Pattern Recognition 48 (2015) 3283–3292

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Non-linear dictionary learning with partially labeled data

Ashish Shrivastava *, Vishal M. Patel, Rama Chellappa

Center for Automation Research, UMIACS, University of Maryland, College Park, MD 20742, United States

ARTICLE INFO

Article history:
Received 2 January 2014
Received in revised form
24 July 2014
Accepted 31 July 2014
Available online 8 August 2014

Keywords:
Weakly supervised learning
Semi-supervised learning
Kernel methods
Dictionary learning
Classification

ABSTRACT

While recent techniques for discriminative dictionary learning have demonstrated tremendous success in image analysis applications, their performance is often limited by the amount of labeled data available for training. Even though labeling images is difficult, it is relatively easy to collect unlabeled images either by querying the web or from public datasets. Using the kernel method, we propose a non-linear discriminative dictionary learning technique which utilizes both labeled and unlabeled data for learning dictionaries in the high-dimensional feature space. Furthermore, we show how this method can be extended for ambiguously labeled classification problem where each training sample has multiple labels and only one of them is correct. Extensive evaluation on existing datasets demonstrates that the proposed method performs significantly better than state of the art dictionary learning approaches when unlabeled images are available for training.

1. Introduction

Sparse and redundant signal representations have recently gained much interest in computer vision field [34,12,27]. This is partly due to the fact that signals or images of interest are often sparse with respect to some dictionary. These dictionaries can be either analytic or they can be learned directly from the data. In fact, it has been observed that learning a dictionary directly from data often leads to improved results in many practical applications such as classification and restoration [34,22,6].

While dictionaries are often trained to obtain good reconstruction, training supervised dictionaries with a specific discriminative criterion has also been considered. For instance, linear discriminant analysis (LDA)-based basis selection and feature extraction algorithm for classification using wavelet packets was proposed by Etemand and Chellappa [14] in the late nineties. Recently, similar algorithms for simultaneous sparse signal representation and discrimination have also been proposed [25,15,24,38,17,18,36,21,37].

Sparse representation and dictionary learning methods for unsupervised learning have also been proposed. In [33], a method for simultaneously learning a set of dictionaries that optimally represent each cluster is proposed. To improve the accuracy of sparse coding, this approach was later extended by adding a block incoherence term in their optimization problem [23]. Some of the other sparsity motivated clustering and subspace clustering methods include [13,8].

The performance of a supervised classification algorithm is often dependent on the quality and diversity of training images, which are mainly hand-labeled. However, labeling images is expensive and time consuming due to the significant human effort involved. On the other hand, one can easily obtain large amounts of unlabeled images from public image datasets like Flickr or by querying image search engines like Bing. This has motivated researchers to develop semi-supervised algorithms, which utilize both labeled and unlabeled data for learning classifier models. Such methods have demonstrated improved performance when the amount of labeled data is limited. See [4] for an excellent survey of recent efforts on semi-supervised learning.

Two of the most popular methods for semi-supervised learning are Co-Training [2] and Semi-Supervised Support Vector Machines (S3VM) [32]. Co-Training assumes the presence of multiple views for each feature and uses the confident samples in one view to update the other. However, in applications such as image classification, one often has just a single feature vector and hence it is difficult to apply Co-Training. S3VM considers the labels of the unlabeled data as additional unknowns and jointly optimizes over the classifier parameters and the unknown labels in the SVM framework [3].

Using the kernel trick, several methods have been proposed in the literature that exploit sparsity of data in the high dimensional feature space. In these methods, a preselected Mercer kernel is used to map the input data onto a features space where dictionaries are trained. It has been shown that such non-linear dictionaries can provide better discrimination than their linear counterparts [20,30,19].

