Available online at www.sciencedirect.com

ScienceDirect

Applied Mathematics and Computation 185 (2007) 272–283

APPLIED MATHEMATICS AND COMPUTATION

www.elsevier.com/locate/amc

Class label versus sample label-based CCA

Tingkai Sun, Songcan Chen *

Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics Nanjing, 210016, PR China

Abstract

When correlating the samples with the corresponding class labels, canonical correlation analysis (CCA) can be used for supervised feature extraction and subsequent classification. Intuitively, different encoding modes for class label can result in different classification performances. However, actually, when the samples in each class share a common class label as in usual cases, a unified formulation of CCA is not only derived naturally, but also more importantly from it, we can get some insight into the shortcoming of the existing feature extraction using CCA for sequent classification: the existing encodings for class label fail to reflect the difference among the samples such as in central region of class and those in mixture overlapping region among classes, consequently resulting in its equivalence to the traditional linear discriminant analysis (LDA) for some commonly-used class-label encodings. To reflect such a difference between the samples, we elaborately design an independent soft label for each sample of each class rather than a common label for all the samples of the same class. A purpose of doing so is to try to promote CCA classification performance. The experiments show that this soft label based CCA is better than or comparable to the original CCA/LDA in terms of the recognition performance.

Keywords: Canonical correlation analysis (CCA); Class label encoding; Separability between classes; Feature extraction

1. Introduction

Feature extraction is widely employed in pattern recognition to obtain the compact representation of the original patterns with limit loss of information. According to the approach in which the class information of the samples is utilized during the feature extraction, the feature extraction methods can be simply classified into unsupervised feature extraction and supervised one. The typical case of the former is principal component analysis (PCA) [1] in which the reconstruction error is minimized while the class information is totally ignored. Two typical cases of supervised feature extraction methods are Fisher linear discriminate analysis (LDA) [1] and canonical correlation analysis (CCA) [2]. In LDA, the class information is incorporated into the within-class scatter matrix and the between-class scatter matrix; in contrast, in CCA, the class information often appears in the form of some numerical label encodings.

CCA was initially developed as a multivariate analysis method [2] and can be seen as the problem of finding basis vectors for two sets of variables such that the correlations between the projections of the variables onto

* Corresponding author.
E-mail address: s.chen@nuaa.edu.cn (S. Chen).

0096-3003/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.amc.2006.06.103

these basis vectors are mutually maximized. For instance, the canonical correlation between image content and the attached texts can be used for Internet image retrieval [3]. Recently CCA gained much attention in various application fields, such as image segment [4], image processing [5], image retrieval [3] and computer vision application [6]. Since CCA is able to directly deal with two sets of variables, so it can directly take one set of variables as class labels, as a result, the class information of samples is explicitly utilized to realize a supervised linear feature extraction and subsequent classification [4,7,8]. The similar method is also straightforwardly used in partial least squares (PLS) for classification [9,10] in which it also involves the correlation between two sets of variables with different constraints.

Utilizing the class label information, CCA can serve as a supervised feature extraction method. Gestel et al. [7] used $\{1, -1\}$ encoding for binary class labels in their kernel CCA. Johansson [11] proposed the following one-of-c label encoding to associate the sample $\boldsymbol{x}$ in class i with the corresponding class label $\boldsymbol{y}^{(i)}$:

$$\boldsymbol{y}^{(i)} = [y_1, y_2, \ldots, y_c]^{\mathrm{T}}, \quad \text{where } y_i = 1 \text{ and } y_j = 0 \quad \text{for all } j \neq i \text{ if } x \in \text{ class } i, i = 1, \ldots, c. \tag{1}$$

This one-of-c encoding is widely employed in multilayer artificial neural network for pattern recognition [13], and it is adopted in kernel CCA for texture classification [8] and also cited in [10]. Note that in [11] the author suggested that this label encoding need not be centered as one modification to ordinary CCA. In [10], the authors cited a different label encoding as follows:

$$\boldsymbol{y}^{(i)} = [y_1, y_2, \ldots, y_{c-1}]^{\mathrm{T}}, \quad \text{where} \begin{cases} y_i = 1, y_j = 0 \quad \text{for all } j \neq i \text{ if } \boldsymbol{x} \in \text{ class } i, i = 1, \ldots, c-1, \\ \boldsymbol{y}^{(i)} = [0, 0, \ldots, 0]^{\mathrm{T}} \quad \text{if } \boldsymbol{x} \in \text{ class } c. \end{cases} \tag{2}$$

Here we name it as $c - 1$ label encoding for convenience. In [9] the authors proposed a slightly different $c - 1$ label encoding that encodes class c as $[1, 1, \ldots, 1]^{\mathrm{T}}$ and keeps others unchanged for discriminant using PLS. Loog et al. [4] took the contextual label information of every pixel into account and developed an elaborate label encoding based on $\{1, 0\}$ and applied it to the medical image segmentation, and this label encoding can be seen as an extension of one-of-c encoding, specifically, it encodes each pixel in a local neighborhood around some pixel in terms of different object classes and then concatenates them into a vector.

Various label encodings have been used in CCA or its variants. Intuitively, different encoding modes for class label can result in different classification performances. In this study, we derive a unified formulation of the existing CCA where the samples in each class share a common class label as in usual encodings in pattern recognition. From such a unified formulation we can get some insight into the shortcoming of the existing CCA for feature extraction: the existing class encodings fail to reflect the difference among the samples such as in the central region of one class and those in (or near) the overlapping region among classes, consequently resulting in its equivalence to the traditional linear discriminant analysis (LDA) for some class-label encodings. It is well known that according to the theory of support vector machine (SVM) [18], the samples near (or in) the overlapping region among classes are more possible to be misclassified than those in central region of that class, so they are more important for the design of the classifier and possibly become the support vectors, which are significant points for SVM.

