Neural Process Lett (2013) 37:335–354
DOI 10.1007/s11063-012-9251-z

Neighborhood Correlation Analysis for Semi-paired Two-View Data

Xudong Zhou · Xiaohong Chen · Songcan Chen

Published online: 23 October 2012

Abstract Canonical correlation analysis (CCA) is a widely used technique for analyzing two datasets (two views of the same objects). However, CCA needs that the samples of the two views are fully-paired. Actually, we are often faced up with the *semi-paired* scenario where the number of available paired samples is limited and yet the number of unpaired samples is sufficient. For such a scenario, CCA is generally prone to overfitting and thus performs poorly, since its definition itself makes it only able to utilize those paired samples. To overcome such a shortcoming, several *semi-paired* variants of CCA have been proposed. However, unpaired samples in these methods are just used in the way of single-view leaning to capture individual views' structure information for regularizing CCA. Intuitively, using unpaired samples in the way of two-view learning should be more natural and more attractive since CCA itself is a two-view learning method. As a result, a novel CCAs *semi-paired* variant named *Ne*ighborhood *C*orrelation *A*nalysis (*NeCA*), which uses unpaired samples in the two-view learning way, is developed through incorporating between-view neighborhood relationships into CCA. The relationships are acquired through leveraging within-view neighborhood relationships of each view's all data (including paired and unpaired data) and between-view paired information. Thus, it can take more sufficient advantage of the unpaired samples and then mitigate overfitting effectively caused by the limited paired data. Promising

X. Zhou · S. Chen (✉)
College of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China
e-mail: s.chen@nuaa.edu.cn

X. Zhou
e-mail: xdzhou@nuaa.edu.cn

X. Zhou
Information Engineering College, Yangzhou University, Yangzhou 225127, China

X. Chen
College of Science, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China

S. Chen
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China

experiments results on several popular multi-view datasets show its feasibility and effectiveness.

Keywords Canonical correlation analysis · Semi-paired learning · Two-view learning · Neighborhood relationship · Neighborhood correlation

1 Introduction

High-dimensional co-occurring data associated with an object frequently and abundantly emerge in the real world. For example, an Internet web page as an object can be represented as (co-occurring) page text and links to the page, and a human can be represented as co-occurring visual and audio contents. A lot of works have been done for analyzing this kind of data [1–7]. Among these works, canonical correlation analysis (CCA) is one of the most widely adopted methods [8–12].

CCA is a classical but useful multivariate statistical analysis method [13]. It aims to find maximally correlated projections between two sets of variables, which can be considered as two views (views **x** and **y**) or representations of the same set of objects. However CCA requires that such two views be *fully-paired*, i.e., each sample in view **x** should have a correspondence in view **y**, and vice versa. Conversely, we are often faced such a scenario where most samples in view **x** have no correspondences in view **y**, and vice versa, thus forming the semi-paired scenario called here. For such a scenario, CCA is generally prone to overfitting and thus performs poorly, since its definition itself makes it only suitful for the paired scenario, so its applications are limited in the real world. Actually, abundant unpaired samples (i.e. **x**- and **y**-only samples) often contain much useful information which will benefit the learning task, just as the unlabeled samples benefit semi-supervised leaning [14,15] by exploiting the intrinsic data structure under clustering assumption or manifold assumption. Recently, several works have concerned such new scenario [16–18]. Blaschko et al. [16] proposed a semi-supervised Laplacian regularization of kernel CCA (SemiLRKCCA), which utilizes intrinsic geometry structure of each view to regularize kernel CCA (KCCA) [19]. As a result, SemiLRKCCA can find a set of meaningful directions which not only make the two view's paired samples highly correlated but also capture each view's manifold structure. SemiCCA [17] utilizes global structure of each view's whole training samples (paired and unpaired samples together) to regularize CCA in order to bridge CCA and principal component analysis (PCA) [20,21] seamlessly. Both SemiLRKCCA and SemiCCA can take sufficient advantage of unpaired samples in addition to paired samples, and consequently achieve better results than CCA just based on the paired samples. It is necessary to mention that the actual meaning of "semi-" in SemiLRKCCA and SemiCCA is "semi-paired" rather than "semi-supervised" in popular semi-supervised learning literature [14,15]. Compared with SemiLRKCCA and SemiCCA, more recent work termed as semi-paired and semi-supervised generalized correlation analysis (S^2GCA) [18] make further research for dealing with semi-paired and semi-supervised scenario. S^2GCA utilizes within-view structural information and within-view discriminant information jointly, to preserve the individual view's structure of unlabeled data and separate labeled data in different classes from each other simultaneously. Without semi-supervised information, S^2GCA is similar to SemiLRKCCA and SemiCCA.