Motivated by the success of non-linear dictionary learning methods [20,30], we propose a novel method to learn kernel discriminative dictionaries for classification in a semi-supervised

* Corresponding author.
E-mail addresses: ashish@umiacs.umd.edu (A. Shrivastava), pvishalm@umiacs.umd.edu (V.M. Patel), rama@umiacs.umd.edu (R. Chellappa).

http://dx.doi.org/10.1016/j.patcog.2014.07.031

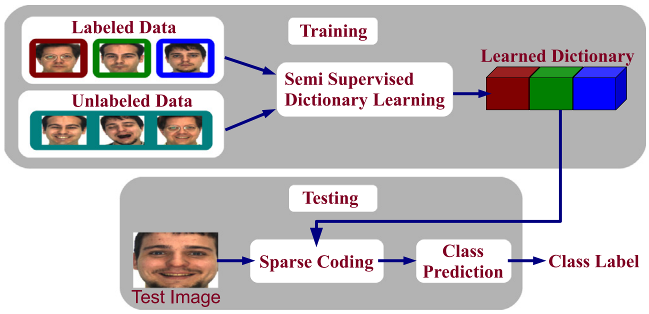

Fig. 1. Block diagram illustrating semi-supervised dictionary learning.

manner. Fig. 1 shows the block diagram of the proposed approach which uses both labeled and unlabeled data. While learning a dictionary, we maintain a probability distribution over class labels for each unlabeled data. The discriminative part of the cost is made proportional to the confidence over the assigned label of the participating training sample. This makes the proposed method robust to label assignment errors.

This paper makes the following contributions:[1]

1. We propose a discriminative dictionary learning method that utilizes both labeled and unlabeled data.
2. Using the kernel trick, we extend the formulation for learning linear dictionaries with labeled and unlabeled data to the non-linear case. An efficient optimization procedure is proposed for solving this non-linear dictionary learning problem.
3. We show how the proposed method can be extended to ambiguously labeled data where each training sample has multiple labels and only one of them is correct.

In our previous work [30], we developed a supervised non-linear discriminative dictionary learning method for image classification. The method proposed in this paper is different from [30] in that it is a general non-linear semi-supervised dictionary learning method. The methods proposed for learning dictionaries form ambiguously labeled data [7] are also different from the one proposed in this paper. Specifically, in [7] two linear methods are proposed – one based on soft decision rules and the other based on hard decision rules. In contrast to linear reconstructive dictionary leaning methods in [7,38], we propose a general discriminative non-linear kernel dictionary learning method for semi-supervised learning.

The rest of the paper is organized as follows. In Section 2, we formulate the problem of non-linear dictionary learning with partially labeled data. The optimization of the proposed framework is presented in Section 3. Experimental results are presented in Section 4 and Section 5 concludes the paper with a brief summary and discussion.

2. Problem formulation

In this section, we formulate the optimization problem for learning discriminative dictionaries with partially labeled data. We first present the linear formulation. We then extend it to the non-linear case.

[1] Preliminary version of this work appeared in [31]. Items 2 and 3 are extensions to [31].

2.1. Linear dictionary learning with partially labeled data

Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^{d \times N}$ be the data matrix where d is the dimension of each data sample $\mathbf{y}_i$ and N is the total number of training samples. We assume that the data is partially labeled and denote the label of the ith sample by l_i. When the sample $\mathbf{y}_i$ is not labeled, we set l_i to 0, i.e., $l_i \in \{0, 1, \ldots C\}$, where C is the total number of classes.

Our goal is to learn a dictionary $\mathbf{D} \in \mathbb{R}^{d \times K}$, where K is the number of unit norm atoms. We represent this dictionary as the concatenation of all the classes' dictionary, i.e. $\mathbf{D} \triangleq [\mathbf{D}_1 | \ldots | \mathbf{D}_C]$ such that each $\mathbf{D}_c \in \mathbb{R}^{d \times K_c}$ can represent the cth class data well while not economically representing the other class data. Here, K_c is the number of atoms in dictionary $\mathbf{D}_c$, and hence, $K = \sum_{c=1}^{C} K_c$. Enforcing each $\mathbf{D}_c$ to represent only its own class c improves the discriminative capability of the learned dictionary. We represent each sample $\mathbf{y}_i$ by sparse linear combination of dictionary $\mathbf{D}$'s atoms and represent the sparse coefficient of the ith sample by $\mathbf{x}_i$. Furthermore, we denote the coefficient matrix for all the samples by $\mathbf{X}$, i.e., $\mathbf{X} \triangleq [\mathbf{x}_1, \ldots, \mathbf{x}_N]$.

