Pattern Recognition Letters 32 (2011) 134–144

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Canonical correlation analysis using within-class coupling ☆

Olcay Kursun [a,*], Ethem Alpaydin [b], Oleg V. Favorov [c]

[a] Department of Computer Engineering, Istanbul University, Avcilar, Istanbul 34320, Turkey
[b] Department of Computer Engineering, Bogazici University, Bebek, Istanbul 34342, Turkey
[c] Biomedical Engineering Department, University of North Carolina, Chapel Hill, NC 27599-7575, USA

ARTICLE INFO

Article history:
Received 27 November 2009
Available online 31 October 2010
Communicated by W. Pedrycz

Keywords:
Temporal contextual guidance
Linear discriminant analysis (LDA)
Samples versus samples canonical correlation analysis (CCA)
Feature extraction

ABSTRACT

Fisher's linear discriminant analysis (LDA) is one of the most popular supervised linear dimensionality reduction methods. Unfortunately, LDA is not suitable for problems where the class labels are not available and only the spatial or temporal association of data samples is implicitly indicative of class membership. In this study, a new strategy for reducing LDA to Hotelling's canonical correlation analysis (CCA) is proposed. CCA seeks prominently correlated projections between two views of data and it has been long known to be equivalent to LDA when the data features are used in one view and the class labels are used in the other view. The basic idea of the new equivalence between LDA and CCA, which we call within-class coupling CCA (WCCCA), is to apply CCA to pairs of data samples that are most likely to belong to the same class. We prove the equivalence between LDA and such an application of CCA. With such an implicit representation of the class labels, WCCCA is applicable both to regular LDA problems and to problems in which only spatial and/or temporal continuity provides clues to the class labels.

1. Introduction

Fisher's linear discriminant analysis (LDA; Fisher, 1936) and Hotelling's canonical correlation analysis (CCA; Hotelling, 1936) are among the oldest, yet the most powerful multivariate data analysis techniques. LDA is one of the most popular supervised dimensionality reduction methods incorporating the categorical class labels of the data samples into a search for linear projections of the data that maximize the between-class variance while minimizing the within-class variance (Rencher, 1997; Alpaydin, 2004; Izenman, 2008).

On the other hand, CCA works with two sets of (related) variables and its goal is to find a linear projection of the first set of variables that maximally correlates with a linear projection of the second set of variables. These sets have recently been also referred to as *views* or *representations* (Hardoon et al., 2004). Finding correlated functions (covariates) of the two views of the same phenomenon by discarding the representation-specific details (noise) is expected to reveal the underlying hidden yet influential semantic factors responsible for the correlation (Hardoon et al., 2004; Becker, 1999; Favorov and Ryder, 2004; Favorov et al., 2003).

Both LDA and CCA have been proposed in 1936, and shortly after, a direct link between them has been shown by Bartlett (1938) as follows. Given a dataset of samples and their class labels, if we consider the features given for the data samples as one view, versus the class labels as the other view (a single binary variable works for the two-class problem but a form of 1-of-*C* coding scheme is typically used for multi-class categorical class labels), this CCA setup is known to be equivalent to LDA (Bartlett, 1938; Hastie et al., 1995). In other words, LDA can be simply said to be a special case of CCA.

The knowledge of this insightful equivalence between LDA and CCA enabled the researchers attempt the use of CCA to surpass the quality of the LDA projections. These attempts used samples versus their class labels using several other forms of representations for the labels (Loog et al., 2005; Barker and Rayens, 2003; Gestel et al., 2001; Johansson, 2001; Sun and Chen, 2007). An interesting example of such a label transformation is by replacing hard categorical labels by soft-labels; in (Sun and Chen, 2007), similar to the support vector idea, the aim was to put more weight on the samples near the class boundaries rather than using a common label for all the samples of a class; thus, more useful projections were found as more focus was placed on the problematic regions in the input space rather than the high-density regions with class centers. Another example is the study on an image segmentation task presented in (Loog et al., 2005), which uses image-pixel features and their associated class labels for learning to classify pixels. Their CCA-based method incorporates the class labels of the neighboring pixels as well, which can naturally be expected to yield LDA-like (but possibly more informative) projections. The method can be applied to other forms of, non-image, data by accounting for the spatial class label configuration in the vicinity of every feature vector (Loog et al., 2005).