In SemiLRKCCA, SemiCCA and S^2GCA, unpaired samples are just used in the way of single-view leaning to capture individual views' structure information for regularizing KCCA or CCA. Consequently, CCA and its variants (SemiLRKCCA, SemiCCA and S^2GCA) only

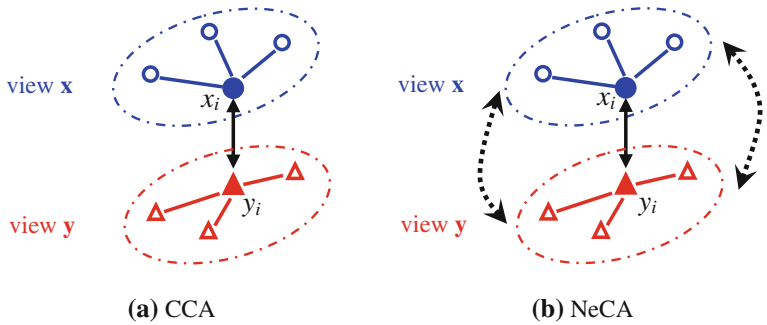

Fig. 1 Comparison of considered correlations between different views in CCA and NeCA. **a** CCA, **b** NeCA. '$\leftrightarrow$' denotes correlation between x_i and y_i, '$\prec \cdots \succ$' denotes correlations between x_i's neighborhood and y_i's neighborhood, where x_i and y_i are paired samples

consider the correlations between paired samples in different views. However, using unpaired samples in the way of two-view learning should be more natural and more attractive since CCA itself is a two-view learning method. Actually, based on the manifold assumption [22], it is reasonable to consider not only the correlations between paired samples in given two views but also the correlations between their respective neighborhoods, as shown in Fig. 1 (see Sect. 3). Therefore, a novel dimension reduction method named *Ne*ighborhood *C*orrelation *A*nalysis (*NeCA*), which uses unpaired samples in the two-view learning way, is developed through incorporating between-view neighborhood relationships into CCA. Note that the construction of between-view neighborhood relationships seems not so straightforward, since the two views (views **x** and **y**) are heterogeneous and semi-paired. Therefore, a method named spectral minimizing-disagreement (SMD) [4] is adopted to generate the relationships, mainly due to that SMD can incorporate the within-view neighborhood relationships and the between-view paired information without artificially equating or relating them naturally.

The recent work S^2GCA is a semi-supervised dimension reduction method. Similar to S^2GCA, local discrimination CCA (LDCCA) [23] also utilizes supervised information. On the other hand, like our work NeCA, LDCCA also incorporates the local manifold geometry structure of data in modeling. However, LDCCA is just appropriate for the fully-paired and fully-supervised scenario, which is different from our semi-paired and unsupervised scenario in NeCA. Thus, in our experiments, we do not make comparison with such two methods. Our work NeCA is developed for adapting CCA to *semi-paired* scenario. Under the scenario, the construction of between-view neighborhood relationships in NeCA is simple, only needing a matrix multiplication (see Sect. 3.2). On the other hand, the works by Aria et al. [24] and Tripathi et al. [25,26] utilize CCA as a dependency measure for solving the problem of cross-domain (or cross-view) object matching (CDOM) [27]. The goal of CDOM is to find correspondence between two sets of objects in different domains in an unsupervised and unpaired way by maximizing pairwise dependency. Usually, solving the pairing problems in [24–26] needs a complicated iterative algorithm, and high computational cost at each iteration. Though [25] uses partial paired information as penalization to improve the accuracy, there is no essential change in computational complexity. Similar to our work, the work [28] for manifold alignment also uses the within-view neighborhood relationships to define the between-view neighborhood relationships, However, it is computationally very complex [25], which has factorial level permutations for a relationship established between different views.