Then a question arises naturally: are these samples locating near the overlapping region among classes also important for feature extraction such as CCA? To our best knowledge, there is no special research about this topic. In the existing class label based CCA where each class share one common label encoding, the samples lying in the overlapping region are treated in the same way as those in non-overlapping regions. To reflect the difference between samples, we take advantages of CCA that can simultaneously deal with two sets of data and elaborately design the soft label encoding for each sample rather than a class. For one class, each sample possesses its private class label that reflects the class distribution in some degree. Then the samples near overlapping region should be assigned its own private class label. The purpose of doing so aims to examine the impact of this label encoding on classification performance, and emphasizes its own different characteristic from LDA (it reduces to LDA when each class has just a single label). The experiments show that this soft label based CCA is better than or comparable to the original CCA/LDA in terms of the recognition performance.

The remaining of this paper is organized as follows: the formulas of CCA and LDA are firstly presented in Section 2; in Section 3 we derive a unified formulation of the existing class label based CCA; from this unified formulation, we can get some insight into the shortcoming of the existing feature extraction using CCA in

Section 4; in Section 5, soft label encoding is proposed; the experiments are described in Section 6 to evaluate the performance of soft label based CCA; finally in Section 7, the paper is summarized and further research topics are given.

2. Review of class label based CCA and LDA

2.1. Description of CCA [12]

Given N pairs of samples $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, \ldots N$, and their means $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$, where $\boldsymbol{x}_i \in R^p$, $\boldsymbol{y}_i \in R^q$, CCA aims to find pairs of projection directions φ and ϕ that maximize the correlation between the random variable $\varphi^{\mathrm{T}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$ and $\phi^{\mathrm{T}}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})$, $i = 1, \ldots, N$, i.e., to maximize

$$r = \frac{\varphi^{\mathrm{T}} \boldsymbol{X} \boldsymbol{Y}^{\mathrm{T}} \phi}{\sqrt{\varphi^{\mathrm{T}} \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}} \varphi} \sqrt{\phi^{\mathrm{T}} \boldsymbol{Y} \boldsymbol{Y}^{\mathrm{T}} \phi}}. \tag{3}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1 - \bar{\boldsymbol{x}}, \ldots, \boldsymbol{x}_N - \bar{\boldsymbol{x}}]$, $Y = [\boldsymbol{y}_1 - \bar{\boldsymbol{y}}, \ldots, \boldsymbol{y}_N - \bar{\boldsymbol{y}}]$, and the symbol T denote the transpose. This can be expressed as the following equivalent optimization problem:

$$\begin{aligned} &\max \varphi^{\mathrm{T}} \boldsymbol{X} \boldsymbol{Y}^{\mathrm{T}} \phi, \\ &\text{s.t.} \varphi^{\mathrm{T}} \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}} \varphi = 1, \phi^{\mathrm{T}} \boldsymbol{Y} \boldsymbol{Y}^{\mathrm{T}} \phi = 1. \end{aligned} \tag{4}$$

Solving this optimization problem, we can obtain the following decoupled generalized eigenvalue equations:

$$\begin{cases} \boldsymbol{X} \boldsymbol{Y}^{\mathrm{T}} (\boldsymbol{Y} \boldsymbol{Y}^{\mathrm{T}})^{-1} \boldsymbol{Y} \boldsymbol{X}^{\mathrm{T}} \varphi = \lambda \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}} \varphi, \\ \boldsymbol{Y} \boldsymbol{X}^{\mathrm{T}} (\boldsymbol{X} \boldsymbol{X}^{\mathrm{T}})^{-1} \boldsymbol{X} \boldsymbol{Y}^{\mathrm{T}} \phi = \lambda \boldsymbol{Y} \boldsymbol{Y}^{\mathrm{T}} \phi, \end{cases} \tag{5}$$

where the eigenvalue λ exactly equals to r^2. From (5), it is observed that when small sample size problem, i.e., $N \ll p$ or $N \ll q$, is encountered, matrix $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ or $\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$ is singular and the involved inverting becomes infeasible. Moreover, the maximal number of the corresponding combinations (λ, φ, ϕ) in (5) is not more than $\min(p, q)$. Correlating samples, say, $\boldsymbol{X}$, with their class labels, say, $\boldsymbol{Y}$, and letting $\boldsymbol{W} = [\varphi_1, \varphi_2, \ldots, \varphi_d]$ denote the eigenvectors corresponding to the first d largest eigenvalues, the extracted features $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$ can be used for classification.

2.2. Description of LDA [1]

We will give a brief description of LDA to help to clarify the relationship between LDA and CCA using some commonly-used class-label encodings hereafter in this paper. Given N samples $\boldsymbol{x}_i$, $i = 1, \ldots, N$, which come from c classes. LDA aims to find discriminative vector $\boldsymbol{w}$ for these samples such that the ratio of the between-class scatter to the within-class scatter of the sample projections onto the basis vector is maximized, i.e., maximize the following criterion defined in (6):

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{w}}, \tag{6}$$

where the between-class scatter matrix, $\boldsymbol{S}_b$, and the within-class scatter matrix, $\boldsymbol{S}_w$, are defined respectively by

$$\boldsymbol{S}_b = \sum_{i=1}^{c} n_i \left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right) \left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)^{\mathrm{T}} = \boldsymbol{H}_b \boldsymbol{D} \boldsymbol{H}_b^{\mathrm{T}} \tag{7}$$

and

$$\boldsymbol{S}_w = \sum_{i=1}^{c} \sum_{\boldsymbol{x} \in \omega_i} \left(\boldsymbol{x} - \bar{\boldsymbol{x}}^{(i)}\right) \left(\boldsymbol{x} - \bar{\boldsymbol{x}}^{(i)}\right)^{\mathrm{T}}, \tag{8}$$

where $\boldsymbol{H}_b = \left[n_1\left(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}\right), \ldots, n_c\left(\bar{\boldsymbol{x}}^{(c)} - \bar{\boldsymbol{x}}\right)\right]$, $\boldsymbol{D} = \mathrm{diag}(1/n_1, \ldots, 1/n_c)$. Solving this optimization problem (6), we can obtain the following generalized eigenvalue equation:

$$\boldsymbol{S}_b\boldsymbol{w} = \lambda_{\mathrm{LDA}}\boldsymbol{S}_w\boldsymbol{w}, \tag{9}$$

where the eigenvalue λ_{LDA} exactly equals to the value of objection function (6). It is worth noting that the maximal number of discriminative vectors, $\boldsymbol{w}$s, is at most $c-1$ due to the rank limit of $\boldsymbol{S}_b$.