☆ The work of O. Kursun was supported by Scientific Research Projects Coordination Unit of Istanbul University under the grant YADOP-5323.

* Corresponding author. Tel.: +90 212 473 7070/17913; fax: +90 212 473 7180.
E-mail addresses: okursun@istanbul.edu.tr (O. Kursun), alpaydin@boun.edu.tr (E. Alpaydin), favorov@bme.unc.edu (O.V. Favorov).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.09.025

In this paper, we present another extension of CCA to LDA along with its equivalence proof. The main idea is to transform the class label of a sample such that it is represented, in a distributed manner, by all the samples in that same class. In other words, CCA is asked to produce correlated outputs (projections) for any pair of samples that belong to the same class, which we called WCCCA that stands for within-class coupling CCA. This extension of CCA to LDA has various advantages despite its increased complexity (see Section 4.2 for a detailed list). One important advantage of the WCCCA idea of using samples versus samples, as the two views, is in its ability to perform a form of implicitly-supervised LDA (see Section 5.2) as sometimes the class labels may be embedded in the patterns of the data rather than being explicitly available, for example, in the patterns of spatial and temporal continuity (Becker, 1999; Favorov and Ryder, 2004; Favorov et al., 2003; Borga and Knutsson, 2001; Stone, 1996). Among exemplary applications on such data, the tasks of division of a video into sequences of relevant frames (scenes), segmentation of an image into image regions sharing certain visual characteristics, identifying sequences of acoustic frames belonging to the same word in speech analysis, or finding sequences of base pairs or amino acids belonging to the same protein in biological sequence analysis can be mentioned. In such settings, the use of LDA is uneasy, if not impossible.

The idea of applying CCA or other forms of mutual information maximization models, for example, between the consecutive frames of a video or between the neighboring image patches for finding correlated functions is not a new one (Becker, 1999; Favorov and Ryder, 2004; Favorov et al., 2003; Borga and Knutsson, 2001; Borga, 1998; Stone, 1996; Kording and Konig, 2000; Phillips et al., 1995; Phillips and Singer, 1997). Many of these attempts are inspired by the learning mechanisms hypothesized to be used by neurons in the cerebral cortex. For example, cortical neurons might tune to correlated functions between their own afferent inputs and the lateral inputs they receive from other neurons with different but functionally related afferent inputs. Thus, groups of neurons receiving such different but related afferent inputs can learn to produce correlated outputs under the contextual guidance of each other (Phillips et al., 1995; Phillips and Singer, 1997). However, it is not mathematically justified whether these correlated functions are good for discrimination. Would the covariates found this way be suitable projections for clustering the frames into scenes or for image segmentation? The results of our study show that such a CCA application would be comparable to performing LDA; and as LDA projections maximize the between-class variance and minimize the within-class variance, the covariates found this way would be useful, for example, to cluster the frames into scenes.

This paper is organized as follows. In Sections 2 and 3, we review the CCA and LDA techniques, respectively. In Section 4, we present the WCCCA idea of using CCA on a samples versus samples basis and provide the proof for its equivalence to LDA; we also show that the theoretically derived equivalence holds also practically on a toy example. In Section 4, we also discuss the advantages and disadvantages of this way of performing LDA; and finally, finish this section by showing the nonlinear kernel extension of WCCCA. In Section 5, we present the experimental results on a face database and show that WCCCA can perform the task of LDA even when the images are made into a movie and the class label information is kept only implicitly through the temporal continuity of the identity of the individual seen in contiguous frames. We conclude in Section 6.

