Neural Networks 49 (2014) 96–106

Contents lists available at ScienceDirect

Neural Networks

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

A trace ratio maximization approach to multiple kernel-based dimensionality reduction

Wenhao Jiang, Fu-lai Chung [*]

Department of Computing, Hong Kong Polytechnic University, Hunghom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 30 August 2012
Received in revised form 3 September 2013
Accepted 18 September 2013

Keywords:
Kernel learning
Dimensionality reduction
Graph embedding
Supervised learning
Unsupervised learning

ABSTRACT

Most dimensionality reduction techniques are based on one metric or one kernel, hence it is necessary to select an appropriate kernel for kernel-based dimensionality reduction. Multiple kernel learning for dimensionality reduction (MKL-DR) has been recently proposed to learn a kernel from a set of base kernels which are seen as different descriptions of data. As MKL-DR does not involve regularization, it might be ill-posed under some conditions and consequently its applications are hindered. This paper proposes a multiple kernel learning framework for dimensionality reduction based on regularized trace ratio, termed as MKL-TR. Our method aims at learning a transformation into a space of lower dimension and a corresponding kernel from the given base kernels among which some may not be suitable for the given data. The solutions for the proposed framework can be found based on trace ratio maximization. The experimental results demonstrate its effectiveness in benchmark datasets, which include text, image and sound datasets, for supervised, unsupervised as well as semi-supervised settings.

1. Introduction

Because of the curse of dimensionality and the requirement of computational efficiency, a lot of dimensionality reduction methods have been developed for particularly high-dimensional data applications. These methods can be classified into two categories, i.e., unsupervised and supervised methods, depending on whether the label information is available or not. Principal component analysis (PCA) (Jolliffe, 1986) is a well-known unsupervised dimensionality reduction technique. It aims at identifying a linear transformation such that the variance of transformed data is maximized. Recently, manifold assumption was introduced into dimensionality reduction. To preserve the manifold structure, nonlinear dimensionality reduction methods such as ISOMAP (Tenenbaum, de Silva, & Langford, 2000), locally linear embedding (LLE) (Roweis & Saul, 2000) and Laplacian Eigenmap (LE) (Belkin & Niyogi, 2001) were proposed. Moreover, a linear approximation of LE called locality preserving projections (LPP) was proposed in He and Niyogi (2004). Unsupervised dimensionality reduction does not involve any label information. However, if there exists label information, supervised dimensionality reduction can be conducted. Linear discriminant analysis (LDA) is a typical supervised dimensionality reduction method, which tries to identify a lower dimensional space minimizing the within-class covariance while maximizing the between-class covariance simultaneously. Its kernelized version called kernel Fisher discriminant analysis (KFDA) has been introduced in Mika, Ratsch, Weston, Scholkopf, and Mullers (1999). Other popular supervised dimensionality reduction methods include partial least squares (PLS) (Wold, 1985) and canonical correlation analysis (CCA) (Hardoon, Szedmak, & Shawe-Taylor, 2004). PLS finds orthogonal projection directions for the input data by maximizing its covariance with the output. The original formulation of CCA is not for dimensionality reduction, however, with label information as one of the two views, CCA becomes a supervised dimensionality reduction method that could extract an effective representation of the object by correlating the linear relationships between the two views of the object.

Many dimensionality reduction techniques could be unified under a common framework. For example, ISOMAP, LLE and LE could be unified by kernel PCA (Ham, Lee, Mika, & Schölkopf, 2004) with specially constructed Gram matrices. Recently, a new framework called graph embedding (Yan et al., 2007) was proposed and it covers many dimensionality reduction technologies, like ISOMAP, LLE, LE, LPP, PCA and LDA. Graph embedding was extended to multiple kernels setting in Lin, Liu, and Fuh (2011) called multiple kernel learning for dimensionality reduction (MKL-DR). The goal of MKL-DR is to learn a transformation matrix from multiple descriptions of data represented by different base kernels. MKL-DR provides the ability of learning a unified space of lower dimension for data in multiple feature representations. It finds the solutions with two relaxation steps and its effectiveness has been demonstrated in image clustering and face recognition tasks. The advantage of using multiple kernels instead of only one kernel in the learning tasks of

[*] Corresponding author. Tel.: +852 2766 7289.
E-mail addresses: cswhjiang@comp.polyu.edu.hk (W. Jiang), cskchung@comp.polyu.edu.hk, cskchung2003@yahoo.com.hk (F.-l. Chung).