3. A unified formulation

Generally in class label based CCA, one common class label is assigned to each class, so the terms $\boldsymbol{XY}^{\mathrm{T}}$ and $\boldsymbol{YY}^{\mathrm{T}}$ can both be simplified through simple mathematical manipulations. This will give rise to a unified formulation as follows.

Let $\boldsymbol{X} = [\boldsymbol{x}_1 - \bar{\boldsymbol{x}}, \ldots, \boldsymbol{x}_N - \bar{\boldsymbol{x}}]$ and $\boldsymbol{Y} = [\boldsymbol{y}_1 - \bar{\boldsymbol{y}}, \ldots, \boldsymbol{y}_N - \bar{\boldsymbol{y}}]$ denote the centered samples set and class labels set, where $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ denote total mean, $\bar{\boldsymbol{x}}^{(i)}$ and $\bar{\boldsymbol{y}}^{(i)}$ mean of samples/labels in the ith class, respectively. According to the above fact that the samples of the same class share a common class label, the covariance matrix $\boldsymbol{XY}^{\mathrm{T}}$ can be derivable as

$$\begin{aligned}\boldsymbol{XY}^{\mathrm{T}} &= \sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^{\mathrm{T}} = \sum_{i=1}^{c}\sum_{\boldsymbol{x},\boldsymbol{y}\in\omega_i}\left(\boldsymbol{x} - \bar{\boldsymbol{x}}^{(i)} + \bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{y} - \bar{\boldsymbol{y}}^{(i)} + \bar{\boldsymbol{y}}^{(i)} - \bar{\boldsymbol{y}}\right)^{\mathrm{T}} \\ &= \sum_{i=1}^{c}\sum_{\boldsymbol{x},\boldsymbol{y}\in\omega_i}\left(\boldsymbol{x} - \bar{\boldsymbol{x}}^{(i)}\right)\left(\boldsymbol{y} - \bar{\boldsymbol{y}}^{(i)}\right)^{\mathrm{T}} + \sum_{i=1}^{c}\sum_{\boldsymbol{x},\boldsymbol{y}\in\omega_i}\left(\boldsymbol{x} - \bar{\boldsymbol{x}}^{(i)}\right)\left(\bar{\boldsymbol{y}}^{(i)} - \bar{\boldsymbol{y}}\right)^{\mathrm{T}} \\ &\quad + \sum_{i=1}^{c}\sum_{\boldsymbol{x},\boldsymbol{y}\in\omega_i}\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{y} - \bar{\boldsymbol{y}}^{(i)}\right)^{\mathrm{T}} + \sum_{i=1}^{c}\sum_{\boldsymbol{x},\boldsymbol{y}\in\omega_i}\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{y}}^{(i)} - \bar{\boldsymbol{y}}\right)^{\mathrm{T}},\end{aligned} \tag{10}$$

where the first and the third term both equal zero due to such a fact that in class i, all labels $\boldsymbol{y}$ are identical so $\boldsymbol{y} - \bar{\boldsymbol{y}}^{(i)}$ is zero. Similarly, the second term is zero due to that

$$\sum_{i=1}^{c}\sum_{\boldsymbol{x},\boldsymbol{y}\in w_i}\left(\boldsymbol{x} - \bar{\boldsymbol{x}}^{(i)}\right)\left(\bar{\boldsymbol{y}}^{(i)} - \bar{\boldsymbol{y}}\right)^{\mathrm{T}} = \sum_{i=1}^{c}\left(\sum_{\boldsymbol{x},\boldsymbol{y}\in w_i}\left(\boldsymbol{x} - \bar{\boldsymbol{x}}^{(i)}\right)\right)\left(\bar{\boldsymbol{y}}^{(i)} - \bar{\boldsymbol{y}}\right)^{\mathrm{T}} = \sum_{i=1}^{c}0\cdot\left(\bar{\boldsymbol{y}}^{(i)} - \bar{\boldsymbol{y}}\right)^{\mathrm{T}} = 0. \tag{11}$$

So finally only the fourth term is kept and hence (10) can be simply rewritten as

$$\boldsymbol{XY}^{\mathrm{T}} = \sum_{i=1}^{c}\sum_{\boldsymbol{x},\boldsymbol{y}\in\omega_i}\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{y}}^{(i)} - \bar{\boldsymbol{y}}\right)^{\mathrm{T}} = \sum_{i=1}^{c}n_i\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{y}}^{(i)} - \bar{\boldsymbol{y}}\right)^{\mathrm{T}} = \sum_{i=1}^{c}n_i\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)\boldsymbol{y}^{(i)\mathrm{T}} = \boldsymbol{H}_b\tilde{\boldsymbol{Y}}^{\mathrm{T}}, \tag{12}$$

where n_i denotes the sample size of class i, $\tilde{\boldsymbol{Y}} = [\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(c)}]$ denotes the class label matrix. The third "=" holds due to the facts that $\sum_{i=1}^{c}n_i\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right) = 0$ and $\bar{\boldsymbol{y}}^{(i)} = \boldsymbol{y}^{(i)}$, where $\boldsymbol{y}^{(i)}$ denotes the uniform label for samples belonging to class i.

On the other hand, for the term $\boldsymbol{YY}^{\mathrm{T}}$ we have:

$$\boldsymbol{YY}^{\mathrm{T}} = \sum_{i=1}^{c}n_i\left(\boldsymbol{y}^{(i)} - \bar{\boldsymbol{y}}\right)\left(\boldsymbol{y}^{(i)} - \bar{\boldsymbol{y}}\right)^{\mathrm{T}}. \tag{13}$$

From (12), we find that the covariance matrix $\boldsymbol{XY}^{\mathrm{T}}$ actually represents some characteristic of separation between classes, i.e., $\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}$ or $\boldsymbol{H}_b$, which is just the constituent of the between-class scatter matrix $\boldsymbol{S}_b$. This indicates that once the samples in each class share a common class label, an association may exist between CCA and LDA. Although the appearance of class label matrix $\tilde{\boldsymbol{Y}}$ makes this association not so immediate, we can still manifest this association by a few special cases, i.e., the usually used label encodings, as will discussed in Section 4.

