Applied Mathematics and Computation 217 (2011) 9041–9052

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Large correlation analysis

Xiaohong Chen [a], Songcan Chen [b,*], Hui Xue [c]

[a] Department of Mathematics, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, PR China
[b] College of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, PR China
[c] School of Computer Science and Engineering, Southeast University, Nanjing 210096, PR China

ARTICLE INFO

Keywords:
Canonical correlation analysis
Large margin learning
Dimensionality reduction
Total correlation
Individual correlation

ABSTRACT

In this paper, a novel supervised dimensionality reduction method is developed based on both the correlation analysis and the idea of large margin learning. The method aims to maximize the minimal correlation between each dimensionality-reduced instance and its class label, thus named as large correlation analysis (LCA). Unlike most existing correlation analysis methods such as CCA, CCAs and CDA, which all maximize the total or ensemble correlation over all training instances, LCA devotes to maximizing the individual correlations between given instances and its associated labels and is established by solving a relaxed quadratic programming with box-constraints. Experimental results on real-world datasets from both UCI and USPS show its effectiveness compared to the existing canonical correlation analysis methods.

1. Introduction

Canonical correlation analysis (CCA) [1], proposed by Hotelling in 1936, is a typical multivariate statistical analysis method. It aims to find a set of canonical basis vectors or projections for given two datasets by maximizing their total or ensemble correlation in a (common) projected space. Nowadays, CCA and its variants have been widely used in many areas such as pattern recognition [2–4], image analysis [5–9], computer vision [10,11], data regression analysis [12], image segmentation [13], climate forecasting[14], and multimedia analysis [15–17] etc. In terms of its mathematical formulation, CCA can be applied in any paired data. Generally, those paired data are usually obtained from different sources or information channels for the same (individual) object such as texts and their associated images contained in the same documents or web pages, sounds (speech) and images (lip feature) from the same person [11], etc. In fact, due to its generality, CCA has been used in dimensionality reduction (DR), feature extraction [2,4–7,13,14,17] and regression [12]. In this paper, we focus on DR and subsequent classification in its reduced space or its extracted features. It is well known that DR aims at seeking a transformation or projection from original potentially high-dimensional data to an alternative lower-dimensional representation to reveal the underlying structure of the data, and thus facilitates the subsequent tasks [5].

Now for a given set of paired data $\{(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_n,\mathbf{y}_n)\}$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ are instances obtained respectively from the two views (or sources) of some objects. For reducing the dimensionality of such a set of data, a natural and straightforward option is using CCA. In particular, when one view of the data, for example, y_i, is artificially-given class label associated with x_i, performing CCA reduces to single-view supervised DR [3–5,8,13]. With the extracted features, various successive tasks such as visualization and classification, etc. can be conducted. In this paper, we focus on supervised (discriminant) correlation analysis. Up to date, there have been many correlation analysis methods proposed for single-view dataset. According to whether

* Corresponding author.
E-mail addresses: lyandcxh@nuaa.edu.cn (X. Chen), s.chen@nuaa.edu.cn (S. Chen), hxue@seu.edu.cn (H. Xue).

0096-3003/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.amc.2011.03.117

class label is encoded into objective functions or not, these dimensionality reduction methods can roughly be divided into two groups: the first group adopts an explicit embedding of the class information [2] of the training instances or/and their neighbor information [13,18] and the second one an implicit embedding [22,23].

For the first group of correlation analysis methods, it is usually the case that the instances in each class share a common class label, it has been proved that using the (hard) one-of-C or one-of C-1 class encodings makes CCA reduce to either LDA or the variants [19–21] of LDA [40], which naturally inherits some limitations from LDA and at the same time weakens strength of CCA itself. In order to exert its strength able to handle any paired data, Sun [18] designed the instance-dependent soft labels according to its neighborhood information to develop a supervised CCA, termed as CCAs. The experiment shows that the CCAs is better than or comparable to CCA in terms of recognition performance. Liu [2] adopted a fuzzy class label for each instance in the form of its fuzzy membership degree to the distribution of the training instances. Loog et al. [13] gave another class label encoding approach specially for individual image pixels through concatenating the standard basis vectors in its neighborhood to segment given image. These correlation analysis methods promote the classification performance to different extents through different *explicit* incorporation of prior (class or neighborhood) information into the process of feature extraction.

Unlike the above *explicit* incorporations, CDA [22] and DCCA [23] *implicitly* incorporate the class information into the framework of correlation analysis to perform supervised DR. CDA [22] seeks a global linear transformation for DR by simultaneously maximizing the difference between the averaged within-class individual correlation and the averaged between–class individual correlation, however, it just suits for single-view data with class information. In DCCA [23], the authors firstly defined the within-class and between-class total correlations for the training instances, proved that the former is exactly equal to the negative of the latter and then obtained a DR projection by maximizing the within-class total correlation. It is worth noting that DCCA mostly analyzes two-view data in which each view corresponds to non-class label information. But when either view is one-of-C class encoding, DCCA is also equivalent to LDA.

CCA, CCAs and DCCA all focus on maximizing so-defined total or ensemble correlations of the training instances. However, CDA pays more attention to maximizing the difference between the averaged within-class individual correlations and the averaged between-class individual correlations. All those methods concern the whole rather than single correlation between the training instances, thus do not necessarily maximize the individual correlation between pairwise data. But the maximization of individual correlation is important in many practical applications such as classification. Recently, MMR [24], as a metamorphosis of CCA, was proposed to maximize the minimal correlation between individual pairs of instances in the feature space instead of the (total) sum of all instance correlations. Specifically, MMR tries to convert original CCA formalization into a convex margin-related optimization problem and find the solution to the problem to get a DR projection matrix, which makes MMR quite similar to support vector machines (SVM) [25,26]. However, such a conversion involves several quite unnatural operations in a strict mathematical sense that will be displayed in Section 1.

Motivated by the effectiveness of correlation analysis and the idea of large margin learning methods [24–26], we develop a new dimensionality reduction method termed as Large Correlation Analysis (LCA). LCA effectively combines both the correlation analysis and the large margin learning together. Firstly, we encode the class label for each training instance according to the class information and generate a two-view paired dataset to cater for the formulation of correlation analysis. Secondly, we try to seek a projection matrix by maximizing the minimal pairwise correlation between each projected instance and its corresponding (hard or soft) class label. Finally, we further using the properties of the matrix norm to relax the pairwise correlation to obtain a large correlation analysis method, *i.e.* LCA. LCA can be estabilished by solving a relaxed convex quadratic problem. It is necessary to point out that unlike the relaxation of MMR, our relaxation is natural and mathematically strict as detailed in Section 2. For the problem of optimizing convex quadratic programming with multivariate output, we first transform the convex quadratic programming into a corresponding box-constrained quadratic programming (BQP) via duality and then solve it by Project Barzilai–Borwein (PBB) method [27–29] effectively.

The rest of the paper is organized as follows. Section 2 gives a brief review of the related works. Section 3 derives our LCA for dimensionality reduction by maximizing the minimal pairwise correlation between each projected instance and its labels. Section 4 presents the experimental results on real-world datasets including UCI and USPS databases, followed by the conclusions and discussions in Section 5.

2. Related works

2.1. Canonical correlation analysis (CCA)

Canonical correlation analysis (CCA) utilizes a given paired dataset respectively from two spaces to simultaneously find a projection matrix for each feature space with aim at maximizing the correlation between the two projected representations [1,5].

