Pattern Recognition 48 (2015) 2667–2675

Contents lists available at ScienceDirect

Pattern Recognition

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

Multiple kernel-based dictionary learning for weakly supervised classification

Ashish Shrivastava [a,*], Jaishanker K. Pillai [b], Vishal M. Patel [a]

[a] Center for Automation Research, UMIACS, University of Maryland, College Park, MD 20742, United States
[b] Google, Mountain View, CA 94043, United States

ARTICLE INFO

Article history:
Received 1 October 2013
Received in revised form
26 January 2015
Accepted 6 March 2015
Available online 14 March 2015

Keywords:
Dictionary learning
Multiple instance learning
Multiple kernel learning

ABSTRACT

In this paper, we develop a multiple instance learning (MIL) algorithm using the dictionary learning framework where the labels are given in the form of positive and negative bags, with each bag containing multiple samples. A positive bag is guaranteed to have only one positive class sample while all the samples in a negative bag belong to the negative class. Given positive and negative bags of data, our method learns appropriate feature space to select positive samples from the positive bags as well as optimal dictionaries to represent data in these bags. We apply this method for digit recognition, action recognition, and gender recognition tasks and demonstrate that the proposed method is robust and can perform significantly better than many competitive two class MIL classification algorithms.

1. Introduction

Acquiring good quality labeled training data is one of the critical steps in building an object recognition system. While human annotation is the popular choice for obtaining labeled data for training, it is expensive and time consuming. However, it is relatively easy to obtain weakly labeled data from the Internet. Such data can be obtained through web search queries, captions, subtitles of movies and from amateur raters without full knowledge about the object categories. This has lead to the development of algorithms for weakly supervised object classification.

A popular machine learning paradigm to handle weakly supervised data is the Multiple Instance Learning (MIL) [7]. In MIL paradigm, examples are not individually labeled but grouped into labeled sets or bags of samples. A positive bag contains at least one positive example, and a negative bag contains only negative examples. Various MIL algorithms have been proposed in the literature for classification [17,30,2]. The MIL algorithms have been used to handle label errors, by collecting multiple samples with possible label errors into positive bags. Effect of alignment errors in training data can be reduced by forming bags with multiple shifted templates. The MIL-based algorithms have also been developed for robust tracking of objects [3].

In recent years, the field of sparse representation and dictionary learning has undergone rapid development, both in theory and in algorithms. It has also been successfully applied to numerous image understanding applications. This is partly due to the fact that signals or images of interest, though high dimensional, can often be coded using few representative atoms in some dictionary. These dictionaries can be either analytic or they can be learned directly from the data. Often, learning a dictionary directly from data usually leads to improved results in many practical applications such as classification and restoration [32,21,22]. This has motivated researchers to develop robust dictionary learning algorithms for various learning scenarios ranging from fully supervised [23,10,25,14–16], to weakly supervised [28], to unsupervised [29,24,5,8]. Note that this work is fundamentally different than the work of Shrivastava et al. [28]. Input to the method of Shrivastava et al. [28] is labeled or unlabeled samples and their method does not handle bags. In other words, their method works under semi-supervised setting and not under the MIL setting. Furthermore, Shrivastava et al. [28] use a predefined kernel, while the proposed method learns an optimal one based on the Multiple Kernel Learning (MKL) method [9,27].

While the MIL algorithms exist for popular classification methods like Support Vector Machines (SVM), logistic regression and boosting, such algorithms have not been studied thoroughly in the literature using the dictionary learning framework. Recently, Huo et al. [12] explored dictionary-based MIL method for detecting abnormal events in videos by predicting the labels of the instances in the positive bag. Similarly, Wang et al. [31] learned a multi-class classification matrix for object representation. In this paper, we develop an MIL algorithm using the non-linear dictionary learning framework by projecting the data into a feature space. We formulate the multiple instance learning problem as a kernel

* Corresponding author.
E-mail addresses: ashish@umiacs.umd.edu (A. Shrivastava), jaypillai@google.com (J.K. Pillai), pvishalm@umiacs.umd.edu (V.M. Patel).

http://dx.doi.org/10.1016/j.patcog.2015.03.005
0031-3203/

learning problem and iteratively learn the dictionary in the embedded space of the learned kernel. Multiple kernel learning essentially combines multiple kernels instead of using a single predefined kernel [9]. Different kernels correspond to different notions of similarity between two data samples. In particular, in a high dimensional feature space, it is not optimal to choose one kernel for all the datasets. In the case of MIL, the kernel is learned in a discriminative manner, ensuring that the negative samples have high reconstruction error on the positive dictionary in the embedded space. This, in turn, reduces the effect of negative samples in the positive bag on the learned positive dictionary. A block diagram of the proposed algorithm is given in Fig. 1.