4. Special cases

In Section 4.1, we study CCA using some commonly-used class-label encodings as the special cases of this unified formulation to clarify the possible association between CCA and LDA, and in Section 4.2 we gain some insight into this type of class label based CCA and point out the disadvantage of CCA using some commonly-used class-label encodings.

In the first part, the discussion focuses on the employment of these label encodings in the following generalized eigenvalue equation of CCA:

$$\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}})^{-1}\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}\varphi = \lambda\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\varphi, \tag{14}$$

The right side, $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$, is just total scatter matrix and equals to $\boldsymbol{S}_b + \boldsymbol{S}_w$, and all computation associated with label encodings, $\boldsymbol{Y}$, are on the left side of (14). Hereafter we focus on the left size of (14) to analyze the employment of label encodings.

4.1. Analysis

(1) one-of-c encoding [11,10]
This label encoding has been depicted in (1). When this label encoding is proposed, the author [11] noted that it need not to be centered during CCA as a modification to ordinary CCA. It seems that doing so violates the definition of CCA and undermines its mathematics foundation. In fact, in this case, $\tilde{\boldsymbol{Y}}$ is an identity matrix so $\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$ becomes just $\boldsymbol{H}_b$ and it is easy to obtain $(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}})^{-1} = \mathrm{diag}(1/n_1,\ldots,1/n_c) = \boldsymbol{D}$, so we can rewritten (14) as

$$\boldsymbol{S}_b\varphi = \lambda(\boldsymbol{S}_b + \boldsymbol{S}_w)\varphi \tag{15}$$

and it can also be derivable as

$$\boldsymbol{S}_b\varphi = \frac{\lambda}{1-\lambda}\boldsymbol{S}_w\varphi \tag{16}$$

which exactly represents its equivalence to LDA by recalling its description formulation (9).When one-of-c encoding is centered by the definition of CCA, $\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$ still equals to $\boldsymbol{H}_b$, which can be seen from (12), yet, $(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}})^{-1}$ becomes a bit complex, but doing so does not prevent (14) from becoming the form of (16). For details proof, please refer to [10].

(2) $c-1$ label encoding [10]
For the $c-1$ label encoding described in (2), we have

Proposition 1. *When $c-1$ label encoding defined in (2) is employed, the following two relationships hold:*

$$\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}} = \left[n_1\left(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}\right),\ldots,n_{c-1}\left(\bar{\boldsymbol{x}}^{(c-1)} - \bar{\boldsymbol{x}}\right)\right], \tag{17}$$

$$\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}})^{-1}\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{S}_b. \tag{18}$$

Proof

$$\begin{aligned}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}} &= \sum_{i=1}^{c} n_i\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)\boldsymbol{y}^{(i)\mathrm{T}} = \left[n_1\left(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}\right),\ldots,n_c\left(\bar{\boldsymbol{x}}^{(c)} - \bar{\boldsymbol{x}}\right)\right]\cdot\tilde{\boldsymbol{Y}}^{\mathrm{T}}\\ &= \left[n_1\left(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}\right),\ldots,n_c\left(\bar{\boldsymbol{x}}^{(c)} - \bar{\boldsymbol{x}}\right)\right]\cdot\begin{bmatrix}\boldsymbol{I}_{c-1}\\ \boldsymbol{0}\end{bmatrix} = \left[n_1\left(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}\right),\ldots,n_{c-1}\left(\bar{\boldsymbol{x}}^{(c-1)} - \bar{\boldsymbol{x}}\right)\right].\end{aligned}$$

The inverse of $\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$ can be written as [10]:

$$(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}})^{-1} = \frac{1}{n_c}\boldsymbol{1}_n\boldsymbol{1}_n^{\mathrm{T}} + \mathrm{diag}(1/n_1,\ldots,1/n_{c-1}).$$

So we have

$$\begin{aligned}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}})^{-1}\boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}} &= \frac{1}{n_c}\left[-n_c\left(\bar{\boldsymbol{x}}^{(c)} - \bar{\boldsymbol{x}}\right)\right]\cdot\left[-n_c\left(\bar{\boldsymbol{x}}^{(c)} - \bar{\boldsymbol{x}}\right)^{\mathrm{T}}\right] + \sum_{i=1}^{c-1} n_i\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)^{\mathrm{T}}\\ &= \sum_{i=1}^{c} n_i\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}\right)^{\mathrm{T}} = \boldsymbol{S}_b.\end{aligned}$$

As can be seen in the proof, $\boldsymbol{XY}^{\mathrm{T}}$ now becomes an altered version of $\boldsymbol{H}_b$, from which the contribution of class c is excluded. Similarly (16) can be obtained. □

(3) Gestel et al.'s scheme $\{1,-1\}$ [7]

In their paper [7], binary classes are encoded as 1 and -1, respectively. In this case, we can easily obtain $\boldsymbol{XY}^{\mathrm{T}} = \frac{2n_1n_2}{N}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ and $\boldsymbol{YY}^{\mathrm{T}} = \frac{4n_1n_2}{N}$, where N is sum of n_1 and n_2, and $\boldsymbol{XY}^{\mathrm{T}}(\boldsymbol{YY}^{\mathrm{T}})^{-1}\boldsymbol{YX}^{\mathrm{T}} = \frac{n_1n_2}{N}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^{\mathrm{T}}$ holds, which is exactly the between-class scatter matrix, $\boldsymbol{S_b}$, in the case of binary classes. It is easy to see that the matrices $\boldsymbol{XY}^{\mathrm{T}}$ and $\boldsymbol{YY}^{\mathrm{T}}$ play a role of constituent of between-class matrix and its corresponding modified factor, i.e., constant $4n_1n_2/N$, respectively. In this case, (16) still holds.