Consider two multidimensional variables $\mathbf{x}$ and $\mathbf{y}$ both with zero mean, CCA tries to seek directions $\mathbf{w}_x$ and $\mathbf{w}_y$ to maximize the correlation between their projections, which can be expressed as

$$\max_{\mathbf{w}_x,\mathbf{w}_y} \frac{E_{x,y}[\mathbf{w}_x^T\mathbf{x}\mathbf{y}^T\mathbf{w}_y]}{\sqrt{E_x[\mathbf{w}_x^T\mathbf{x}\mathbf{x}^T\mathbf{w}_x]E_y[\mathbf{w}_y^T\mathbf{y}\mathbf{y}^T\mathbf{w}_y]}}, \tag{1}$$

where $E[\cdot]$ denotes the expectation of the variables. Usually, we are given a set of pairwise data $\{(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_n,\mathbf{y}_n)\}$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ correspond to the i-th object respectively. Denote $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_n]$, $\mathbf{Y} = [\mathbf{y}_1,\ldots,\mathbf{y}_n]$. Now empirically problem (1) can be approximated as

$$\max_{\mathbf{w}_x,\mathbf{w}_y} \frac{\mathbf{w}_x^T\mathbf{X}\mathbf{Y}^T\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{X}\mathbf{X}^T\mathbf{w}_x\mathbf{w}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{w}_y}}. \tag{2}$$

Obviously, what CCA maximizes is the *total* or *ensemble* correlation on the training instances. However, it is intuitive that such a correlation may not guarantee the maximization of individual correlations. Conversely, maximization of individual correlations can more likely lead to the total ensemble correlation. It is likely such an intuition that motivates Szedmak et al. [24] to develop their MMR as a variation of CCA.

2.2. Maximum margin robot (MMR)

In their implementation, Szedmak et al. [24] made the following relaxations:
Firstly,

$$\frac{\mathbf{w}_x^T\mathbf{X}\mathbf{Y}^T\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{X}\mathbf{X}^T\mathbf{w}_x\mathbf{w}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{w}_y}} = \frac{\left\langle \mathbf{Y}\mathbf{w}_y^T, \mathbf{X}\mathbf{w}_x^T\right\rangle_F}{\left\|\mathbf{w}_x^T\mathbf{X}\right\|\left\|\mathbf{w}_y^T\mathbf{Y}\right\|} \Rightarrow \frac{\left\langle \mathbf{Y}, \mathbf{X}\mathbf{w}_x^T\mathbf{w}_y\right\rangle_F}{\|\mathbf{Y}\|_F\left\|\mathbf{X}\mathbf{w}_x^T\mathbf{w}_y\right\|_F}. \tag{3}$$

Next, problem (3) is converted into a linearly constrained convex optimization problem with the form of maximizing the margin in SVM. Concretely, (3) is reformulated as

$$\begin{aligned} &\min \left\|\mathbf{X}\mathbf{w}_x^T\mathbf{w}_y\right\|_F \\ &\text{s.t.}\left\langle \mathbf{Y}, \mathbf{X}\mathbf{w}_x^T\mathbf{w}_y\right\rangle_F \geqslant \lambda. \end{aligned} \tag{4}$$

Furthermore, (1) a data independent objective $\|\mathbf{w}_x^T\mathbf{w}_y\|_F$ is used to replace the original data dependent objective function $\left\|\mathbf{X}\mathbf{w}_x^T\mathbf{w}_y\right\|_F$ in (4); (2) the total constraint in (4) is decomposed into a set of individual constraints corresponding to each instance, consequently, the MMR optimization problem is finally formed below:

$$\begin{aligned} &\min \quad \tfrac{1}{2}\|\mathbf{w}_x^T\mathbf{w}_y\|_F^2 \\ &\text{s.t.}\left\langle \mathbf{y}_i, \mathbf{w}_y\mathbf{w}_x^T\mathbf{x}_i\right\rangle_F \geqslant 1, \quad i = 1,2,\ldots,n, \end{aligned} \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\langle \mathbf{A}, B\rangle_F = tr(\mathbf{A}^T\mathbf{B})$ is the Frobenius inner product [29]. Such a relaxation from (3) to (5) involves several quite unnatural steps in a strict mathematical sense: 1) the denominator conversion from the left side to the right side of " $\Rightarrow$ " is mathematically neither natural nor strict; 2) the conversion of the objective functions from (4) to (5) is mathematically not so strict. More crucially, the solution resulted from (5) is the product of $\mathbf{w}_x$ and $\mathbf{w}_y$, which is not convenient for subsequent learning tasks such as visualizations. Though formally different from MMR, Ma et al. [22] also use individual correlations to develop CDA [22].

2.3. Correlation discriminant analysis (CDA)

Let $\mathbf{x}_i \in R^D (i = 1,2,\ldots,n)$ be D-dimensional instances and $y_i \in \{1,2,\ldots,C\}$ be associated hard class labels, where C is the number of classes. Unlike MMR, CDA [22] seeks an optimal DR transformation matrix $\mathbf{W}$ to maximize the difference between the averaged within-class individual correlations and the averaged between-class individual correlations.

$$\max_W \frac{1}{N_w}\sum_{(i,j)|_{y_i=y_j}}^{n} \frac{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_j}{\sqrt{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i\mathbf{x}_j^T\mathbf{A}\mathbf{x}_j}} - \frac{1}{N_b}\sum_{(i,j)|_{y_i\neq y_j}}^{n} \frac{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_j}{\sqrt{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i\mathbf{x}_j^T\mathbf{A}\mathbf{x}_j}} \tag{6}$$

or

$$\max_W \frac{1}{N_w}\sum_{(i,j)|_{y_i=y_j}}^{n} \frac{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_j}{\sqrt{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i\mathbf{x}_j^T\mathbf{A}\mathbf{x}_j}} - \frac{1}{n^2}\sum_{(i,j)}^{n} \frac{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_j}{\sqrt{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i\mathbf{x}_j^T\mathbf{A}\mathbf{x}_j}}. \tag{7}$$

Here, N_w and N_b are the numbers of instance pairs from the same classes or different classes respectively, and $\mathbf{A} = \mathbf{W}^T\mathbf{W}$. Similar to CCA, the maximization of the *averaged* individual correlations does not ensure the maximization of *single* individual correlation on each pair instances. On the other hand, the CDA objective function in (6) or (7) is not convex and has multiple local maxima. Thus it has to be solved by iterative optimization techniques such as the gradient ascent approach. In addition, CDA has higher computational cost due to involvement of many parameters.

3. Large correlation analysis

3.1. Motivation

The goal of CCA is to find a pair of optimal projections ($\mathbf{w}_x$,$\mathbf{w}_y$) to solve problem (1), or equivalently, minimize $E_{x,y}\left[\left\|\mathbf{w}_x^T\mathbf{x}-\mathbf{w}_y^T\mathbf{y}\right\|^2\right]$. It can be proved that CCA can attain the optimal correlation for Gaussian joint distribution of $\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}$ [see Appendix A]. However, it is difficult for CCA to obtain an ideal result for non-Gaussian data. One of the reasons behind this may be due to that the maximization of the *total* correlation between all given paired training data in the projected space. Thus CCA can not ensure the maximization of *individual* correlations in the projected space and that the difference among individual are more important for the subsequent tasks such as classification. In contrast, CDA [22] pays more attention to the *averaged individual* correlations and improves classification performance in its experiments. MMR [24] turns the *individual* correlations into corresponding constraints and maximizes the correlations in the maximum margin sense. Its experiments show improvement of subsequent classification performance. Inspired by the success of CDA and MMR, we propose a novel DR method based on maximum margin learning and (individual) correlation analysis, termed as large correlation analysis (LCA).

3.2. Large correlation analysis

3.2.1. Principle and deduction of large correlation analysis

Suppose that we are given the paired training instances

$$(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_n,\mathbf{y}_n)\in\mathscr{X}\times\mathscr{Y}, \tag{8}$$

where domain $\mathscr{X}\in R^D$ is a nonempty set which is fixed with unknown probability distribution, where instance $\mathbf{x}_i$ is drawn, and $\mathbf{y}_i$ is the corresponding class label encoded as in Appendix B, and thus $\|\mathbf{y}_i\|=1$, $i=1,2,\ldots,n$. The instances of the same class share a common class label as usual. Here it is necessary to point out that LCA can use either hard- or soft-encoded class labels and has no worry about the reduction to LDA resulted from the hard labels due to use of individual correlations. Denote $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n]\in R^{D\times n}$ and label matrix $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n]\in R^{(C-1)\times n}$. Without loss of generality, set all $\|\mathbf{x}_i\|=1$, $i=1,2,\ldots,n$. Let $\mathbf{W}$ be the projection matrix which projects each $\mathbf{x}_i$ to $\mathbf{W}^T\mathbf{x}_i$. Our goal is to find the $\mathbf{W}$ by maximizing the minimum of all individual correlations $\rho_i=\frac{\mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)}{\|\mathbf{y}_i^T\|\|\mathbf{W}^T\mathbf{x}_i\|}$ between each projection $\mathbf{W}^T\mathbf{x}_i$ and its class label $\mathbf{y}_i$, $i=1,2,\ldots,n$. This problem can be formulated as

$$\max_W \min_{1\leqslant i\leqslant n}\frac{\mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)}{\|\mathbf{y}_i^T\|\|\mathbf{W}^T\mathbf{x}_i\|}. \tag{9}$$