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.09.004

classification and dimensionality reduction has also been demonstrated in Choi, Choi, Katake, Kang, and Choe (2010) and a nonlinear way to combine the kernels has been proposed.

In this paper, we propose a regularized multiple kernel learning framework for dimensionality reduction. Based on this framework, a method to learn a kernel and a linear transformation matrix is introduced and algorithms to find approximate solutions by trace ratio maximization (Ngo, Bellalij, & Saad, 2012) are derived. By focusing on techniques pertaining to dimensionality reduction, the proposed formulation introduces a new class of applications with the multiple kernel learning framework to address not only supervised learning problems but also unsupervised and semi-supervised ones.

The rest of the paper is structured as follows. A review of related works is presented in Section 2. In Section 3, a unified framework is introduced and based upon which our new dimensionality reduction method (MKL-TR) is proposed. The experimental results in supervised, unsupervised and semi-supervised datasets are reported in Section 4. Concluding remarks and a discussion of future works are given in Section 5.

Notation: In the rest of this paper, we denote the sample set as matrix $X = [x_1, x_2, \dots, x_n]$, where $x_i \in R^m$ is an m-dimensional vector. For supervised dimension reduction task, the class label of the sample x_i is assumed to be $y_i \in \{1, \dots, c\}$ and c denotes the number of classes. The M base kernels are denoted as K_i and the corresponding nonnegative coefficients are $\alpha = [\alpha_1, \dots, \alpha_M]^T$. For a given dimensionality reduction task, we seek to find a kernel $K_\alpha = \sum_i^M \alpha_i K_i$ and a transformation matrix P such that the resulting data $z_i = P^T \Phi(x_i)$ have a lower dimensionality d. For the kernel feature space induced, P could be expressed as $P = \Phi(X)A$, where $\Phi(X)$ is the data matrix in the feature space and $A \in R^{n\times d}$ is the coefficient matrix for P, the reduced data matrix could be expressed as $Z = P^T\Phi(X) = (\Phi(X)A)^T\Phi(X) = A^T\Phi(X)^T\Phi(X) = A^TK$.

2. Related works

2.1. Multiple kernel learning for discriminant analysis

In Lanckriet, Cristianini, Bartlett, Ghaoui, and Jordan (2004), Lanckriet et al. pioneered the work of multiple kernel learning which integrates the tuning of kernels into the learning process. The idea has been applied to discriminant analysis with kernel Fisher discriminant analysis (KFDA) and regularized kernel discriminant analysis (RKDA) being used in optimal kernel selection (OKS) (Kim, Magnani, & Boyd, 2006) and discriminant kernel learning (DKL) (Ye, Ji, & Chen, 2008) respectively. Both of them are supervised techniques, with OKS for binary-class data and DKL extending it to multiple-class settings. They aim at learning an optimal kernel based on regularized kernel Fisher discriminant analysis (KFDA) for which the following objective function is commonly used:

$$\max_A \operatorname{tr}((A^T(KHK + \lambda K)A)^{-1}A^TKS_bKA), \tag{1}$$

where K is a kernel matrix, $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix, $\lambda > 0$ is the regularization parameter, and S_b is defined as

$$S_b(i,j) = \begin{cases} n/n_{y_i} - 1 & \text{if } y_i = y_j \\ -1 & \text{if } y_i \neq y_j, \end{cases} \tag{2}$$

in which n_{y_i} is the total number of samples in the class of data i. With the optimal A found, the optimal value of problem (1) becomes

$$\operatorname{tr}((KHK + \lambda K)^{-1}KS_bK). \tag{3}$$

DKL aims to find a combination of base kernels such that the above value is maximized.

