Neurocomputing 191 (2016) 263–272

Contents lists available at ScienceDirect

Neurocomputing

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

Canonical sparse cross-view correlation analysis

Chen Zu, Daoqiang Zhang *

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ARTICLE INFO

Article history:
Received 18 September 2015
Received in revised form
8 January 2016
Accepted 19 January 2016
Communicated by Yue Gao
Available online 23 February 2016

Keywords:
Canonical correlation analysis
Feature extraction
Sparse representation
Cross correlation
Dimensionality reduction

ABSTRACT

Recently, multi-view feature extraction has attracted great interest and Canonical Correlation Analysis (CCA) is a powerful technique for finding the linear correlation between two view variable sets. However, CCA does not consider the structure and cross view information in feature extraction, which is very important for subsequence tasks. In this paper, a new approach called Canonical Sparse Cross-view Correlation Analysis (CSCCA) is proposed to address this problem. We first construct similarity matrices by performing sparse representation between within-class samples. Then local manifold information and cross-view correlations are incorporated into CCA. Furthermore, a kernel version of CSCCA (KCSCCA) is proposed to reveal the nonlinear correlation relationship between two sets of features. We compare CSCCA and KCSCCA with existing multi-view feature extraction methods and perform experiments on both artificial data set and real world databases including multiple features and face data sets. The experimental results demonstrate the merits of our proposed method.

1. Introduction

Multi-view learning [1,2] which learns patterns or features from instances with multiple representations has been one of the hotspots in machine learning community. It has been shown that learning from multiple representations of data often achieves better performance than traditional single view learning methods. Recently, multi-view learning techniques have been extended to multi-view regression [3] and multi-view clustering [4].

Canonical correlation analysis (CCA) [5] is a learning method to find linear relationship between two groups of multidimensional variables. The goal of CCA is to seek two bases which would maximize the correlation of data by projecting two-view data obtained from various information sources, e.g. sound and image. In the past decades, CCA and its variants have been successfully applied to many fields such as image processing [6], pattern recognition [7,8], medical image analysis [9,10] and data regression analysis [11].

Standard CCA is an unsupervised linear dimensionality reduction method. It cannot preserve local structure in canonical subspaces and either cannot reveal nonlinear correlation relationship. In order to extract features with discriminant information, variants of CCA called discriminant CCA (DCCA) [12], random correlation ensemble (RCE) [13] and discriminative extended CCA (DECCA) [14] are proposed. For example, DCCA not only considers the correlation between two corresponding views of a sample, but also uses all the cross-view correlation between within-class examples. Several works have also discussed the relationships between CCA and LDA, especially when the data features are used in one view and the class labels are used in the other view [15,16]. It is shown that CCA and LDA have some equivalent relations [17]. In order to deal with nonlinear circumstance, some nonlinear CCA algorithms have been proposed in the literature [18]. Kernel methods [19,20] are widely used to reveal nonlinear structure in the original input space, and have been introduced into CCA (e.g., Kernel CCA (KCCA)) [21]. KCCA first maps the data into high dimensional feature space by implicit nonlinear mappings, and then traditional CCA is performed in the feature space in which the nonlinear problem in the original space is converted into a linear one. However, like many other kernel methods, one disadvantage of KCCA is the choice of appropriate kernel and kernel parameters. Neural networks based nonlinear CCA suffers from some intrinsic limitations such as long-time training, slow convergence and local minima [18].

