Neurocomputing 155 (2015) 205–218

Contents lists available at ScienceDirect

Neurocomputing

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

Instance-specific canonical correlation analysis

Deming Zhai [a,*], Yu Zhang [b], Dit-Yan Yeung [c], Hong Chang [d], Xilin Chen [d], Wen Gao [a,e]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[b] Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
[c] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China
[d] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China
[e] Institute of Digital Media, Peking University, Beijing, China

ARTICLE INFO

Article history:
Received 2 October 2013
Received in revised form
2 November 2014
Accepted 12 December 2014
Communicated by D. Zhang
Available online 2 January 2015

Keywords:
Canonical correlation analysis
Least squares regression
Multi-view statistical learning

ABSTRACT

Canonical Correlation Analysis (CCA) is one of the most popular statistical methods to capture the correlations between two variables. However, it has limitations as a linear and global algorithm. Although some variants have been proposed to overcome the limitations, neither of them achieves locality and nonlinearity at the same time. In this paper, we propose a novel algorithm called Instance-Specific Canonical Correlation Analysis (ISCCA), which approximates the nonlinear data by computing the instance-specific projections along the smooth curve of the manifold. First, we propose a least squares solution for CCA which will set the stage for the proposed method. Second, based on the framework of least squares regression, CCA is extended to the instance-specific case which obtains a set of locally linear smooth but globally nonlinear transformations. Third, ISCCA can be extended to semi-supervised setting by exploiting the unlabeled data to further improve the performance. The optimization problem is proved to be convex and could be solved efficiently by alternating optimization. And the globally optimal solutions could be achieved with theoretical guarantee. Moreover, for large scale applications, iterative conjugate gradient algorithm can be used to speed up the computation procedure. Experimental results demonstrate the effectiveness of our proposed method.

1. Introduction

In many computer vision and pattern recognition applications, an object has multiple observations corresponding to multiple views which are related yet different. This scenario can be found in a wide range of applications, including pose estimation, facial expression recognition, object recognition involving images taken at different camera angles, and identity recognition in video and audio streams. Some recently emerged applications including face matching with both near-infrared and visual images [16] and the alignment of face images taken at different resolutions [18] also belong to this category. In such applications, modeling the statistical correlations or commonalities between multi-view observations plays a very crucial role.

Canonical Correlation Analysis (CCA) [14] is one of the most popular statistical methods for capturing the linear correlation between two multivariate random vectors corresponding to two different views. The goal of CCA is to seek a pair of linear transformations that maximize the correlation between two views after projecting the data to a common lower-dimensional space by applying the transformations. In recent years, CCA has been successfully applied to a wide range of computer vision applications, such as image processing and analysis [23], facial expression recognition [37], pose estimation [21], and image retrieval [12]. However, a limitation of CCA is that it can only reveal the linear correlation relationship between different views in a global way, making it inadequate for some more complicated applications.

As pointed out by Bottou and Vapnik [6], it is usually not easy to find a single function which holds good predictability for the entire data space, but it is much easier to seek some functions that are capable of producing good predictions on some specified regions. For many computer vision applications in particular, it has been demonstrated that the idea of local learning is very useful, e.g., [2,31,33,32,36]. For CCA, instead of seeking a common transformation for all instances from the same view, the model can exhibit much higher flexibility if each instance is allowed to have a specific transformation for projecting it to a common lower-dimensional space. To control the model complexity, the transformations for neighboring instances may be constrained to have the same or similar representation forms. This local learning approach can provide greater flexibility than introducing global nonlinearity by using the kernel trick [27].

To overcome the limitations of CCA, several variants of CCA have been proposed in the literature. Kernel CCA (KCCA) [1] is a kernel

* Corresponding author. Tel.: +86 10 62600523.
E-mail address: zhaideming@hit.edu.cn (D. Zhai).

http://dx.doi.org/10.1016/j.neucom.2014.12.028
0925-2312/

extension of CCA which essentially applies CCA to data from different views after applying kernel-induced feature mappings. However, like most kernel methods, the nonlinearity in KCCA is applied in a global way in the sense that the nonlinear mapping is uniform anywhere, i.e., the induced kernel function with the same parameter is applied to all data pairs. Moreover, to deliver good performance, the choice of both the problem-dependent kernel function and its parameters is still a sticky problem. Non-Consolidating Correlation Analysis (NCCA) [10] also extends CCA by learning additional non-shared transformations for each view. NCCA comprises two steps: applying CCA to find shared embedded data and applying NCCA to find non-shared embedded data. In spirit, the learning of shared embedding space in the first step is in the same way as that in CCA. Another extension, called neural-network CCA (NNCCA) [15], is to discover nonlinear correlation with the help of the nonlinear processing ability of a neural network. Unfortunately, learning in neural networks suffers from some intrinsic problems such as long-time training, slow convergence and local optima.

While KCCA, NCCA and NNCCA extend CCA by providing nonlinear processing ability to some extent, neither of them takes the local structure of data into account and hence cannot satisfactorily deal with data with complex and nonlinear manifold structure. To tackle this problem, locality preserving CCA (LPCCA) [29] was proposed to introduce local manifold structure into CCA. In LPCCA, the globally nonlinear problem is decomposed into a series of locally linear sub-problems whose solutions can be combined to give the mapping vectors. Although local information is considered, the transformation learned in LPCCA is still global in the sense that the same transformation is applied to all instances from the same view.