4.2. Review and insight

Up to now, we have analyzed four kinds of label encodings and surprisingly found that when these label encodings are used: (a) $\boldsymbol{XY}^{\mathrm{T}}$ is just the constituent of the between-class scatter matrix, $\boldsymbol{S}_b$; and (b) the left hand side of CCA Eq. (14), $\boldsymbol{XY}^{\mathrm{T}}(\boldsymbol{YY}^{\mathrm{T}})^{-1}\boldsymbol{YX}^{\mathrm{T}}$ is just $\boldsymbol{S}_b$, in which some important discriminant information is contained. So (14) can be rewritten as (15), and can also be derivable as $\boldsymbol{S}_b\varphi = \frac{\lambda}{1-\lambda}\boldsymbol{S}_w\varphi$, which exactly represents its equivalence to LDA. In summary, we propose a unified formulation of class label based CCA and interpret this unified formulation in terms of separability between classes, and this indicates a possible association of CCA with LDA. We further clarify this implicit, not immediate association by analysis of CCA using some commonly-used class-label encodings as the special cases of this unified formulation.

In fact, the discovery of this relationship between CCA and LDA is not occasional. As long as we employ the class label encoding that is assigned to one class rather than one sample, we can always rewritten $\boldsymbol{XY}^{\mathrm{T}}$ as $\boldsymbol{H}_b$ or its altered version (e.g., see proof of Proposition 1), which is the constituent of the between-class scatter matrix $\boldsymbol{S}_b$, resulting in an equivalent or similar relationship between CCA and LDA.

It is well known that in classical LDA the essential discriminant information is contained in between-class scatter matrix $\boldsymbol{S}_b$, where class mean $\bar{\boldsymbol{x}}^{(i)}$ acts as the prototype of class i, and $\boldsymbol{w}^{\mathrm{T}}(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}})$ or $\boldsymbol{w}^{\mathrm{T}}(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}}^{(j)})$ can be deemed as the metric of the separability between classes, where $\boldsymbol{w}$ denotes the discriminant vector. Then a question will naturally arise: is class mean $\bar{\boldsymbol{x}}^{(i)}$ a "good" prototype of the samples in class i? In other words, is the class mean $\bar{\boldsymbol{x}}^{(i)}$ representative for the samples in class i? Generally we take it for granted that the data distribution is Gaussian, i.e, a hyper-sphere can be used for a fair good approximation of the data distribution, so the centroid of this hyper-sphere can be a representative prototype. However, the assumption of Gaussian distribution is not always the case. In [17] the authors illustrate an example of non-Gaussian distribution. On the other hand, for linearly separable case, $\boldsymbol{w}^{\mathrm{T}}(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}})$ may be a reasonable metric for separability between classes. However, in linearly non-separable case, $\boldsymbol{w}^{\mathrm{T}}(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}})$ is unnecessarily a good metric, especially for the cases where the distribution areas of different classes intersect heavily. In these two kinds of cases, i.e., non-Gaussian distribution and linearly non-separable cases, if the samples in each class still share a common class label as in usual cases, then class label based CCA will becomes classical LDA or its variant, then the advantage of CCA that can changes its supervised mode of feature extraction flexibly will be lost, comparing to LDA. If we pursue the reason of this phenomenon, we can find the distribution of the samples is not considered by the existing CCA using common class label in one class. In other words, the samples in (or near) overlapping region among classes are deemed to be as important as the samples in the central areas of that class. This sharply conflicts with the idea of support vectors machine (SVM) which emphasizes that the samples close to the class boundary are more important than others [18], because these samples are more easily misclassified and thus probably possess more weight in the design of the classifier. To overcome this shortcoming of the existing class label based CCA, we elaborately design a soft label for each sample rather than a class according to the region it lies in to reflect the difference between samples in the same class, as will be discussed in Section 5.

5. Soft label encoding

Up to now, we have derived a unified formulation of CCA using some commonly-used class-label encodings. In doing so we do not aim to deduce some new class label encodings from this unified formulation, instead, we aim to gain some insight of the existing CCA for feature extraction and point out its shortcomings. The label encodings mentioned above can be named as hard class labels because the samples of each class share one common label encoding that crisply indicates the class membership of each sample. For example, the class label of the sample coming from class 1 can be encoded as [1, 0, ..., 0], which can be interpreted that degree of membership of sample to class 1 is 100%, whereas 0% for the other classes. This hard class label is assigned to the samples coming from one class wherever they locate, and the impact of sample distribution on the feature extraction for classification is not considered. However, for machine learning problems such as SVM, it is just the samples near the overlapping region among classes that play more important roles than those in the central region of each class. So it is necessary to embody the difference between the samples in the process of feature extraction. Specially, the main difference between samples near the overlapping regions is their class information, and it is deemed that half of the class information for one sample is hidden in its neighbors [14], so one convenient way to describe the class distribution is to observe the class memberships of the neighbors of each sample. To this end, the original hard class label can reasonably be modified into soft one and embed to CCA to reflect the impact of the sample distribution on the feature extraction. Here, we can utilize fuzzy k-nearest neighbor method in [15] to yield a soft class label encoding for each example, specifically, for training sample $\boldsymbol{x}_i$ coming from class j, its corresponding soft class label $\tilde{\boldsymbol{y}}_i$ can be defined by

$$\tilde{\boldsymbol{y}}_i = [y_{i1}, y_{i2}, \ldots, y_{ic}]^{\mathrm{T}}, \tag{19}$$