Equivalently,

$$\begin{aligned}&\max_W \quad t\\ &\text{s.t.}\quad \frac{\mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)}{\|\mathbf{y}_i^T\|\|\mathbf{W}^T\mathbf{x}_i\|}\geqslant t\quad i=1,2,\ldots,n.\end{aligned} \tag{10}$$

This is a non-convex optimization problem in $\mathbf{W}$ and can be solved by the gradient ascent algorithm. But the gradient ascent algorithm is usually time consuming. According to the Cauchy–Schwarz inequality of matrix norm, *i.e.*, $\|\mathbf{W}^T\mathbf{x}\|_2=\|\mathbf{W}^T\mathbf{x}\|_F\leqslant\|\mathbf{W}\|_F\|\mathbf{x}\|_2$, we relax the constraints of (10) and obtain

$$\frac{\mathbf{y}^T(\mathbf{W}^T\mathbf{x})}{\|\mathbf{y}\|\|\mathbf{W}^T\mathbf{x}\|}\geqslant\frac{\mathbf{y}^T(\mathbf{W}^T\mathbf{x})}{\|\mathbf{W}\|_F\|\mathbf{x}\|}=\frac{\mathbf{y}^T(\mathbf{W}^T\mathbf{x})}{\|\mathbf{W}\|_F}. \tag{11}$$

As a result, the above problem (10) can be translated to

$$\begin{aligned}&\max_W \quad t\\ &\text{s.t.}\quad \mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)\geqslant t\|\mathbf{W}\|_F\quad i=1,2,\ldots,n.\end{aligned} \tag{12}$$

Letting $t\|\mathbf{W}\|_F=\lambda$, we obtaining the following equivalent constrained optimization problem:

$$\begin{aligned}&\min_W \tfrac{1}{\lambda}\|\mathbf{W}\|_F\\ &\text{s.t.}\quad \mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)\geqslant\lambda\quad i=1,2,\ldots,n.\end{aligned} \tag{13}$$

Now, problem (13) can naturally be expressed as a convex quadratic programming as the following

$$\min_W \frac{1}{2}\|\mathbf{W}\|_F^2 \tag{14}$$

$$\text{s.t.}\quad \mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)\geqslant 1\quad i=1,2,\ldots,n. \tag{15}$$

Compared to the relaxation of MMR, our relaxation from (10) to (14) and (15) is natural and mathematically strict. Form the formulation (14), we can find that it has the same objective function as SVM of the quadratic form with linear constraints, except for: (1) using vector output to replace scalar output in SVM, and (2) using the DR matrix **W** just as an intermediate step of designing a final classifier. Thus we call the resulting algorithm (14) and (15) as large correlation analysis (LCA).

3.2.2. Solution and properties of large correlation analysis

In order to solve problem (14) and (15), we apply the Lagrange technique to define the following Lagrange function

$$L(\mathbf{W},\alpha)=\frac{1}{2}\|\mathbf{W}\|_F^2-\sum_{i=1}^{n}\alpha_i\left[\mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)-1\right]. \tag{16}$$

Zeroing the derivative of the objective (16) with respect to the **W** and using the Karush–Kuhn–Tucher (KKT) conditions, we have

$$\frac{\partial L(\mathbf{W},\alpha)}{\partial \mathbf{W}}=\mathbf{W}-\sum_{i=1}^{n}\alpha_i\mathbf{x}_i\mathbf{y}_i^T=0, \tag{17}$$

$$\alpha_i\left[\mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)-1\right]=0,\quad i=1,2,\ldots,n, \tag{18}$$

$$\alpha_i\geqslant 0,\quad i=1,2,\ldots,n, \tag{19}$$

where $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_n)^T$ is a vector consisting of the Lagrange multipliers corresponding to the constraints of (15). By (17), we obtain $\mathbf{W}=\sum_{i=1}^{n}\alpha_i\mathbf{x}_i\mathbf{y}_i^T$. It is a tensor product sum of the feature vectors and corresponding label vectors.

Note that $\mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)\geqslant 1,\ i=1,2,\ldots,n$ are a set of rigorous conditions for all training instances and generally do not always hold due to the existence of noise or outliers which will lead to no feasible solution for optimization (14) and (15). To make the problem still solvable and its solution robust, we introduce slack variables $\xi_i\geqslant 0$, $i=1,2,\ldots,n$ to relax the corresponding constraints as done in the soft margin formulation of support vector machine (SVM) [25,26]. With such a relaxation, the original constraints are changed into

$$\mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)\geqslant 1-\xi_i,\quad i=1,2,\ldots,n. \tag{20}$$

Usually, we adopt $\sum_{i=1}^{n}\xi_i$ as a metric to measure the degree of instances violating the constraints. Consequently, the relaxed LCA is reformulated as

$$\min\quad \frac{1}{2}\|\mathbf{W}\|_F^2+\eta\sum_{i=1}^{n}\xi_i, \tag{21}$$

$$\text{s.t.}\quad \mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)\geqslant 1-\xi_i,\quad i=1,2,\ldots,n, \tag{22}$$

$$\xi_i\geqslant 0,\quad i=1,2,\ldots,n, \tag{23}$$

where $\eta>0$ is a penalty constant that controls the trade-off between training error and margin. Solving problem (21)–(23) aims to make the minimal correlation as large as possible and at the same time the number of the constraints violated as small as possible. So-designed algorithm is termed as soft LCA, or LCA for short.

Again with the Lagrangian technique of the constrained optimization problem (21)–(23), we define the Lagrange function as follows:

$$L(W,\xi,\alpha,\mu)=\frac{1}{2}\|\mathbf{W}\|_F^2+\eta\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i\left[\mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)-1+\xi_i\right]-\sum_{i=1}^{n}\mu_i\xi_i, \tag{24}$$

where $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_n)^T$ and $\mu=(\mu_1,\mu_2,\ldots,\mu_n)^T$ are Lagrangian multipliers. Now zeroing the derivatives with respect to these variables yields

$$\frac{\partial L(\mathbf{W},\alpha)}{\partial \mathbf{W}}=\mathbf{W}-\sum_{i=1}^{n}\alpha_i\mathbf{x}_i\mathbf{y}_i^T=0, \tag{25}$$

$$\frac{\partial L}{\partial \xi_i}=\eta-\alpha_i-\mu_i=0,\quad i=1,2,\ldots,n \tag{26}$$

And the corresponding KKT conditions are

$$\alpha_i\left[\mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i)-1+\xi_i\right]=0,\quad i=1,2,\ldots,n, \tag{27}$$

$$\mu_i\geqslant 0,\quad i=1,2,\ldots,n, \tag{28}$$

$$\alpha_i\geqslant 0,\quad i=1,2,\ldots,n. \tag{29}$$

By (25), the projection matrix **W** can be characterized again as the sum of the tensor products of the instance vectors and their corresponding class label vectors, that is

$$\mathbf{W} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i \mathbf{y}_i^T \tag{30}$$

Substituting the equality constraints of (26) into (24), we obtain its Wolfe dual objective below

$$DL(\mathbf{W}, \xi, \alpha, \mu) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \mathbf{y}_i^T \mathbf{y}_j \mathbf{x}_i^T \mathbf{x}_j. \tag{31}$$

For more detailed derivation, please see Appendix C. Finally, solving problem (21)–(23) is equivalent to solving the following optimization problem:

$$\begin{aligned} &\max \quad \sum_{i=1}^{n} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \mathbf{y}_i^T \mathbf{y}_j \mathbf{x}_i^T \mathbf{x}_j, \\ &\text{s.t.} \quad 0 \leqslant \alpha_i \leqslant \eta \quad i = 1, 2, \ldots, n. \end{aligned} \tag{32}$$

