring: A majorization-minimization approach. *Signal Processing* 89:1683–1693.

Protter, M., and Elad, M. 2009. Image sequence denoising via sparse and redundant representations. *IEEE Transactions on Image Processing* 18(1):27–35.

Ren, X., and Lin, Z. 2013. Linearized alternating direction method with adaptive penalty and warm starts for fast solving transform invariant low-rank textures. *International Journal of Computer Vision* 104(1):1–14.

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

Shen, L.; Wang, S.; Sun, G.; Jiang, S.; and Huang, Q. 2013. Multi-level discriminative dictionary learning towards hierarchical visual categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 383–390.

Shen, L.; Sun, G.; Huang, Q.; Wang, S.; Lin, Z.; and Wu, E. 2015. Multi-level discriminative dictionary learning with application to large scale image classification. *IEEE Transactions on Image Processing* 24(10):3109–3123.

Shrivastava, A.; Pillai, J. K.; Patel, V. M.; and Chellappa, R. 2012. Learning discriminative dictionaries with partially labeled data. In *Proceedings of IEEE Conference on Image Processing*, 3113–3116.

Toh, K., and S.Yun. 2010. An accelerated proximal gradient algorithmfor nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization* 6:615–640.

Wang, H.; Nie, F.; Cai, W.; and Huang, H. 2013. Semi-supervised robust dictionary learning via efficient $\ell_{2,0+}$-norms minimization. In *Proceedings of IEEE International Conference on Computer Vision*, 1145–1152.

Wechsler, H.; Phillips, P. J.; Bruce, V.; Fogelman-Soulie, F.; and Huang, T. S. 1998. *Face Recognition: From Theory to Applications*. Springer-Verlag Berlin Heidelberg.

Wright, J.; Yang, M.; Ganesh, A.; Sastry, S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2):210–227.

Yang, M.; Zhang, L.; Feng, X.; and Zhang, D. 2011. Fisher discrimination dictionary learning for sparse representation. In *Proceedings of IEEE International Conference on Computer Vision*, 543–550.

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

Zhang, X.; Wang, D.; Zhou, Z.; and Ma, Y. 2013. Simultaneous rectification and alignment via robust recovery of low-rank tensors. In *Advances in Neural Information Processing Systems*, 1637–1645.

Zhang, X.; Zhou, Z.; Wang, D.; and Ma, Y. 2014. Hybrid singular value thresholding for tensor completion. In *Proceedings of AAAI Conference on Artificial Intelligence*, 1362–1368.

Zhang, G.; Jiang, Z.; and Davis, L. 2012. Online semi-supervised disciminative dictionary learning for sparse representation. In *Proceedings of Asian Conference on Computer Vision*, 259–273.