where $y_{ij} = 0.51 + 0.49n_{ij}/k$ and $y_{im} = 0.49n_{im}/k$, $m \neq j$, n_{ij} and n_{im} denote the number of $\boldsymbol{x}_i$'s neighbors coming from class j and m, respectively, and k the parameter of k-nearest neighbor method (k–NN). Let us consider two extreme cases: if $n_{ij} = k$, it means that all k-nearest neighbors of $\boldsymbol{x}_i$ come from the same class as $\boldsymbol{x}_i$, resulting in a hard class label $[0, \ldots, 1, \ldots, 0]^{\mathrm{T}}$ where only the jth entry is one; in contrast, when $n_{ij} = 0$, resulting in a label like $[y_{i1}, \ldots, 0.51, \ldots, y_{ic}]^{\mathrm{T}}$ where the jth entry, y_{ij}, is 0.51 and $0 \leqslant y_{im} \leqslant 0.49$ otherwise. If the latter case happens, it means that this sample exhibits a high degree of membership to several classes. In this way, the significance of the samples in mixture region between classes to the classifier is emphasized. So one-of-c label encoding becomes one special case of this soft label encoding, and the index of the dominated entry (i.e., 0.51 or more) in $\tilde{\boldsymbol{y}}_i$ still indicates its original class membership. In real application, the samples coming from the same class may yield different class labels, so the proposed soft label distinguishes itself obviously from one-of-c label and others such as those mentioned above. We anticipate that CCA incorporating this soft class label (referred as CCA-s later) could enhance its recognition performance.

6. Experiments and analysis

To evaluate the performance of proposed soft label method, some classification experiments are performed on ORL human face dataset (free available at: <http://www.uk.research.att.com/facedatabase.html>) and 10 UCI machine learning repository (free available at: <http://www.ics.uci.edu/~mlearn/MLSummary.html>).

6.1. Experimental setup

The respective name of the dataset, its class number c and the sample dimensionality d are tabulated in Table 1. The experiment of common CCA that employs one-of-c label encoding (still referred as CCA later) is also performed. The experiments are randomly repeated 10 times for ORL and 100 times for UCI. All contrastive experiments are based on the identical partition of the training/test set for each dataset.

For ORL dataset, the training set size 200 is far less than the sample dimension 112 × 92 (=10304), small sample size problem occurs for both CCA and CCA-s. To overcome this problem, both of CCA and CCA-s are preceded by PCA and then performed in the transformed 60-dimensional PCA subspace. All of the finally obtained discriminative vectors are used to extract features for recognition by the nearest neighbor classifier. On each dataset the parameter k is determined by cross-validation based on the dataset partition with (approximately) equal sizes.

Table 1
Some attributes of the dataset and the comparison of the recognition accuracies (%)

Dataset	*c*	*d*	Recognition accuracy		*dev*
			CCA/LDA	CCA-s/*k*	
ORL	40	112 × 92	95.65	96.50/17	0.7841
Balance	3	4	87.96	**90.16**/30	0.4434
Bupa	2	6	58.93	59.79/25	0.4480
E. coli	6	6	80.23	**82.12**/12	0.4181
Glass	6	9	56.97	**59.00**/3	0.4875
Lenses	3	4	73.92	**75.15**/5	0.6816
Sonar	2	60	69.68	70.58/30	0.4570
Thyroid	3	5	93.61	**95.46**/15	0.3982
Vehicle	4	18	73.48	71.78/95	0.6223
Wine	3	13	98.14	98.17/80	0.6386
Wpbc	2	32	66.25	**67.62**/15	0.4910

6.2. Result analysis

The recognition results on all datasets are given in Table 1, from which we can observe that: (a) on 9 datasets of total 11 ones, CCA-s outperforms CCA, especially on ORL, balance, *E. coli*, glass, lenses, thyroid and wpbc datasets, CCA-s outperforms CCA more than one percent; (b) on wine datasets, the former is comparable to the latter; (c) on vehicle dataset, the former is outperformed by the latter; (d) for some datasets, the parameter k may exceed the number of training samples in some classes. For instance, for ORL dataset, $k = 17 > 5$ (the number of training samples per class). So when the soft label is computed by means of k–NN, the neighbors of each sample may inevitably share multiple classes, consequently resulting in a soft label totally distinct from one-of-c. Fig. 1 illustrates the change of the recognition accuracy w.r.t. the parameter k on ORL, *E. coli*, thyroid, wine and vehicle datasets, where k exhausts a large range of possible values from 1 to 100 for all dataset except for wine dataset due to the limitation of the training sample size. These representative datasets are selected because on the first three, wine and vehicle datasets CCA-s outperforms CCA, is comparable to CCA and is outperformed by CCA, respectively. Meanwhile the recognition accuracy

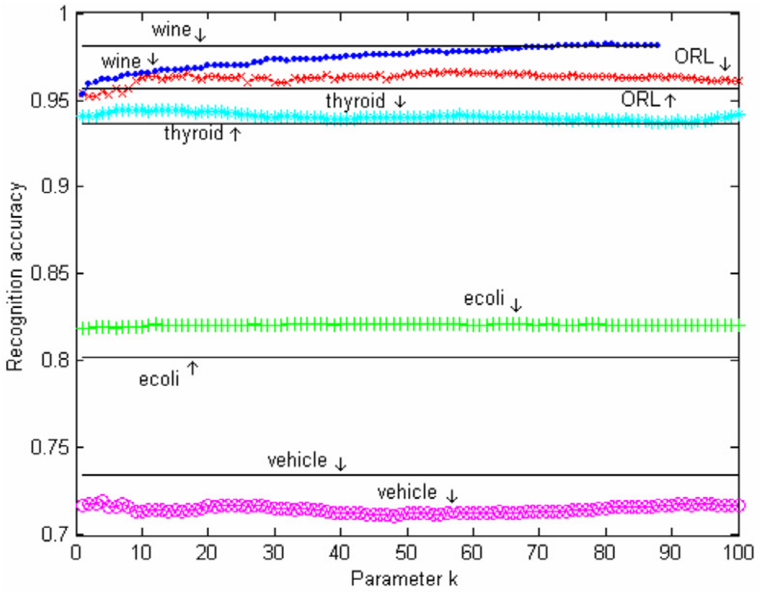

Fig. 1. The change of recognition accuracy w.r.t. the parameter k. The horizontal solid lines denote the recognition accuracies using CCA, which are independent of k.