In what follows, we summarize favorable and attractive characteristics of the proposed algorithm:

(1) NeCA considers the correlations not only between those paired samples but also between each sample in one view and its corresponding between-view local neighbors in the other view. This way, it uses unpaired samples in the way of two-view learning. Thus, it can take more sufficient advantage of the unpaired samples and then mitigate overfitting effectively caused by the limited paired data.
(2) NeCA has no regularization in its optimization problem, due to that NeCA utilizes those unpaired samples in the way of two-view learning. As a result, NeCA can be boosted in performance by the regularizing individual views with structure information as employed in SemiLRKCCA and SemiCCA. Thus, two extensions of NeCA are developed: Laplacian-regularization NeCA (LRNeCA) and PCA-regularization of NeCA (PRNeCA).
(3) CCA, SemiLRKCCA and SemiCCA are special cases of NeCA, LRNeCA and PRNeCA respectively when they only consider the correlations between paired data.
(4) Despite all these advantages, NeCA still maintains the characteristic of being easily solved by a generalized eigenvalue problem (GEP) similar to regular CCA, and can directly be generalized to more than two views through the way of CCA to multi-set CCA [29].

The remainder of this paper is structured as follows: In Sect. 2, we briefly review CCA and its semi-paired variants. We propose NeCA and formulate specific learning algorithm in Sect. 3. NeCA can be regularized by individual views' structure information, therefore corresponding generalizations of NeCA are presented in Sect. 4. In Sect. 5, we present the experimental results on several popular multi-view datasets, including two web page datasets (Internet Advertisements and WebKB) and two image datasets (Multiple Feature Handwritten Digit Database [MFD] and Yale). Finally, we conclude this paper in Sect. 6.

2 CCA and Its Semi-paired Variants

2.1 Canonical Correlation Analysis (CCA)

CCA [8,13], proposed by Hotelling [13], is a well-known technique for finding pairs of vectors that maximize the correlation between two sets of paired variables. The two sets of variables can be considered as two views of the same objects.

To be specific, given $X = [x_1, \ldots, x_n] \in R^{d_x \times n}$ and $Y = [y_1, \ldots, y_n] \in R^{d_y \times n}$ are the two sets of variables of the same objects, where $x_i \in R^{d_x}$ and $y_i \in R^{d_y}$ (both with zero mean) correspond to the ith object. CCA can be defined as the problem of finding a pair of canonical basis vectors: w_x of size $d_x \times 1$ and w_y of size $d_y \times 1$, such that the projected variables $w_x^T X$ and $w_y^T Y$ are maximally correlated. The two projection vectors can be acquired by maximizing the following correlation coefficient:

$$\rho = \frac{w_x^T X Y^T w_y}{\sqrt{w_x^T X X^T w_x w_y^T Y Y^T w_y}}. \tag{1}$$

Since ρ is invariant to the scaling of w_x and w_y, CCA can be expressed equivalently as the following constrained optimization problem:

$$\begin{aligned} &\max_{w_x, w_y} \; w_x^T X Y^T w_y \\ &s.t. \;\; w_x^T X X^T w_x = 1, \quad w_y^T Y Y^T w_y = 1. \end{aligned} \tag{2}$$

The solution of formulation Eq. (2) can be obtained by solving the following GEP:

$$\begin{bmatrix} 0 & XY^T \\ YX^T & 0 \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda \begin{bmatrix} XX^T & 0 \\ 0 & YY^T \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix}, \tag{3}$$

where λ is the generalized eigenvalue corresponding to the generalized eigenvector (w_x, w_y). Usually, r-dimensional mappings ($W_x \in R^{d_x \times r}$, $W_y \in R^{d_y \times r}$) are given by selecting the top r generalized eigenvectors of Eq. (3) and $r \leq \min(d_x, d_y)$.

Additionally, the optimization problem of Eq. (2) can be rewritten in another equivalent form as

$$\begin{aligned} &\min_{w_x, w_y} \sum_{i=1}^{n} \left(w_x^T x_i - w_y^T y_i \right)^2 \\ &s.t. \sum_{i=1}^{n} \left(w_x^T x_i \right)^2 = 1, \quad \sum_{i=1}^{n} \left(w_y^T y_i \right)^2 = 1. \end{aligned} \tag{4}$$

CCA is not suitable for a semi-paired scenario according to its mathematical formulation (1). In order to make it adapt this scenario, recently, two semi-paired variants of CCA (i.e. SemiLRKCCA and SemiCCA) were separately developed by utilizing the structural information hidden in unpaired data in two views to regularize KCCA or CCA. And it has been shown that both have better performance than KCCA or CCA which just relies on a small amount of paired samples.