Let $\mathbf{A}_{ij} = \mathbf{y}_i^T \mathbf{y}_j \mathbf{x}_i^T \mathbf{x}_j$ and use them to form a matrix $\mathbf{A}$. It is easy to prove that $\mathbf{A}$ is positive semi-definitiness. As a result, (32) can be rewritten in matrix form as

$$\begin{aligned} &\min \quad \tfrac{1}{2} \boldsymbol{\alpha}^T \mathbf{A} \alpha - \mathbf{1}^T \boldsymbol{\alpha}, \\ &\text{s.t.} \quad \mathbf{0} \leqslant \boldsymbol{\alpha} \leqslant \eta \mathbf{1}, \end{aligned} \tag{33}$$

where $\mathbf{1} = [1, \ldots, 1]^T \in R^n$, $\mathbf{0} = [0, \ldots, 0]^T \in R^n$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^T$. Obviously, (33) is a standard box-constrained quadratic programming problem (BQP) [27,28] in $\boldsymbol{\alpha}$ and can be solved using any off-the-shelf software packages such as SMO [30] solving SVM objective. However, slightly different from SVM formulation, our formulation involves multivariate rather than univariate output. Thus instead, we adopt the Projected Barzilai–Borwein Method (PBB) [29,31,32]. PBB method [31,32] is a projected gradient descent method incorporated with Barzilai–Borwein (BB) method [33] and the Grippo–Lampariello–Lucidi (GLL) line search [34]. Dai [29] proved its global convergence. In its each iteration, the constraints can be added or deleted from the working set according to whether they are violated. For more details, please refer to [29].

Now for our optimization problem (33), denote Ω the feasible set, that is

$$\Omega = \{\boldsymbol{\alpha} \in R^n | \mathbf{0} \leqslant \boldsymbol{\alpha} \leqslant \eta \mathbf{1}\}. \tag{34}$$

Let P_Ω be the projection operator onto Ω

$$P_\Omega(\alpha_i) = \begin{cases} \alpha_i & 0 \leqslant \alpha_i \leqslant \eta, \\ 0 & \alpha_i < 0, \\ \eta & \alpha_i > \eta. \end{cases} \tag{35}$$

Assume that the current $\boldsymbol{\alpha}^{(k)}$ is feasible and $\mathbf{g}_k = \mathbf{A}\boldsymbol{\alpha}^{(k)} - \mathbf{1}$, PBB method updates it to $\boldsymbol{\alpha}^{(k+1)}$ in terms of

$$\boldsymbol{\alpha}^{(k+1)} = P_\Omega(\boldsymbol{\alpha}^{(k)} - \lambda_k \mathbf{g}_k), \tag{36}$$

where $\lambda_k > 0$ is the step length. Once the $\boldsymbol{\alpha}$ is obtained and substituted into (30), a final projection matrix $\mathbf{W}$ can be produced and expressed as

$$\mathbf{W} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i \mathbf{y}_i^T = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \begin{pmatrix} \alpha_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha_n \end{pmatrix} \begin{pmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \end{pmatrix} = \mathbf{X} \Lambda \mathbf{Y}^T, \tag{37}$$

where $\boldsymbol{\Lambda} = diag(\boldsymbol{\alpha})$ is a diagonal matrix with its diagonal entries corresponding components of $\boldsymbol{\alpha}$. Now for classifying a test instance x_{test}, we first use the obtained matrix $\mathbf{W}$ to get its projection $\mathbf{W}^T x_{test} = \mathbf{Y} \Lambda^T \mathbf{X}^T x_{test}$ and then use the k-nearest neighbor classifier (k-NN) to test LCA classification performance and make comparison with some closely-related methods such as CCA, CCAs and CDA.

For clearness, Algorithm LCA can be formally stated in Table 1 below:

We now enumerate several characteristics of our LCA as follows:

(1) Compared with CCA and CCAs, LCA pays attention to maximizing the mimimum *individual* correlation between instance and its class label rather than the *total or ensemble* correlation (in fact, the sum of all *individual* correlations) of the training instances. Although both CDA and MMR treat *individual* correlations, the former focuses on maximizing the difference between the *averaged within-class individual* correlations and the *averaged between-class individual* correlations. Maximizing the *total* or *averaged* individual correlations does not necessarily ensure the maximization of *single* individual correlation on each pair instances. However, the latter, MMR, attains its objective in an unnatural and mathematically unstrict way.

Table 1
Algorithm LCA.

Input: Training instances matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in R^{D\times n}$; class label matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in R^{(C-1)\times n}$; parameter ε, k is the index of iterations.
output: $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T, \mathbf{W} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i \mathbf{y}_i^T$.
Step 1: Compute matrix $\mathbf{A}$ with elements $\mathbf{A}_{ij} = \mathbf{y}_i^T \mathbf{y}_j \mathbf{x}_i^T \mathbf{x}_j, i, j = 1, 2, \ldots, n$;
Step 2: Using PBB method to solve problem (33) to obtain $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$
2.0 Let $\boldsymbol{\alpha}^{(1)} = \left(\alpha_1^{(1)}, \ldots, \alpha_n^{(1)}\right)^T \in R^n$ and $\lambda_1 > 0$. If $\boldsymbol{\alpha}^{(1)} \notin \Omega$, replace $\boldsymbol{\alpha}^{(1)}$ by $P_\Omega(\boldsymbol{\alpha}^{(1)})$. Set $k = 1$.
2.1 Compute gradient vector $\mathbf{g}_k = \mathbf{A}\boldsymbol{\alpha}^{(k)} - \mathbf{1}$. If $\|P_\Omega(\alpha^{(k)} - \mathbf{g}_k) - \alpha^{(k)}\|_2 < \varepsilon$, *stop*.
2.2 Compute $\boldsymbol{\alpha}^{(k+1)} = P_\Omega\left(\boldsymbol{\alpha}^{(k)} - \lambda_k \mathbf{g}_k\right)$
2.3 Compute $\mathbf{s}_k = \boldsymbol{\alpha}^{(k+1)} - \boldsymbol{\alpha}^{(k)}$, and let step-length $\lambda_k = \frac{\mathbf{s}_k^T \mathbf{s}_k}{\mathbf{s}_k^T (\mathbf{g}_{k+1}^T - \mathbf{g}_k)}$.
2.4 Set $k = k + 1$, and goto step 2.1.
Step 3: Output $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ and $\mathbf{W} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i \mathbf{y}_i^T$.

(2) The objective function of LCA can be converted to its dual form by its BQP formulation and the latter can in turn be effectively solved by PBB method that does not suffer from the singularity problems which causes serious instability problems for CCA and CCAs. Using PBB method [29] to solve the problem does not need matrix inversion involved in CCA and CCAs and has been proved to be effective and easily coded. In addition, CCAs is usually time-consuming due to choice of an additional neighbor number k in determining soft labels for each instance by cross-validation, whereas CDA needs $O(Idn^2)$ computations [22] due to its many parameters involved and random initialization, and thus has higher cost than LCA which takes $O(Idn)$, where I is the number of iteration, d and n are the dimension and number of training instances, respectively.
(3) Although MMR has a similar formulation to LCA, its transformation matrix $\mathbf{W}$ needs meet decomposability, *i.e.*, it should be an tensor product of $\mathbf{W}_x$ and $\mathbf{W}_y$ ($\mathbf{W}_x$ and $\mathbf{W}_y$ are the projection matrices corresponding to $\mathbf{X}$ and $\mathbf{Y}$ respectively). The derivation of MMR is somewhat inconsistent with its initial motivation due to several unnatural relaxations involved. It is difficult to apply the matrix $\mathbf{W}$ obtained by MMR in classification because we don't know the exact value of $\mathbf{W}_x$ and $\mathbf{W}_y$. In contrast, our LCA is designed specially for single-view datasets with class information. Thus its projection transformation $\mathbf{W}$ can be directly used for subsequent classification.
(4) As we have known, the maximum extractable number of canonical variables by CCA is equal to $\min\{rank(\mathbf{X}\mathbf{X}^T), rank(\mathbf{Y}\mathbf{Y}^T)\}$. When the same class instances share a common class label, CCA reduces to LDA, which leads to that the maximally- reducible dimensional number is C-1. Although the reduced dimensionality of LCA is also C-1, it depends on the encoding way of class labels. The final dimension number varies with the encoding way and LCA is more flexible than both LDA and CCA in this sense.
(5) Both CCA and CCAs use the centered data to define the correlation termed as Pearson's correlation [36], while the CDA and our LCA directly base on original un-centered data to define the individual correlation, so-called the cosine correlation [36]. And it has been shown [37,38] that the Pearson's correlation is *less* discriminant than the cosine correlation due to the fact that the centered data are less informative than the original data. Though both CDA and LCA share a common starting point in the definition of correlation, our LCA is desired to have better classification performance due to the maximization of the minimal individual cosine correlation which is empirically validated in the following experiments.