Table 2
The distribution of the most possibly misclassified samples

Dataset	Class #	Misclassified by CCA	Misclassified by CCA-s
ORL	36, 38, 40	0.6, 0.3, 1.2	0.5, 0.1, 1.0
E. coli	1, 2, 4	4.24, 10.08, 8.98	4.44, 10.4, 8.39
Thyroid	1, 2, 3	2.4, 2.15, 1.18	2, 1.87, 1.16
Wine	1, 2, 3	0.24, 1.31, 0.33	0.13, 1.18, 0.34
Vehicle	1, 2, 4	44.59, 47.38, 10.86	45.94, 48.83, 12.01

of CCA on these datasets is also shown as lines due to its independence of k. On all datasets but wine, the slight fluctuation of the curves of the recognition accuracy w.r.t. the parameter k reflects the robustness of CCA-s in some degree.

Table 2 shows the distribution of the most possibly misclassified classes and the corresponding mean misclassification frequency on ORL, *E. coli*, thyroid, wine and vehicle datasets. On ORL, *E. coli*, thyroid and wine datasets, the mean misclassification frequency of CCA-s are lower than those of CCA for most of the selected

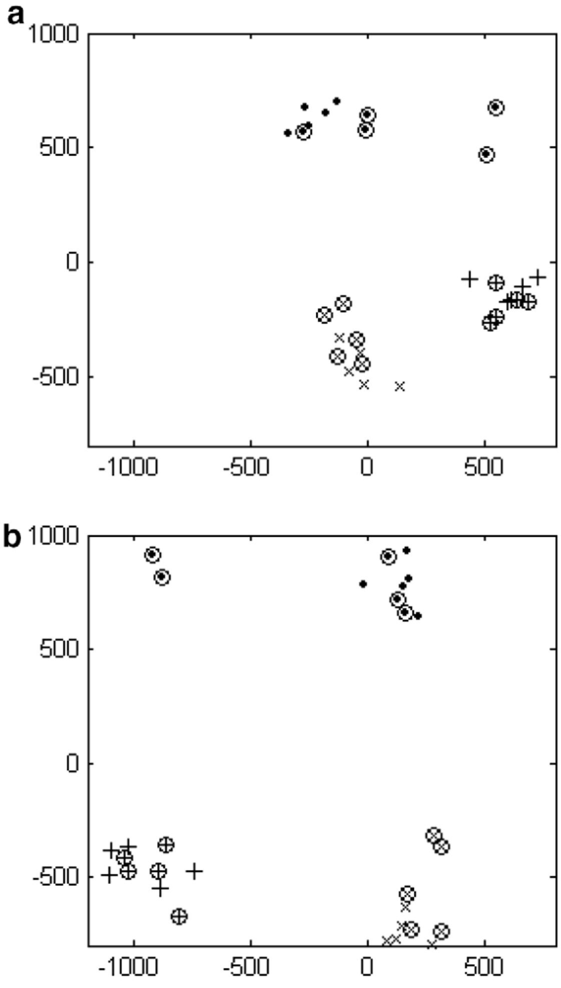

Fig. 2. Two-dimensional visualization of ORL data with dimensionality reduced by (a) CCA and (b) CCA-s. The symbols dot, (×) and (+) denote training samples in class 36, 38 and 40, respectively, and circled dot, (×) and (+) denote testing samples in class 36, 38 and 40, respectively.

classes, however, on vehicle dataset, CCA-s is higher than CCA. These comparisons are approximately consistent to the corresponding differences of the recognition accuracies between CCA and CCA-s.

To visually illustrate the difference between the discriminant performance of features extracted respectively by CCA and CCA-s, we only select the samples in class 36, 38 and 40 of ORL dataset, which are most possibly misclassified, and project them onto two eigenvectors corresponding to the first two largest eigenvalues of CCA and CCA-s to get their plots, as shown in Fig. 2(a) and (b), respectively. From the figure, we can find that the projected between-class scatter by CCA-s (Fig. 2(b)) seems larger than that by CCA (Fig. 2(a)), though the projected within-class scatter seems no obvious difference between them. In this set of contrastive experiments, 2, 1, and 3 samples belonging to class 36, 38 and 40, respectively, are misclassified by CCA, resulting in recognition accuracy of just 95.0%, in contrast, using proposed CCA-s, the numbers of the misclassification for the same classes are reduced to 1, 0 and 2, respectively, resulting in recognition accuracy of 97.5%. It seems that the improvement of recognition accuracy benefits from larger between-class scatter brought by CCA-s.

The separability criterion between classes is generally taken as $J = |\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{W}|/|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{W}|$ [1] or $J = \mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{W})/\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{W})$ [16], where $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$ denote between-class scatter matrix and within-class scatter matrix, respectively, $\boldsymbol{W}$ the projective matrix with the discriminant vectors $\boldsymbol{w}$ as its columns, and $\mathrm{tr}(\cdot)$ the trace of a matrix. Fig. 3 illustrates $\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{W})$ and $\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{W})$ for ORL data in CCA and CCA-s in 10 times of experiments. It is obvious that $\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{W})$ of CCA-s is much greater than that of CCA/LDA, yet $\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{W})$ of CCA-s is only slightly greater than that of CCA/LDA. It is just the enhancement of the separability criterion of CCA-s against CCA/LDA that promotes the recognition performance evidently. In other words, in this case CCA-s can extract more discriminative features for classification by incorporating soft class label into CCA.