To achieve locality and nonlinearity simultaneously, we propose in this paper a novel algorithm called Instance-Specific Canonical Correlation Analysis (ISCCA), which approximates the nonlinear data by computing the instance-specific projections along the smooth curve of the manifold. First, we propose a least squares solution for CCA to set the stage for the proposed method. From the least squares solution, we can recover the Euclidean distance in the common lower-dimensional space of CCA. Although similar results have been reported by [28], our analysis is under a more general framework than that in [28] which requires some assumption about the data dimensionality. Second, based on the framework of least squares regression (LSR), we extend CCA to the instance-specific case. Specially, each instance has its own specific transformation and the transformations are locally linear smooth but globally nonlinear. As a consequence, the learned model is more robust and flexible. Furthermore, under situations when large quantities of unlabeled data are also available, ISCCA can be extended for semi-supervised learning by exploiting the unlabeled data to further boost performance. Semi-supervised ISCCA is particularly useful when the labeled data is scarce. The last but not the least, our method can be formulated as a convex optimization problem and can be solved efficiently via an alternating optimization procedure. Irrespective of the initial value in iteration, the globally optimal solutions could be achieved with theoretical guarantee. Moreover, the proposed method combined with the iterative conjugate gradient algorithm LSQR [24] can be used for handling large-scale problems.

The contribution of the paper is highlighted in the following: (1) we model locality and nonlinearity jointly for multi-view correlation learning; (2) instance-specific projections are computed along the smooth curve of the manifold; (3) the convex objective functions are solved efficiently with global optimal solution.

The rest of this paper is organized as follows. In Section 2, some preliminary knowledge is introduced. In Section 3, we discuss the relationship between CCA and least squares regression under a more general framework. The proposed instance-specific CCA method is detailed in Section 4. Section 5 shows the experimental results with applications to manifold alignment for pose estimation and facial expression recognition, and cross-modal retrieval tasks. Finally, Section 6 gives some concluding remarks.

2. Preliminary work

We briefly review the CCA algorithm in this section. Let us represent two datasets $\mathcal{X}$ and $\mathcal{Y}$ with their matrix forms $\mathbf{X} \in \mathbb{R}^{d_x \times l}$ and $\mathbf{Y} \in \mathbb{R}^{d_y \times l}$, where l is the number of instances and each column vector $\mathbf{x}_i$ ($\mathbf{y}_j$) in $\mathbf{X}$ ($\mathbf{Y}$) denotes an instance with d_x (d_y) dimensions. For the case of reduction to one dimension, CCA computes two projection vectors $\mathbf{w}_x \in \mathbb{R}^{d_x}$ and $\mathbf{w}_y \in \mathbb{R}^{d_y}$ to maximize the correlation coefficient ρ between the two datasets:

$$\rho = \max_{\mathbf{w}_x,\mathbf{w}_y} \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y}}, \tag{1}$$

where $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ denote the sample covariance matrices of each view and $\mathbf{C}_{xy}$ denotes the sample covariance between the two views. Since ρ is invariant to the scaling of $\mathbf{w}_x$ and $\mathbf{w}_y$, CCA is equivalent to solving the following problem:

$$\max_{\mathbf{w}_x,\mathbf{w}_y} \ \mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y \quad \text{s.t.} \ \ \mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x = 1, \ \ \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y = 1. \tag{2}$$

By introducing Lagrange multipliers, $\mathbf{w}_x$ can be finally obtained by solving a generalized eigenvalue problem:

$$\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{w}_x = \lambda^2 \mathbf{C}_{xx}\mathbf{w}_x. \tag{3}$$

We can then compute $\mathbf{w}_y$ from $\mathbf{w}_x$ as

$$\mathbf{w}_y = \frac{\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{w}_x}{\lambda}. \tag{4}$$

If CCA is to project to multiple dimensions, multiple projection vectors under certain orthonormality constraints can be computed simultaneously by solving the following generalized eigenvalue problem:

$$\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{W}_x = \mathbf{C}_{xx}\mathbf{W}_x\mathbf{\Lambda}^2, \tag{5}$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing different λ's as diagonal entries. After obtaining the optimal $\mathbf{W}_x$, we can recover $\mathbf{W}_y$ as follows:

$$\mathbf{W}_y = \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{W}_x\mathbf{\Lambda}^{-1}. \tag{6}$$

CCA can also be transformed into the following problem where maximizing the canonical correlation in Eq. (1) is equivalent to minimizing the distance between the two observations:

$$\min_{\mathbf{W}_x,\mathbf{W}_y} \sum_{i=1}^{l} \|\mathbf{W}_x^T\mathbf{x}_i - \mathbf{W}_y^T\mathbf{y}_i\|_2^2 = \|\mathbf{W}_x^T\mathbf{X} - \mathbf{W}_y^T\mathbf{Y}\|_F^2 \quad \text{s.t.} \ \ \mathbf{W}_x^T\mathbf{X}\mathbf{X}^T\mathbf{W}_x = \mathbf{I}, \ \ \mathbf{W}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{W}_y = \mathbf{I}, \tag{7}$$

where $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the ℓ_2 vector norm and Frobenius matrix norm, respectively. The Frobenius matrix norm of an $m \times n$ matrix $\mathbf{A}$ with its elements a_{ij} is defined as $\|\mathbf{A}\|_F = \sqrt{\text{tr}(\mathbf{A}\mathbf{A}^T)} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$, where $\text{tr}(\cdot)$ denotes the trace of a square matrix.

3. Relationship between CCA and least squares regression

In this section, we first derive the solutions of CCA and least squares regression. Based on the derived solutions, we then establish the relationship between the solutions and verify their Euclidean distance based equivalence.