4. Experiments and analysis

In this section, to investigate LCA classification performance, we conduct a systematical comparison with the state-of-art related dimensionality reduction algorithms such as LDA, CCA, CDA, CCAs in UCI database and USPS database. For both LCA and CCA, we select the same C-simplex vertices encoding [35] as class labels in the following experiments.

4.1. UCI Datasets

4.1.1. Datasets description

The UCI Repository of machine learning datasets [39] includes various datasets consisting of both artificial and real data, and can be accessed from the web site (http://archive.ics.uci.edu/ml/). Among the data sets in UCI machine learning repository, we selected 19 data sets which contain thirteen multi-class data sets and six binary-class data sets cited in Table 2.

4.1.2. Experimental setting and evaluation

For each dataset, we randomly select half of each class as training and the rest as testing and repeat the experiments ten times, and then report their average results in Table 2. For CCAs, the neighbor number k involved is selected from 1 to $\min_i\{|C_i|\}$, where $|C_i|$ is the number of training instances of class C_i.The reduced dimension of CCA is $\min\{p, q\}$, where p and q denote the dimension of data X and Y, respectively. The reduced dimension of LDA is C-1. CDA just seeks a transformation matrix without reduction. The parameter η in (33) of LCA is searched by cross-validation from $\{2^{-10}, 2^{-9}, \ldots 2^{0}, \ldots, 2^{9}, 2^{10}\}$ for optimizing performance. The parameter sought corresponding to the best results in the validation is used in testing.

Table 2
Comparison of average (%) and variance (10^{-4}) among CCA, LDA, CDA, CCAs and LCA on the 19 UCI datasets.

Dataset (Class/Dim/Num)	CCA	LDA	CDA	CCAs (k)	LCA
Banlance (3/4/625)	87.60 ± 5.16	87.53 ± 4.83	92.24 ± 0.77	89.13 ± 3.12 (1)	***93.94*** ± 3.66
Bupa (2/6/345)	57.44 ± 7.79	60.52 ± 14.0	63.26 ± 17.0	60.41 ± 18.74 (76)	***69.65*** ± 1.94
Cmc (3/9/1473)	42.61 ± 4.22	42.46 ± 4.34	***46.61*** ± 3.95	44.95 ± 2.32 (24)	44.84 ± 1.47
Dermat (6/33/366)	96.65 ± 0.77	96.32 ± 0.48	85.77 ± 4.39	96. 48 ± 0.47 (2)	***97.31*** ± 0.49
Ecoli (6/6/332)	81.10 ± 5.45	80.37 ± 4.77	80.30 ± 3.39	81.46 ± 7.12 (2)	***81.74*** ± 3.50
Glass (6/29/214)	58.57 ± 29.0	54.76 ± 40.0	58.01 ± 8.96	60.28 ± 6.54 (4)	***65.33*** ± 4.27
Iris (3/4/150)	92.80 ± 4.82	93.33 ± 3.56	91.33 ± 9.18	93.60 ± 3.48 (11)	***97.07*** ± 1.50
Ionosphere (2/34/351)	82.40 ± 6.30	84.29 ± 5.09	83.43 ± 4.14	83.14 ± 6.62 (3)	***89.83*** ± 4.34
Lense (3/4/24)	79.09 ± 50.0	79.09 ± 56.0	57.27 ± 221.	78.18 ± 77.14 (2)	***85.45*** ± 44.2
Pid (2/8/768)	68.88 ± 4.12	68.88 ± 4.12	66.09 ± 3.81	***70.44*** ± 5.93 (156)	65.91 ± 4.69
Sonar (2/60/208)	75.92 ± 29.0	74.56 ± 23.0	80.39 ± 12.0	77.48 ± 35.78 (29)	***83.69*** ± 9.80
Soybean (4/35/47)	97.39 ± 5.04	97.39 ± 5.04	92.17 ± 46.0	97.83 ± 5.25(2)	***99.13*** ± 7.56
Teaching (3/5/151)	55.07 ± 38.0	55.60 ± 79.0	50.67 ± 23.0	53.86 ± 40.38 (20)	***58.40*** ± 20.0
Thyroid (3/5/215)	92.71 ± 5.20	93.27 ± 7.14	78.32 ± 13.0	93.46 ± 3.88 (9)	***96.73*** ± 1.79
Vehicle (4/18/846)	73.44 ± 4.82	***73.84*** ± 3.05	58.01 ± 2.71	73.01 ± 2.66 (9)	52.86 ± 5.37
Wine (3/13/178)	***98.07*** ± 0.87	97.84 ± 1.56	63.64 ± 16.0	97.38 ± 3.75 (5)	79.89 ± 22.0
Water (2/38/116)	70.88 ± 51.0	71.40 ± 32.0	84.91 ± 9.71	75.09 ± 14.91(4)	***86.67*** ± 14.0
Wdbc (2/30/569)	93.45 ± 2.38	93.45 ± 2.38	90.25 ± 6.12	***94.08*** ± 1.12(53)	92.01 ± 2.84
Waveform (3/21/5000)	81.23 ± 0.47	81.22 ± 7.96	74.64 ± 0.24	81.84 ± 0.39 (170)	***82.01*** ± 0.18
Average	78.17 ± 13.39	78.22 ± 15.7	73.54 ± 21.33	79.05 ± 12.61	***80.12*** ± ***12.1***

4.1.3. Experimental results and discussion

We give the comparison of classification performance of the aforementioned methods in Table 2. In the last 5 columns, the mean and standard deviation of the accuracy are provided in percent. The bold digits indicate that the corresponding best mean accuracy is obtained from features extracted by the corresponding linear dimensionality reduction methods. In CCAs, the optimal number of neighbor $k$is also listed in Table 2.

The recognition results on the 19 datasets are given in Table 2. The best performances are highlighted and several attractive observations can be obtained:

(1) LCA outperforms LDA/CCA on 15 of the 19 datasets, specifically achieving the maximum improvement of 15.27% on Water, of more than 5% on Banlance, Bupa, Glass, Ionosphere, Lense and Sonar and slight improvement on the other 7 of the 15 datasets. However, on some outside of the 15 datasets, such as Vehicle and Wine, like CDA, LCA degrades significantly in performance.
(2) Compared with CDA, LCA provides better results on 16 out of the total 19 datasets, especially on the 11 datasets including Bupa, Dermat, Glass, Iris, Ionosphere, Lense, Soybean, Teaching, Thyroid, Wine and Waveform, LCA improves more than 6% in classification accuracies and obtains comparable results on the other 2 datasets. Furthermore, we need to point out that LCA has less computational cost than CDA in training stage.
(3) Although overall better than or comparable to the standard CCA in recognition performance, CCAs outperforms LCA only on four datasets such as Vehicle, Wine, Pid and Wdbc, to some degrees from 1% to 21%. Compared with CCA and CCAs, LCA performs best on the rest datasets.
(4) From the obtained accuracy variances of the five compared methods on each dataset, we can see that our LCA achieves the best stability on most of the datasets.

4.2. USPS Database

4.2.1. Datasets description and experimental setting

The USPS database (http://www.cs.toronto.edu/~roweis/data.html) consists of grayscale handwritten digit images from 0 to 9, as shown in Fig. 1. Each digit contains 1100 images with 256 gray levels and their sizes are all 16×16. Now we select five pairwise digits with varying difficulty for odd vs. even digits as usual in [43] to conduct our experiment.