The key contributions of our work are

1. We develop a multiple instance dictionary learning framework to handle weakly supervised data.
2. We also demonstrate how kernel learning can be incorporated into the dictionary learning framework so that data from negative and positive bags are well represented at the same time positive and negative classes are separated in the feature space.
3. We propose a novel classification procedure based on the proposed multiple instance dictionary framework.
4. We demonstrate the effectiveness our approach on three publicly available image classification datasets.

1.1. Organization of the paper

This paper is organized as follows. Section 2 defines and formulates the multiple instance dictionary learning problem. Details of the optimization problem are presented in Section 3. A classification procedure using our proposed dictionary learning method is presented in Section 4. Experimental results are presented in Section 5 and Section 6 concludes the paper with a brief summary and discussion.

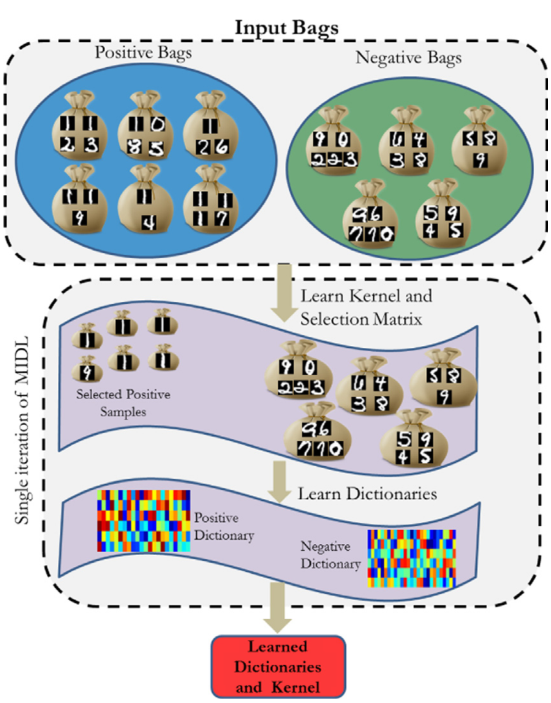

Fig. 1. Overview of our method. Given positive and negative bags of data, our method learns appropriate feature space to select positive samples from the positive bags as well as optimal dictionaries to represent data in these bags.

2. Problem formulation

Given a set of training samples, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^{d \times N}$, one can learn a dictionary $\mathbf{D} \in \mathbb{R}^{d \times K}$ with K atoms that leads to the best representation for each member in this set, under strict sparsity constraints by solving the following optimization problem:

$$\arg\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \quad \text{subject to } \forall i \ \|\mathbf{x}_i\|_0 \le T_0, \tag{1}$$

where $\mathbf{x}_i$ represents the ith column of coefficient matrix $\mathbf{X} \in \mathbb{R}^{K \times N}$ and T_0 is the sparsity parameter. Here, the Frobenius norm is defined as $\|\mathbf{A}\|_F = \sqrt{\sum_{ij} A_{ij}^2}$ and the norm $\|\mathbf{x}\|_0$ counts the number of non-zero elements in $\mathbf{x}$. Various algorithms have been proposed in the literature that can solve the above optimization problem [1,32].

Using the kernel trick, one can also make the dictionary learning model (1) non-linear [19,20]. Let $\mathbf{\Phi} : \mathbb{R}^d \to G$ be a non-linear mapping from a d dimensional space into a dot product space G. A non-linear dictionary can be trained in the feature space G by solving the following optimization problem:

$$\arg\min_{\mathbf{A},\mathbf{X}} \|\mathbf{\Phi}(\mathbf{Y}) - \mathbf{\Phi}(\mathbf{Y})\mathbf{A}\mathbf{X}\|_F^2 \quad \text{s.t. } \forall i \ \|\mathbf{x}_i\|_0 \le T_0, \tag{2}$$