2.2 SemiLRKCCA: Semi-supervised Laplacian Regularization of KCCA

Based on KCCA [19] which is the kernelization of CCA for dealing with nonlinearly-correlated data, Blaschko et al. [16] developed a SemiLRKCCA through the manifold regularization technique [22] to tackle semi-paired scenario. In this paper without loss of generality in comparison to both our proposed NeCA and SemiCCA, we just concern SemiLRKCCAs linear version (named SemiLRCCA). However, their kernelizations to nonlinear counterparts are straightforward.

Now suppose we are given two sets of training samples: $X = [X_P X_U] = [x_1, \ldots, x_p, x_{p+1}, \ldots, x_{N_X}] \in R^{d_x \times N_x}$ and $Y = [Y_P Y_U] = [y_1, \ldots, y_p, y_{p+1}, \ldots, y_{N_Y}] \in R^{d_y \times N_y}$, where $X_P = [x_1, \ldots, x_p]$ and $Y_P = [y_1, \ldots, y_p]$ are paired samples; $X_U = [x_{p+1}, \ldots, x_{N_x}]$ and $Y_U = [y_{p+1}, \ldots, y_{N_y}]$ are unpaired samples; N_x (resp. N_y) is the sample number of X (resp. Y); d_x (resp. d_y) is the dimensionality of X (resp. Y). Then, SemiLRCCA can be expressed as the following optimization problem:

$$\begin{aligned} &\max \;\; w_x^T X_P Y_P^T w_y \\ &s.t. \;\; w_x^T \left(X_P X_P^T + \frac{\gamma_x}{N_x^2} X \hat{L}_X X^T \right) w_x + \varepsilon_x w_x^T w_x = 1, \\ &\quad w_y^T \left(Y_P Y_P^T + \frac{\gamma_y}{N_y^2} Y \hat{L}_Y Y^T \right) w_y + \varepsilon_y w_y^T w_y = 1. \end{aligned} \tag{5}$$

where $\hat{L}_x \left(= \left(D^X\right)^{-\frac{1}{2}}(D^X - S^X)(D^X)^{-\frac{1}{2}} \right)$ is the empirical graph Laplacian as defined in manifold learning and constructed by the N_x samples (paired and unpaired samples) of view $\mathbf{x}$, γ_x is regularization parameter, S^X is a similarity matrix of view $\mathbf{x}$, D^X is the diagonal matrix whose entries are the row or column sums of S^X; L_y and γ_y are defined similarly.

From Eq. (5), we can know that $\varepsilon_x w_x^T w_x$ and $\varepsilon_y w_y^T w_y$ both are Tikhonov regularization terms. Since Tikhonov regularization is the most commonly used method for overcoming the singularity problem, we can compactly rewrite Eq. (5) as Eq. (6) by omitting Tikhonov regularization terms $\varepsilon_x w_x^T w_x$ and $\varepsilon_y w_y^T w_y$.

$$\begin{aligned} &\max \quad w_x^T X_P Y_P^T w_y \\ &s.t. \quad w_x^T \left(X_P X_P^T + \gamma_x X L_x X^T\right) w_x = 1, \quad w_y^T \left(Y_P Y_P^T + \gamma_y Y L_y Y^T\right) w_y = 1. \end{aligned} \tag{6}$$

where $L_X = \frac{1}{N_x^2}\hat{L}_X$ and $L_Y = \frac{1}{N_y^2}\hat{L}_Y$. It is necessary to mention that the algorithms involved (CCA, SemiLRCCA (i.e. Eq. (6)), SemiCCA and our NCA) all use Tikhonov regularization in our experiments. Therefore, the performance of Eq. (6) is the same with that of Eq. (5).

Furthermore, for consistency and contrast with our work later, Eq. (6) can be rewritten as follows:

$$\begin{aligned} &\max_{w_x,\, w_y} \quad w_x^T X S^{Semi} Y^T w_y \\ &s.t. \quad w_x^T X D^{SemiX} X^T w_x = 1, \quad w_y^T Y D^{SemiY} Y^T w_y = 1. \end{aligned} \tag{7}$$

where $S^{Semi} = \begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix}$, $D^{SemiX} = \begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix} + \gamma_x L_x$, $D^{SemiY} = \begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix} + \gamma_y L_y$ and I_p is the $p \times p$ identity matrix.