4.2.2. Experimental results and discussion

In the experiment, we concretely select 110, 220, 330, 440, 550 instances respectively from each class for training and the rest for test, perform 10 runs and show their averaged accuracies in Fig. 2 for classification task of each pair of selected digits.

From Fig. 2, we can obtain several observations as follows:

(1) When the number of the training instances per class is small, such as 110, LCA shows an obvious performance superiority to the other four compared methods.
(2) With the increase of the training instances per class, the test accuracies of all the 5 methods are respectively improved to different extents.

Fig. 1. An illustration of 10 subjects in the USPS database.

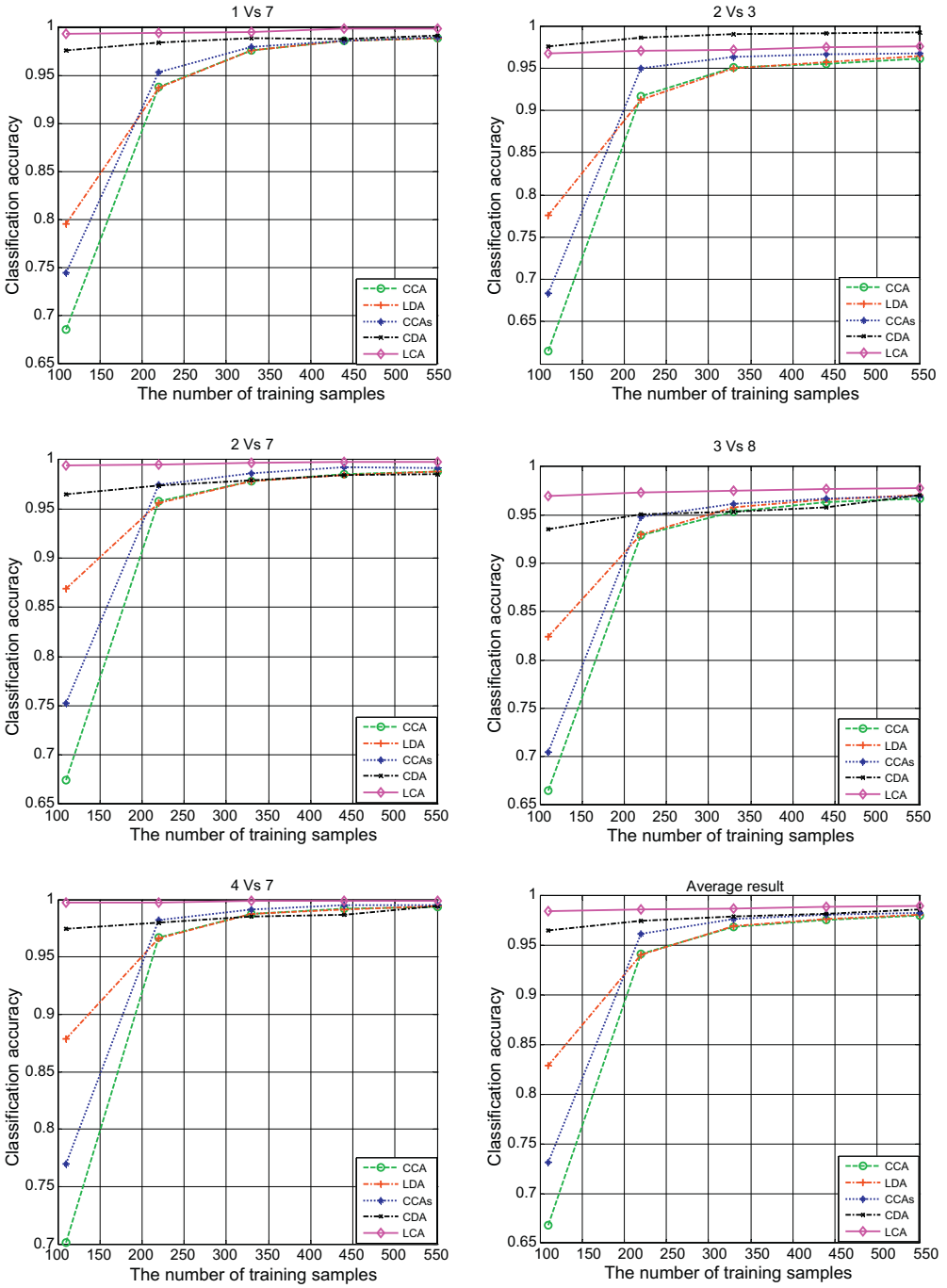

Fig. 2. Comparison among LDA, CCA, CCAs, CDA and LCA in single and average performances on the USPS database.

(3) When the number of the training instances per class is large enough to be able to reflect the data distribution, such as 440 and 550, most of the methods can achieve almost similar classification accuracies. However, for some relatively difficult recognition digits such as 1vs.7, 2vs.7 and 3vs.8, LCA still keeps its superiority among the compared methods.

Finally jointly from both the *average* accuracies respectively in Sections 4.1 and 4.2 , we can relatively safely claim that on the whole, our LCA maximizing the minimal *individual* correlation on training data yields better classification performance than the other correlation analysis methods which maximize the *total* or *ensemble* correlation on training data.

5. Conclusion and discussion

In this paper, inspired by the large margin learning methods, we propose a new linear dimensionality reduction method based on correlation analysis, namely large correlation analysis (LCA). Different from LDA, CCA, CCAs and CDA, LCA pays more attention to maximizing the minimal *individual* correlation between each projected training instance and the corresponding class label rather than the *total* or *ensemble* correlation. Then we extend it to the relaxed version and obtain a box-constrained (convex and multivariate) quadratic programming (BQP) problem in its dual space, as a result, with strong duality and projected Barzilai–Borwein (PBB) method, we resolve the dual problem efficiently and thus get the solution of the original problem. The experimental results on UCI machine learning repository and USPS database show encouraging classification performance compared with the state-of-art related works.

There are several directions of future study given as follow:

(1) **Kernelization**: Although all of the deduction and experimental results are based on the linear feature extraction, in fact, they can be easily generalized to nonlinear version via the kernel tricks [41] which is fit for linearly inseparable data.
(2) **Partially paired learning:** At present, CCA-type methods mainly involve paired instances (x,y), however, in practice, a large number of additional unpaired samples (*i.e.* x-only instances and y-only instances) may generate due to certain reasons such as camera occlusion in surveillance. To further utilize the prior information hidden in the additional unpaired instances, several semi-supervised [42] extensions of CCA have been proposed, *e.g.,* based on Tikhonov regularization [5] and Graph-Laplacian regularization [43]. We intend to further investigate partially paired data with the idea of LCA.
(3) **Sparse Learning**: Sparse learning such as the l1-regularization has attracted a lot of interests in recent years in statistics, machine learning and signal processing. Some works in [44,45] proposed sparse canonical correlation analysis (SCCA) which examines the relationship between two sets of variables and provides sparse solutions. Hence how to incorporate sparse learning with our LCA is another interesting topic for future work.

Acknowledgements

Partially and respectively Supported by NSFC Grant Nos. 60905002, 60973097 and 1101128, NUAA Research Funding (NS2010201).