In order to deal with unlabeled data, we introduce a probability matrix $\mathbf{P} \in \mathbb{R}^{C \times N}$ such that each column of $\mathbf{P}$ represents the class distribution of the corresponding data sample. In other words, (c, i)th element P_{ci} of $\mathbf{P}$ denotes the probability of the ith sample belonging to class c. Hence, by definition,

$$
\begin{aligned}
&P_{ci} = 1 \quad \text{if } \mathrm{y}_i \text{ is labeled with one class and } l_i = c.\\
&P_{ci} = 0 \quad \text{if } \mathrm{y}_i \text{ is labeled with one class and } l_i \neq c.\\
&0 \le P_{ci} \le 1 \quad \text{if } \mathbf{y}_i \text{ is unlabeled or ambiguously labeled.}
\end{aligned} \tag{1}
$$

We denote the probability of all the samples belonging to class c by a diagonal matrix $\mathbf{P}_c \in \mathbb{R}^{N \times N}$ such that $\mathbf{P}_c(i,i) = P_{ci}$ and the non-diagonal elements of $\mathbf{P}_c$ are set equal to zeros. Also, we define a matrix $\mathbf{Q}_c \triangleq 1 - \mathbf{P}_c$ to denote the probability of all the samples not belonging to the cth class. Furthermore, we define $\mathbf{P}_c^{sqrt}$ and $\mathbf{Q}_c^{sqrt}$ the square root of $\mathbf{P}_c$ and $\mathbf{Q}_c$, respectively, i.e., $\mathbf{P}_c = \mathbf{P}_c^{sqrt}\mathbf{P}_c^{sqrt}$ and $\mathbf{Q}_c = \mathbf{Q}_c^{sqrt}\mathbf{Q}_c^{sqrt}$. The Frobenius norm and the sparsity promoting ℓ_1 norm of a matrix $\mathbf{A}$ are denoted as $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_1$, respectively.

Equipped with these notations, we formulate the dictionary learning problem as one of optimizing

$$
\mathcal{J}_0(\mathbf{D}, \mathbf{X}, \mathbf{P}) = \mathcal{F}_0(\mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{P}) + \mathcal{H}(\mathbf{X}, \mathbf{P}) + \lambda_1 \|\mathbf{X}\|_1, \tag{2}
$$

where

$$
\begin{aligned}
\mathcal{F}_0(\mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{P}) = {} & \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \\
& + \tau_1 \sum_{c=1}^{C} \|(\mathbf{Y} - \mathbf{D}_c \mathbf{X}^c)\mathbf{P}_c^{sqrt}\|_F^2 \\
& + \tau_2 \sum_{c=1}^{C} \|\mathbf{D}_c \mathbf{X}^c \mathbf{Q}_c^{sqrt}\|_F^2,
\end{aligned} \tag{3}
$$

$$
\mathcal{H}(\mathbf{X}, \mathbf{P}) = \lambda_2(\mathrm{tr}(S_w(\mathbf{X}, \mathbf{P}) - S_b(\mathbf{X}, \mathbf{P}))) + \eta \|\mathbf{X}\|_F^2, \tag{4}
$$

and $\mathbf{X}^c$ is the coefficient matrix corresponding to the cth class. Here, the first term of $\mathcal{F}_0$ encourages $\mathbf{D}$ to be a good representative of the data matrix $\mathbf{Y}$ without needing any label information. The second term of $\mathcal{F}_0$ enforces that the cth class dictionary $\mathbf{D}_c$ represents well those samples which are likely to belong to class c. Note that $\mathbf{P}_c^{sqrt}$ is a diagonal matrix and hence the contribution of each sample in this part of the cost is proportional to the probability of it having come from the cth class. The third part of $\mathcal{F}_0$ enlarges the reconstruction error of those samples which are less likely to have come from the cth class. The parameters τ_1 and τ_2 control the discriminative capability of the learned dictionary.

The second term $\mathcal{H}$ of $\mathcal{J}_0$ in (2) makes the sparse coefficients of samples discriminative by decreasing the trace of within-class