On the other hand, to find the reasons why CCA-s is outperformed by CCA on a few datasets, we define the following variable *dev* as

$$dev = \frac{1}{c}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{\omega_i}\|\tilde{y} - e_i\|_1, \tag{20}$$

to measure the deviation degree of this soft label from the "standard" one-of-c label, where $\tilde{y}$ and $\boldsymbol{e}_i$ denote the soft label corresponding to the samples in class i and the one-of-c label encoding corresponding to class i,

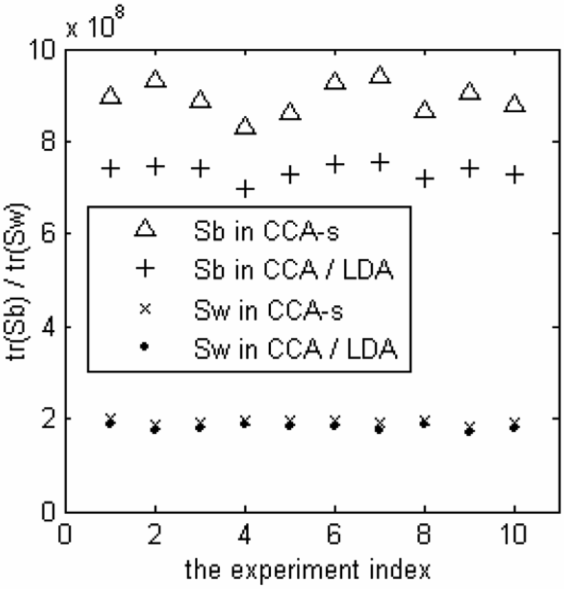

Fig. 3. The between-class scatter ($\boldsymbol{S}_b$) and within-class scatter ($\boldsymbol{S}_w$) for ORL data with dimensionality reduced by CCA and CCA-s in 10 times of experiments.

respectively, and $\|\cdot\|_1$ denotes L_1-norm of a vector. The larger *dev* means more "fuzzy" class information is introduced into the soft label encodings; and *dev* being zero means the soft labels totally degenerates into one-of-c label encodings. The mean values of *dev* for every dataset in the repeated experiments are also tabulated in Table 1. We can observe that: (a) on 7 of total 11 datasets, the devs are all less than 0.5, correspondingly, CCA-s outperforms CCA on these datasets; (b) of other four datasets where the *dev* is greater than 0.5, CCA-s is outperformed by, or comparable to CCA on vehicle and wine datasets, respectively, except for datasets of ORL and lenses. It seems that the moderate introduction of "fuzzy" class information by the proposed soft label encoding will help to improve the recognition performance. However, the exceptional cases on ORL and lenses datasets seem to indicate that the effect of the soft label encoding is data dependent.

7. Discussion and some further study topics

In this paper, we derive first a unified formulation of CCA when the samples in each class share a common class label as in usual cases, then from this unified formulation of CCA, we can get some insight into the shortcoming of the existing feature extraction using CCA for sequent classification: the existing encodings for class fail to reflect the difference among the samples such as in central region of class and those in mixture overlapping region among classes, consequently resulting in its equivalence to the traditional linear discriminant analysis (LDA) for some commonly-used class-label encodings. To reflect such a difference between the samples and its impact on feature extraction, we elaborately design an independent soft label for each sample of each class rather than a common label for all the samples of the same class. In doing so we try to break the performance limitation imposed on CCA due to its equivalence to LDA and to promote its classification performance. The experiments show that this soft label based CCA is better than or comparable to the original CCA/LDA in terms of the recognition performance.

Although in this paper all of the mathematical deductions and the experiments are based on the linear feature extraction, in fact, they can be easily generalized to the corresponding kernel versions via the kernel trick [18]. In fact, some label encodings, e.g., one-of-c and binary scalar encoding, have been successfully applied in kernel CCA [7,8]. On the other hand, PLS for discriminant also involves the same correlation between samples and the corresponding class labels, however, with different constraints. So our proposed soft class label based CCA can also generalized to kernel CCA and PLS, which can be further studied.

Acknowledgements

We wish to thank Natural Science Foundations of Jiangsu under Grant No. BK2005122 for support, and also thank the suppliers who provided the UCI machine learning repository.

References

[1] R.O. Duda, P.E. Hart, D.G. Stock, Pattern Classification, second ed., John Wiley and Sons, New York, 2001.

[2] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321–377.

[3] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning method, Neural Computation 16 (2004) 2639–2664.

[4] Marco Loog , Bram van Ginneken, R.P.W. Duin, Dimensionality reduction by canonical contextual correlation projections, in: Proceedings of the eighth European Conference on Computer Vision, May 2004, pp. 562–573.

[5] Y. Hel-Or, The canonical correlations of color images and their use for demosaicing, HP Labs Technical Report, HPL-2003-164 (R.1), 2004.

[6] T. Melzer, M. Reiter, H. Bischof, Appearance models based on kernel canonical correlation analysis, Pattern Recognition 36 (2003) 1961–1971.

[7] T.V. Gestel, J.A.K. Suykens, J. De Brabanter, B. De Moor, J. Vandewalle, Kernel canonical correlation analysis and least squares support vector machines, in: Proceedings of the International Conference on Artificial Neural Networks (ICANN 2001) 2001, pp. 384–389.

[8] Yo Horikawa, Use of Autocorrelation Kernels, in: Kernel Canonical Correlation Analysis for Texture Classification, ICONIP 2004, LNCS 3316, Springer-Verlag, Berlin, 2004, pp. 1235–1240.

[9] J. Baek, M. Kim, Face recognition using partial least squares components, Pattern Recognition 37 (2004) 303–1306.

[10] M. Barker, W. Rayens, Partial least squares for discrimination, Journal of Chemometrics 17 (2003) 166–173.

[11] B. Johansson, On classification: simultaneously reducing dimensionality and finding automatic representation using canonical correlation, Technical report LiTH-ISY-R-2375, ISSN 1400-3902, Linköping University, 2001.

[12] M. Borga, Canonical correlation: A tutorial. Available online from: <http://people.imt.liu.se/~magnus/cca/tutorial/>, 1999.

[13] C.M. Bishop, Neural Network for Pattern Recognition, Clarendon Press, Oxford, 1995.

[14] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Transactions of Information Theory 13 (1967) 21–27.

[15] J.M. Keller, M.R. Gray, J.A. Givens, A fuzzy *k*-nearest neighbor algorithm, IEEE Transactions on Systems Man and Cybernetics 15 (1985) 580–585.

[16] S.C. Chen, J. Liu, Z.H. Zhou, Making FLDA applicable to face recognition with one sample per person, Pattern Recognition 37 (2004) 1553–1555.

[17] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Graph embedding: A general framework for dimensionality reduction, in: Proceeding of IEEE CVPR'05, 2005.

[18] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and other Kernel-based Learning Methods, Cambridge University Press, 2000.