In recent years, locality preserving methods such as locally linear embedding (LLE) [22], Isomap [23] and locality preserving projections (LPP) [24] have achieved a remarkable flourish in single-view dimensionality reduction. These methods preserve the neighborhood information so as to discover the low dimensional manifold structure embedded in the original high dimensional space. Inspired by the similar idea, Sun and Chen proposed a locality preserving CCA (LPCCA) [25]. LPCCA incorporates the local structure information into CCA and decomposes the global nonlinear problem into many local linear ones, consequently, in each

* Corresponding author.
E-mail address: dqzhang@nuaa.edu.cn (D. Zhang).

http://dx.doi.org/10.1016/j.neucom.2016.01.053
0925-2312/

small neighborhood field the problem can be treated as linear CCA and the global problem can be solved by optimizing the combination or integration of these local sub-problems. It has been shown that LPCCA performs better than CCA in discovering intrinsic structure of data for some applications, e.g., data visualization and pose estimation. Nevertheless, LPCCA only concerns the correlation between sample pairs and the discrimination of the extracted features which is important in subsequent classification task, while LPCCA is dependent on the parameter k which is manually chosen through experience.

Several supervised multi-view feature extraction methods have been proposed in recent researches. For example, Diethe et al. extended the convex formulation for Kernel Fisher Discriminant Analysis to multiple views [26]. Chen et al. proposed Hierarchical Multi-view Fisher Discriminant Analysis to improve the performance in classification and dimensionality reduction of multi-view task [27]. Sharma proposed a general multi-view feature extraction approach called Generalized Multiview Analysis (GMA) [28]. Although these works can do well in supervised learning situations, but they do not consider the intrinsic structures of the data, such as manifold structure. Local discrimination CCA (LDCCA) [29] and discriminative locality preserving CCA (DLPCCA) [30] can be seemed as extensions of CCA which use label and neighborhood information. Specifically, LDCCA not only considers the correlations between sample pairs but also the correlations between samples and their local neighborhoods. DLPCCA based on LDCCA can use label discriminative information to improve classification performance and preserve the geometric structure of data to enhance the smoothness of the extracted features. It worth noting that LPCCA, LDCCA and DLPCCA directly use the standard Euclidean distance to measure the similarity between data points which may be affected by outliers for the deficiency of robustness of Euclidean distance. In LPCCA, the locality means that the global nonlinear problem is decomposed into local linear ones. So the local structure information can be preserved in the canonical subspace. In LDCCA, the locality means the local neighborhood sample pairs are used to compute the correlations while DLPCCA considers the local structure information in two views separately.

In this paper, we propose a novel learning method for multi-view data called canonical sparse cross-view correlation analysis (CSCCA). We first construct similarity matrices by performing ℓ_1 norm sparse representation on within-class samples. Then local manifold structure information and cross-view correlation are incorporated into CCA. Here, we use sparse reconstruction because ℓ_1 norm is more robust to noises than Euclidean distance. The proposed method not only preserves the local structure information in two views separately, but also the structure information in the cross view. It is worth noting that many works have investigated sparse CCA [31,32]. In those papers, the projective vectors obtained by CCA need to be sparse that means there are a lot of zero values in it, while the sparse in our method means sparse reconstruction. The sparse representation achieved by minimizing a ℓ_1 regularization related objective function chooses its neighborhood automatically and does not have to encounter model parameters. The sparse in our method is totally different from those sparse CCA. The proposed method can be efficiently solved via generalized eigenvalue decomposition. Although the solution is similar to that of [33], they are derived from different motivation. Moreover, we extend CSCCA to kernel version (KCSCCA) to find nonlinear correlation. Experimental results on both synthetic and real world data sets including multiple features data set and face databases validate the effectiveness of the proposed method.

The rest of the paper is organized as follows: In Section 2, CCA is briefly reviewed, the proposed CSCCA and KCSCCA are then introduced in detail. The experiments and results on various data sets are given in Section 3. Finally, we conclude this paper in Section 4.