2.2. Multiple kernel learning for dimensionality reduction

To learn the transformation matrix from multiple sources, MKL-DR (Lin et al., 2011) extended graph embedding to multiple kernel learning setting.

2.2.1. Graph embedding

In Yan et al. (2007), graph embedding is proposed to provide a unified framework for dimensionality reduction. Graph embedding defines two graphs G and G', and W and W' are the corresponding affinity matrices. The projection vector v could be obtained by solving

$$\begin{aligned} &\min_v \ v^TXLX^Tv \\ &\text{s.t. } v^TXL'X^Tv = 1, \quad \text{or} \\ &v^TXDX^Tv = 1, \end{aligned} \tag{4}$$

where $L = \operatorname{diag}(W\mathbf{1}) - W$ and $L' = \operatorname{diag}(W'\mathbf{1}) - W'$ are graph Laplacian of G and G' respectively. To find d such vectors, the following generalized eigenvalue problem has to be solved

$$\begin{aligned} &\min_v \ \operatorname{tr}(V^TXLX^TV) \\ &\text{s.t. } V^TXL'X^TV = I. \end{aligned} \tag{5}$$

The PCA (Jolliffe, 1986), ISOMAP (Tenenbaum et al., 2000), LLE (Roweis & Saul, 2000), LPP (He & Niyogi, 2004), LDA, local discriminant embedding (LDE) (Chen, Chang, & Liu, 2005), and marginal Fisher analysis (MFA) (Yan et al., 2007) can be expressed by graph embedding.

2.2.2. MKL-DR

Multiple kernel learning for dimensionality reduction (MKL-DR) made an attempt to combine multiple kernel learning with dimensionality reduction. It aims to find a linear combination of base kernels $K_\alpha = \sum_i^M \alpha_i K_i$, where $\alpha_i \geq 0$, and a transformation matrix such that the following objective function is optimized

$$\begin{aligned} &\min_{A,\alpha} \sum_{i,j=1}^n \|A^T\mathbb{K}^{(i)}\alpha - A^T\mathbb{K}^{(j)}\alpha\|^2 W_{ij} \\ &\text{s.t. } \sum_{i,j=1}^n \|A^T\mathbb{K}^{(i)}\alpha - A^T\mathbb{K}^{(j)}\alpha\|^2 W'_{ij} = 1, \\ &\alpha_i \geq 0, \quad i = 1, 2, \dots, M, \end{aligned} \tag{6}$$

where

$$\mathbb{K}^{(i)} = \begin{bmatrix} K_1(1,i) & \cdots & K_M(1,i) \\ \vdots & \ddots & \vdots \\ K_1(n,i) & \cdots & K_M(n,i) \end{bmatrix} \in R^{n\times M}. \tag{7}$$

It can be easily shown that such a formulation of MKL-DR is equivalent to

$$\begin{aligned} &\min_{A,\alpha} \ \operatorname{tr}(A^TK_\alpha L_1 K_\alpha A) \\ &\text{s.t. } \operatorname{tr}(A^TK_\alpha L_2 K_\alpha A) = 1, \\ &\alpha_i \geq 0, \quad i = 1, 2, \dots, M, \end{aligned} \tag{8}$$

where $L_1 = \operatorname{diag}(W\mathbf{1}) - W$ and $L_2 = \operatorname{diag}(W'\mathbf{1}) - W'$. In Lin et al. (2011), it is solved by updating A and α alternately. Given α, problem (8) is relaxed into a generalized eigenvalue problem:

$$\begin{aligned} &\min_A \ \operatorname{tr}(A^TK_\alpha L_1 K_\alpha A) \\ &\text{s.t. } A^TK_\alpha L_2 K_\alpha A = I \end{aligned} \tag{9}$$