Appendix A. Derivation of the CCA for Gaussian distribution data

Proposition. *If the random variable* $x \sim N(\mu_x, \Sigma_{11})$, $y \sim N(\mu_y, \Sigma_{22})$ *and* $z = \begin{pmatrix} x \\ y \end{pmatrix} \sim N(\mu, \Sigma)$, *where* $\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. *Based on the expectation risk minimization and the assumption that* $p(x,y)$ *is a joint unknown density function, we can perform dimensionality reduction for the two sets of variables according to the following rule:*

$$
\begin{aligned}
&\iint_{\mathbf{XY}} \left\| \mathbf{W}_x^T \mathbf{x} - \mathbf{W}_y^T \mathbf{y} \right\|^2 p(x,y)\,dxdy \\
&= \iint_{\mathbf{XY}} \left\| \mathbf{W}_x^T \mathbf{x} - \mathbf{W}_y^T \mathbf{y} \right\|^2 p(x|y)p(y)\,dxdy = \int_{\mathbf{Y}} p(y)dy \int_X \left[\mathbf{W}_x^T \mathbf{x}\mathbf{x}^T \mathbf{W}_x - 2\mathbf{W}_x^T \mathbf{x}\mathbf{y}^T \mathbf{W}_y + \mathbf{W}_y^T \mathbf{y}\mathbf{y}^T \mathbf{W}_y \right] p(x|y)\,dx \\
&= \mathbf{W}_x^T \left[\int_{\mathbf{Y}} \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} + \mu_x \mu_x^T \right) p(y)dy \right] \mathbf{W}_x - 2\mathbf{W}_x^T \left[\int_{\mathbf{Y}} \left(\mu_x + \Sigma_{12}\Sigma_{22}^{-1}(y - \mu_y) \right) \mathbf{y}^T p(y)dy \right] \mathbf{W}_y + \mathbf{W}_y^T \left[\int_{\mathbf{Y}} \mathbf{y}\mathbf{y}^T p(y)dy \right] \mathbf{W}_y \\
&= \mathbf{W}_x^T \left[\Sigma_{11} + \mu_x \mu_x^T \right] \mathbf{W}_x - 2\mathbf{W}_x^T \left[\Sigma_{12} + \mu_x \mu_y^T \right] \mathbf{W}_y + \mathbf{W}_y^T \left[\Sigma_{22} + \mu_y \mu_y^T \right] \mathbf{W}_y .
\end{aligned}
$$

In fact, we don't know exact the mean and the variance of the two variables, via replacing them with the sample mean and the variance, we obtain

$$\iint_{\mathbf{XY}} \left\|\mathbf{W}_x^T\mathbf{x}-\mathbf{W}_y^T\mathbf{y}\right\|^2 p(x,y)dxdy = \frac{1}{n}\left[\mathbf{W}_x^T\mathbf{X}\mathbf{X}^T\mathbf{W}_x - 2\mathbf{W}_x^T\mathbf{X}\mathbf{Y}^T\mathbf{W}_y + \mathbf{W}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{W}_y\right].$$

With the constraints $\iint_{\mathbf{XY}} \left\|\mathbf{W}_x^T\mathbf{x}\right\|^2 p(x,y)dxdy = 1$ and $\iint_{\mathbf{XY}} \left\|\mathbf{W}_y^T\mathbf{y}\right\|^2 p(x,y)dxdy = 1$, we obtain the subsequent optimization problem:

$$\begin{aligned}
&\min \quad \mathbf{W}_x^T\mathbf{X}\mathbf{X}^T\mathbf{W}_x - 2\mathbf{W}_x^T\mathbf{X}\mathbf{Y}^T\mathbf{W}_y + \mathbf{W}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{W}_y,\\
&\text{s.t.} \quad \mathbf{W}_x^T\mathbf{X}\mathbf{X}^T\mathbf{W}_x = 1,\\
&\qquad \mathbf{W}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{W}_y = 1.
\end{aligned}$$

For its optimization, firstly using the Lagrangian multiplier method to define the following function:

$$L(\mathbf{W}_x,\mathbf{W}_y) = \mathbf{W}_x^T\mathbf{X}\mathbf{X}^T\mathbf{W}_x - 2\mathbf{W}_x^T\mathbf{X}\mathbf{Y}^T\mathbf{W}_y + \mathbf{W}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{W}_y - \lambda_1\left(\mathbf{W}_x^T\mathbf{X}\mathbf{X}^T\mathbf{W}_x - 1\right) - \lambda_2\left(\mathbf{W}_y^T\mathbf{Y}\mathbf{Y}^T\mathbf{W}_y - 1\right).$$

Then zeroing the derivatives with respect to $\mathbf{W}_x$ and $\mathbf{W}_y$ yields

$$\begin{aligned}
&\frac{\partial L(\mathbf{W}_x,\mathbf{W}_y)}{\partial \mathbf{W}_x} = 2\mathbf{X}\mathbf{Y}^T\mathbf{W}_y - 2\mathbf{X}\mathbf{X}^T\mathbf{W}_x - 2\lambda_1\mathbf{X}\mathbf{X}^T\mathbf{W}_x = 0 \Rightarrow (1+\lambda_1)\mathbf{X}\mathbf{X}^T\mathbf{W}_x = \mathbf{X}\mathbf{Y}^T\mathbf{W}_y,\\
&\frac{\partial L(\mathbf{W}_x,\mathbf{W}_y)}{\partial \mathbf{W}_y} = 2\mathbf{Y}\mathbf{X}^T\mathbf{W}_x - 2\mathbf{Y}\mathbf{Y}^T\mathbf{W}_y - 2\lambda_2\mathbf{Y}\mathbf{Y}^T\mathbf{W}_y = 0 \Rightarrow (1+\lambda_2)\mathbf{Y}\mathbf{Y}^T\mathbf{W}_y = \mathbf{Y}\mathbf{X}^T\mathbf{W}_x.
\end{aligned}$$

Incorporating above two constraints leads to $\lambda_1 = \lambda_2$. Using λ to denote both, we can find that under the Gaussian assumption, dimensionality reduction for two sets of variables based on the above risk minimization principle is equivalent to CCA.

Appendix B. Class label encoding [35]

Motivated by the fact that the C vertices of a regular C-simplex are the most balanced and symmetricly separate points in the (C-1)-dimensional space, here we choose the regular C-simplex vertices as the multivariate label. Let $\mathbf{L}_i \in R^{C-1}$, $i = 1, 2, \ldots, C$ be the vertices of the regular C-simplex in R^{C-1} and $\mathbf{L} = [\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_C]$, where $\mathbf{L}_{i,j}$ is an element of $\mathbf{L}$ in the ith row and jth column. Firstly, let $\mathbf{L}_1 = (1, 0, \ldots, 0)^T \in R^{C-1}$ and $\mathbf{L}_{1,i} = \frac{-1}{C-1}$ for $i = 2, \ldots, C$. Then we get the first row and the first column of the $\mathbf{L}$. Secondly, we compute the next rows and columns by the following formula

$$\mathbf{L}_{k+1,k+1} = \sqrt{1 - \sum_{i=1}^{k} \mathbf{L}_{i,k}^2}, \tag{B1}$$

$$\mathbf{L}_{k+1,j} = -\frac{\mathbf{L}_{k+1,k+1}}{C-k-1}, \quad j = k+2, \ldots, C, \tag{B2}$$

$$\mathbf{L}_{i,k+1} = 0, \quad k+1 < i \leqslant C-1, \tag{B3}$$

where $k = 1, 2, \ldots, C-2$. It is easy to prove that $\sum_{i=1}^{C} \mathbf{L}_i = 0, \|\mathbf{L}_i\| = 1, i = 1, 2, \ldots, C$ and

$$\|\mathbf{L}_i - \mathbf{L}_j\| = \sqrt{\frac{2C}{C-1}} \tag{B4}$$

i.e., these class labels have zero mean, unit norms and equal pairwise distances. For example, the correspondence class labels for three class instances are $\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}-1/2\\\sqrt{3}/2\end{bmatrix}$ and $\begin{bmatrix}-1/2\\-\sqrt{3}/2\end{bmatrix}$ respectively. If $\mathbf{x}_i \in C_k$,$i = 1, 2, \ldots, n$, $\mathbf{y}_i = \mathbf{L}_k$, $k = 1, 2, \ldots, C$. The instances in each class share a common class label as usual. So we obtain a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in R^{D\times n}$ and an associated class label matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n] \in R^{(C-1)\times n}$ which can be used for correlation analysis in this paper.