where $\mathbf{\Phi}(\mathbf{Y}) = [\mathbf{\Phi}(\mathbf{y}_1), \ldots, \mathbf{\Phi}(\mathbf{y}_N)]$. Since the dictionary lies in the linear span of the samples $\mathbf{\Phi}(\mathbf{Y})$, in (2) we have used the following model for the dictionary in the feature space, $\tilde{\mathbf{D}} = \mathbf{\Phi}(\mathbf{Y})\mathbf{A}$, where $\mathbf{A} \in \mathbb{R}^{N \times K}$ is a matrix with K atoms [19]. This model provides adaptivity via modification of the matrix $\mathbf{A}$. After some algebraic manipulations, the cost function in (2) can be rewritten as

$$\|\mathbf{\Phi}(\mathbf{Y}) - \mathbf{\Phi}(\mathbf{Y})\mathbf{A}\mathbf{X}\|_F^2 = \mathrm{tr}((\mathbf{I} - \mathbf{A}\mathbf{X})^T \mathcal{K}(\mathbf{Y},\mathbf{Y})(\mathbf{I} - \mathbf{A}\mathbf{X})),$$

where $\mathcal{K}(\mathbf{Y},\mathbf{Y})$ is a kernel matrix whose elements are computed from $\kappa(i,j) = \mathbf{\Phi}(\mathbf{y}_i)^T \mathbf{\Phi}(\mathbf{y}_j)$. It is apparent that the objective function is feasible since it only involves a matrix of finite dimension $\mathcal{K} \in \mathbb{R}^{N \times N}$, instead of dealing with a possibly infinite dimensional dictionary.

An important property of this formulation is that the computation of $\mathcal{K}$ only requires dot products. Therefore, we are able to employ Mercer kernel functions to compute these dot products without carrying out the mapping $\mathbf{\Phi}$. Some commonly used kernels include polynomial kernels $\kappa(\mathbf{x},\mathbf{y}) = \langle(\mathbf{x},\mathbf{y}\rangle + c)^d$ and Gaussian kernels $\kappa(\mathbf{x},\mathbf{y}) = \exp\left(-\|\mathbf{x}-\mathbf{y}\|^2/c\right)$, where c and d are the parameters.

In multiple instance learning setting, we have labeled bags instead of samples for training. Each bag is labeled either $+1$ or -1, called a positive or a negative bag, respectively. A negative bag will have only negative samples. A positive bag is guaranteed to have at least one positive sample, while the remaining ones can be either positive or negative. We denote the bth positive bag by a matrix $\mathbf{Y}_b^p \triangleq [\mathbf{y}_{b,1}^p, \ldots, \mathbf{y}_{b,m_b}^p] \in \mathbb{R}^{d \times m_b}$, whose columns are the m_b positive samples. Here, d is the dimension of data sample. Similarly, let $\mathbf{Y}_b^n \triangleq [\mathbf{y}_{b,1}^n, \ldots, \mathbf{y}_{b,n_b}^n] \in \mathbb{R}^{d \times n_b}$ be the bth negative bag containing the n_b negative samples. We denote concatenation of all positive bags with $\mathbf{Y}_p$ and that of negative bags with $\mathbf{Y}_n$, i.e. $\mathbf{Y}_p \triangleq [\mathbf{Y}_1^p, \ldots, \mathbf{Y}_{N_p}^p] = [\mathbf{y}_1^p, \ldots, \mathbf{y}_M^p] \in \mathbb{R}^{d \times M}$ and $\mathbf{Y}_n \triangleq [\mathbf{Y}_1^n, \ldots, \mathbf{Y}_{N_n}^n] = [\mathbf{y}_1^n, \ldots, \mathbf{y}_N^n] \in \mathbb{R}^{d \times N}$, where $M \triangleq \sum_{b=1}^{N_p} m_b$ is the total number of positive samples in all the positive bags and $N \triangleq \sum_{b=1}^{N_n} n_b$ is the total number of negative samples in all the negative bags. There are N_p positive bags and N_n negative bags in total. The bth positive bag contains m_b samples, while the bth negative bag has n_b samples. Given $\mathbf{Y}_n$ and $\mathbf{Y}_p$, the objective is to learn a non-linear dictionary-based model that can classify a novel test sample to a positive or a negative class.

We denote the negative dictionary, in feature space, as $\tilde{\mathbf{D}}_n = \mathbf{\Phi}(\mathbf{Y}_n)\mathbf{A}_n$ and positive dictionary as $\tilde{\mathbf{D}}_p = \mathbf{\Phi}(\mathbf{Y}_p)\mathbf{A}_p$, where $\mathbf{A}_n \in \mathbb{R}^{N \times K_n}$ and $\mathbf{A}_p \in \mathbb{R}^{M \times K_p}$ are matrices with K_n and K_p being the