2. Proposed method

2.1. Canonical correlation analysis

Given a set of pair-wise data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}^q$, where $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ are samples from different views. Define that $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{p \times n}$ and $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n] \in \mathbb{R}^{q \times n}$. We assume that the data have been preprocessed with zero mean. CCA seeks to find two basis or projection vectors $\boldsymbol{w}_x \in \mathbb{R}^p$ and $\boldsymbol{w}_y \in \mathbb{R}^q$, such that the canonical variables $x = \boldsymbol{w}_x^T \boldsymbol{x}_i$ and $y = \boldsymbol{w}_y^T \boldsymbol{y}_i$ would be maximally correlated. The objective function of CCA can be formulated as

$$\max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\boldsymbol{w}_x^T \boldsymbol{C}_{xy} \boldsymbol{w}_y}{\sqrt{(\boldsymbol{w}_x^T \boldsymbol{C}_{xx} \boldsymbol{w}_x)(\boldsymbol{w}_y^T \boldsymbol{C}_{yy} \boldsymbol{w}_y)}} \tag{1}$$

where $\boldsymbol{C}_{xx} = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T = \boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{C}_{yy} = \sum_{i=1}^n \boldsymbol{y}_i \boldsymbol{y}_i^T = \boldsymbol{Y}\boldsymbol{Y}^T$ are within-sets covariance matrices and $\boldsymbol{C}_{xy} = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{y}_i^T = \boldsymbol{X}\boldsymbol{Y}^T$ is between-sets covariance matrix. Since the two basis vectors are scale independent, $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ can be obtained by solving the following optimization problem with constraints:

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

The optimization problem of CCA can be solved by applying Lagrangian equation to Eq. (2) and we can obtain the following generalized eigenvalue decomposition problem:

$$\begin{bmatrix} & \boldsymbol{C}_{xy} \\ \boldsymbol{C}_{xy}^T & \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{C}_{xx} & \\ & \boldsymbol{C}_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix} \tag{3}$$

2.2. Canonical sparse cross-view correlation analysis

In this section, in order to cope with the nonlinear problems and improve the performance of CCA in subsequent classification task, we propose a novel feature extraction method called canonical sparse cross-view correlation analysis (CSCCA). In CSCCA, the local structure information is incorporated and the cross correlations between two views from within-class samples are used by sparse representation.

The optimization problem of CCA can be written in the equivalent form [25] as:

$$\begin{aligned} \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \quad & \boldsymbol{w}_x^T \cdot \sum_{i=1}^n \sum_{j=1}^n (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^T \cdot \boldsymbol{w}_y \\ \text{s.t.} \quad & \boldsymbol{w}_x^T \cdot \sum_{i=1}^n \sum_{j=1}^n (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \cdot \boldsymbol{w}_x = 1 \\ & \boldsymbol{w}_y^T \cdot \sum_{i=1}^n \sum_{j=1}^n (\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^T \cdot \boldsymbol{w}_y = 1 \end{aligned} \tag{4}$$

We incorporate the local structure information and the within-class cross correlations into Eq. (4). The objective function of CSCCA can be formulated as follows:

$$\begin{aligned} \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \quad & \boldsymbol{w}_x^T \cdot \left(\sum_{i=1}^n \sum_{j=1}^n S_{ij}^x (\boldsymbol{x}_i - \boldsymbol{x}_j) S_{ij}^y (\boldsymbol{y}_i - \boldsymbol{y}_j)^T + \sum_{i=1}^n \sum_{j=1}^n (S_{ij}^x + S_{ij}^y) \boldsymbol{x}_i \boldsymbol{y}_j^T \right) \cdot \boldsymbol{w}_y \\ \text{s.t.} \quad & \boldsymbol{w}_x^T \cdot \sum_{i=1}^n \sum_{j=1}^n S_{ij}^x (\boldsymbol{x}_i - \boldsymbol{x}_j) S_{ij}^x (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \cdot \boldsymbol{w}_x = 1 \\ & \boldsymbol{w}_y^T \cdot \sum_{i=1}^n \sum_{j=1}^n S_{ij}^y (\boldsymbol{y}_i - \boldsymbol{y}_j) S_{ij}^y (\boldsymbol{y}_i - \boldsymbol{y}_j)^T \cdot \boldsymbol{w}_y = 1 \end{aligned} \tag{5}$$