Appendix C. The derivation of the Wolfe dual representation (31) and (32) for the Lagrangian function

$$\begin{aligned}
L(\mathbf{W},\xi,\alpha,\mu) &= \frac{1}{2}\|\mathbf{W}\|_F^2 + \eta\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[\mathbf{y}_i^T(\mathbf{W}^T\mathbf{x}_i) - 1 + \xi_i\right] - \sum_{i=1}^{n}\mu_i\xi_i\\
&= \frac{1}{2}\|\mathbf{W}\|_F^2 + \eta\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\mathbf{x}_i^T\mathbf{W}\mathbf{y}_i + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\mu_i\xi_i = \frac{1}{2}\|\mathbf{W}\|_F^2 + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\mathbf{x}_i^T\mathbf{W}\mathbf{y}_i\\
&= \sum_{i=1}^{n}\alpha_i + \frac{1}{2}tr[\mathbf{W}\mathbf{W}^T] - tr\left[\mathbf{W}\sum_{i=1}^{n}\alpha_i\mathbf{y}_i\mathbf{x}_i^T\right] = \sum_{i=1}^{n}\alpha_i + \frac{1}{2}tr[\mathbf{W}\mathbf{W}^T] - tr[\mathbf{W}\mathbf{W}^T]\\
&= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\mathbf{y}_i^T\mathbf{y}_j\mathbf{x}_i^T\mathbf{x}_j.
\end{aligned}$$

References

[1] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321–377.
[2] Y.Y. Liu, X.P. Liu, Z.X. Su, A new fuzzy approach for handling class labels in canonical correlation analysis, Neurocomputing 71 (2008) 1735–1740.
[3] T.V. Gestel, J.A.K. Suykens, J. De Brabanter, et al. Kernel canonical correlation analysis and least squares support vector machines, in: Proceedings of the International Conference on Artificial Neural Networks, 2001, pp. 384–389.
[4] Q.S. Sun, S.G. Zeng, Y. Liu, et al, A new method of feature fusion and its application in image recognition, Pattern Recognition (38) (2005) 2437–2448.
[5] D.R. Hardoon, S. Szedmak, J.S. Taylor, Canonical correlation analysis: an overview with application to learning method, Neural Computation 16 (2004) 2639–2664.
[6] O. Friman, M. Borga, P. Lundberg, et al, Canonical correlation as a tool in functional MRI data analysis, Proceedings of SSAB Symposium on Image Analysis, Norrkoping, Sweden, 2001.
[7] A.A. Nielsen, Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data, IEEE Transactions in Image Processing 11 (2002) 293–305.
[8] Y. Horikawa, Use of autocorrelation kernels in kernel canonical correlation analysis for text classification, in: International Conference on Neural Information Processing, 2004, pp. 1235–1240.
[9] Y. Fu, T.S. Huang, Image classification using correlation tensor analysis, IEEE Transaction on Image Processing. 17 (2) (2008) 226–234.
[10] T. Melzer, M. Reiter, H. Bischof, Appearance models based on kernel canonical correlation analysis, Pattern Recognition 36 (2003) 1961–1971.
[11] E. Kidron, Y.Y. Schechner, M. Elad, Pixels and sound, IEEE Proceedings of Computer Vision and Pattern Recognition 1 (2005) 88–95.
[12] B. Abraham, G. Merola, Dimensionality reduction approach to multivariate prediction, Computational Statistics and Data Analysis 48 (2005) 5–16.
[13] M. Loog, B.V. Ginneken, R.P.W. Duin, Dimensionality reduction by canonical contextual correlation projections, Pattern Recognition 38 (2005) 2409–2418.
[14] A.J. Cannon, W.W. Hsieh, Robust nonlinear canonical correlation analysis: application for seasonal climate forecasting, Nonlinear Processes in Geophysics 12 (2008) 221–232.
[15] W.W. Feng, B.U. Kim, Y. Yu, Real-time data-driven deformation using kernel canonical correlation analysis, ACM Transactions on Graphics 27 (3) (2008) 1–9.
[16] M.E. Sargin, Y. Yemez, E. Erzin, et al, Audiovisual synchronization and fusion using canonical correlation analysis, IEEE Transactions on Multimedia 9 (7) (2007) 1396–1403.
[17] T.K. Kim, R. Cipolla, Canonical correlation analysis of video volume tensors for action categorization and detection, IEEE Transaction Pattern Analysis and Machine Intelligence. 31 (8) (2009) 1415–1428.
[18] T.K. Sun, S.C. Chen, Class label versus samples label-based CCA, Applied Mathematics and Computation 185 (2007) 272–283.
[19] M. Barker, W. Rayens, Partial least squares for discrimination, Journal of Chemometrics 17 (2003) 166–173.
[20] Y.H. He, L. Zhao, C.R. Zou, Face recognition based on PCA/KPCA plus CCA, in: Proceedings of the ICNC 2005, Lecture Notes in Computer Science, (3611), Springer, Berlin, 2005, pp. 71–74.
[21] L. Sun, S.W. Ji, J.P. Ye, A least squares linear discriminant analysis, International Conference on Machine Learning (2006) 1087–1094.
[22] Y. Ma, S.H. Lao, E. Takikawa, et al, Discriminant analysis in correlation similarity measure space, Proceeding of the 24 th International Conference on Machine Learning 227 (2007) 577–584.
[23] T.K. Sun, S.C. Chen, J.Y. Yang, et al, A novel method of combined feature extraction for recognition, IEEE Conferences on Data Minging (2008) 1043–1048.
[24] S. Szedmak, T.D. Bie, D. Hardoon, A metamorphosis of canonical correlation analysis into multivariate maximum margin learning, European Symposium Artificial Neural Networks (2007).
[25] J. Share-Taylor, N. Cristianini, Support Vector Machines, Cambridge University Press, 2000.
[26] C.J. C Burgest, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (1998) 121–167.
[27] J.J. Moré, G. Toraldo, Algorithms for bound constrained quadratic programming problems, Numerische Mathematik 55 (1989) 377–400.
[28] J.J. Moré, G. Toraldo, On the solution of large quadratic programming problems with bound constraints, SIAM Journal on Optimization 1 (1991) 93–113.
[29] Y. H Dai, R. Fletcher, Projected Barzilai–Borwein methods for large-scale box-constrained quadratic programming, Numerische Mathematik 100 (2005) 21–47.
[30] J.C. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Scholkopf, C. Burges, A. Smola (Eds.), Advances in KernelMethods: Support Vector Machines, Cambridge MA; MIT Press, Cambridge, 1998, pp. 185–208.
[31] E.G. Birgin, J.M. Martinez, M. Raydan, Nonmonotone spectral projected gradient methods on convex sets, SIAM Journal on Optimization. 10 (2000) 1196–1211.
[32] E.G. Birgin, J.M. Martinez, M. Raydan, Algorithm 813: SPG-software for convex-constrained optimization, ACM Transaction on Mathematical Software 27 (2001) 340–349.
[33] J. Barzilai, J.M. Borwein, Two-point step gradient methods, IMA Journal of Numerical Analysis 8 (1988) 141–148.
[34] L. Grippo, F. Lampariello, S. Lucidi, A nonmonotone line search techniques for Newton's method, SIMA Journal on Numerical Analysis 23 (1986) 07–716.
[35] S.J. An, W.Q. Liu, S. Venkatesh, Face recognition using kernel ridge regression, Computer Vision and Pattern Recognition (2007) 1–7.
[36] J. Rodgers, W. Nicewander, Thirteen ways to look at the correlation coefficient, The American Statistician 42 (1) (1988) 59–66.
[37] J.W. Schneider, P. Bouiund, Matrix comparison, Part 1. Motivation and important issues for measuring the resemblance between proximity measures or ordination results, Journal of the American Society for Information Science and Technology 58 (11) (2007) 1586–1595.
[38] P. Ahlgren, B. Jarneving, R. Rousseau, Requirement for a cocitation similarity measure, with special reference to peasron's correlation coefficient, Journal of the American Society for Information Science and Technology 54 (6) (2003) 550–560.
[39] C.L. Blake, C.J. Merz, UCI repository of machine learning databases, University of California, Department of Information and Computer Sciences, Irvine, 1998.
[40] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics. 7 (1936).
[41] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.
[42] O. Chapelle, B. Scholkopf, A. Zien (Eds.), Semi-Supervised Learning, MIT Press, 2006.
[43] M.B. Blaschko, C.H. Lampert, A. Gretton, Semi-supervised laplacian regularization of kernel canonical analysis, In ECML PKDD'08 (2008) 133–145.
[44] D.R. Hardoon, J. Shawe-Taylor, Sparse canonical correlation analysis, Technical report, University College London, 2007.
[45] E. Parkhomenko, D. Tritchler, J. Beyent, Sparse canonical correlation analysis with application to genomic data integration, Statistical Applications in Genetics and Molecular Biology 8 (1) (2009) (article 1).