2. Canonical correlation analysis (CCA)

Canonical correlation analysis (CCA) is introduced by Hotelling (1936) to describe the linear relation between two multidimensional (or two sets of) variables as the problem of finding basis vectors for each set such that the projections of the two variables on their respective basis vectors are maximally correlated (Hotelling, 1936; Rencher, 1997; Hardoon et al., 2004; Izenman, 2008). These two sets of variables, for example, may correspond to different views of the same semantic object (e.g. audio versus video of a person speaking, two cameras viewing the same object as in binocular vision, text versus links or images in webpages, etc). Let u-dimensional X and v-dimensional Y denote corresponding two sets of real-valued random variables (i.e., $X \in \mathbb{R}^u$ and $Y \in \mathbb{R}^v$), the canonical correlation is defined as:

$$\rho(X;Y) = \sup_{f,g} \operatorname{corr}(f^T X; g^T Y), \tag{1}$$

where, corr(X;Y) stands for Pearson's correlation coefficient. Let u-dimensional column vector $X = x_i$ denote the ith sample (row) of the first view (dataset), v-dimensional column vector $Y = y_i$ denote the ith sample of the second dataset, and N denote the total number of samples. Then, the first dataset $\boldsymbol{D}_1$ is an $N \times u$ matrix that can be expressed as:

$$\boldsymbol{D}_1 = [x_1, x_2, \ldots, x_N]^T \tag{2}$$

and similarly, the $N \times v$ matrix for the second dataset $\boldsymbol{D}_2$ can be written as:

$$\boldsymbol{D}_2 = [y_1, y_2, \ldots, y_N]^T. \tag{3}$$

Assuming that each dataset has zero mean, the total covariance matrix of (X, Y) can be written as a block matrix:

$$\boldsymbol{C}(X,Y) = \mathrm{E}\left\{ \begin{pmatrix} X \\ Y \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}^T \right\} = \begin{bmatrix} \boldsymbol{C}_{XX} & \boldsymbol{C}_{XY} \\ \boldsymbol{C}_{YX} & \boldsymbol{C}_{YY} \end{bmatrix}, \tag{4}$$

where the within-sets covariance matrices are given as:

$$\begin{aligned} \boldsymbol{C}_{XX} &= E\{XX^T\}, \\ \boldsymbol{C}_{YY} &= E\{YY^T\} \end{aligned} \tag{5}$$

and the between-sets covariance matrix is given as:

$$\boldsymbol{C}_{XY} = E\{XY^T\} = \boldsymbol{C}_{YX}^T \tag{6}$$

and now the canonical correlation is the maximum of ρ with respect to f and g:

$$\rho(X;Y) = \sup_{f,g} \frac{f^T \boldsymbol{C}_{XY} g}{\sqrt{f^T \boldsymbol{C}_{XX} f g^T \boldsymbol{C}_{YY} g}}. \tag{7}$$

The problem of finding the orthogonal projections that achieve the top correlations reduces to a generalized eigenproblem, where the projection f (and the projection g can be solved for similarly) corresponds to the top eigenvector of the following (Hardoon et al., 2004):

$$\boldsymbol{C}_{XX}^{-1} \boldsymbol{C}_{XY} \boldsymbol{C}_{YY}^{-1} \boldsymbol{C}_{YX} f = \lambda_{CCA} f \tag{8}$$

and

$$\rho(X;Y) = \sqrt{\lambda_{CCA}}, \tag{9}$$

where λ_{CCA} corresponds to the largest eigenvalue of Eq. (8).

3. Fisher linear discriminant analysis (LDA)

Fisher linear discriminant analysis (LDA) is a variance preserving approach with the goal of finding the optimal linear discriminant function (Fisher, 1936; Rencher, 1997; Raudys and Duin, 1998; Alpaydin, 2004; Izenman, 2008). As opposed to unsupervised methods such as principal component analysis (PCA), independent component analysis (ICA), or the two view counterpart