Similar to CCA, through the Lagrange multiplier method and some mathematical manipulations, the solution of formulation (6) can be reduced to the following GEP:

$$\begin{bmatrix} 0 & X_P Y_P^T \\ Y_P X_P^T & 0 \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda \begin{bmatrix} X_P X_P^T + \gamma_x X L_x X^T & 0 \\ 0 & Y_P Y_P^T + \gamma_y Y L_y Y^T \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix}. \tag{8}$$

2.3 SemiCCA: Semi-supervised Learning of CCA

In order to mitigate the overfitting of CCA due to limited paired data, Kimura et al. [17] developed SemiCCA by combining CCA with PCA for utilizing unpaired samples. As a result, it is simply formulated as a combined eigenvalue problem of both CCA and PCA. This way, similar to CCA and SemiLRCCA, the solution of SemiCCA can also be attributed to solving the following GEP:

$$A \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda B \begin{bmatrix} w_x \\ w_y \end{bmatrix}, \tag{9}$$

where $A = \beta \begin{bmatrix} 0 & C_{xy}^P \\ (C_{xy}^P)^T & 0 \end{bmatrix} + (1-\beta) \begin{bmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{bmatrix}$, $B = \beta \begin{bmatrix} C_{xx}^P & 0 \\ 0 & C_{yy}^P \end{bmatrix} + (1-\beta) \begin{bmatrix} I_{d_x} & 0 \\ 0 & I_{d_y} \end{bmatrix}$; $C_{xx} = \frac{1}{n_x} X X^T$, $C_{yy} = \frac{1}{n_y} Y Y^T$, $C_{xx}^p = \frac{1}{n_p} X_P X_P^T$, $C_{yy}^p = \frac{1}{n_p} Y_P Y_P^T$, $C_{xy}^p = \frac{1}{n_p} X_P Y_P^T$; I_d is the $d \times d$ identity matrix, and β is the trade-off parameter

$(0 < \beta < 1)$. The first terms of A and B ensure the correlations between the paired data to be maximized and the second terms ensure the covariances of X and Y to be maximized respectively.

Furthermore, A and B can be rewritten as $A = \begin{bmatrix} 0 & C_{xy}^P \\ (C_{xy}^P)^T & 0 \end{bmatrix} + \eta \begin{bmatrix} C_{xx} & 0 \\ 0 & C_{\mathrm{yy}} \end{bmatrix}$ and $B = \begin{bmatrix} C_{xx}^P & 0 \\ 0 & C_{yy}^P \end{bmatrix} + \eta \begin{bmatrix} I_{d_x} & 0 \\ 0 & I_{d_y} \end{bmatrix}$, where $\eta = \frac{1-\beta}{\beta}$. Though no objective function was explicitly given in [17]. In fact, we still deduce the corresponding objective function to Eq. (9) as

$$\begin{aligned} \max_{w_x,\, w_y} \quad & w_x^T C_{xy}^P w_y + \eta \left(\frac{1}{2} w_x^T C_{xx} w_x + \frac{1}{2} w_y^T C_{yy} w_y \right) \\ s.t. \quad & w_x^T C_{xx}^P w_x + w_y^T C_{yy}^P w_y + \eta \left(w_x^T w_x + w_y^T w_y \right) = 1. \end{aligned} \tag{10}$$

Though Eqs. (3), (8) and (9) are all GEPs, Eqs. (3) and (8) can be solved in a decoupled way but Eq. (9) cannot, due to that the left-side matrix of the former is anti-diagonal but that of the latter is not. Consequently, solving the latter is more complex than solving the former. In the following section, we desire that our NeCA still has such a decouplable property in solution.

3 Neighborhood Correlation Analysis (NeCA)

In this section, we formally introduce our NeCA method. CCA and its variants (SemiLRCCA and SemiCCA) mainly consider the correlations between paired samples in different views (views **x** and **y**). Actually, in terms of the manifold assumption, it is reasonable to consider not only the correlations between paired samples in given two views but also the correlations between their respective neighborhoods, as shown in Fig. 1. To this end, the between-view neighborhood relationships are constructed by leveraging within-view neighborhood relationships of individual views and between-view paired data. Then NeCA is developed through incorporating between-view neighborhood relationships into CCA, which can take more sufficient advantage of unpaired samples in the way of two-view leaning. As a result, NeCA can consider the correlations between the samples in one view and their corresponding between-view local neighbors in the other view, in addition to the correlations between those paired samples. This way, it can work with partially-paired data with large number of the unpaired samples.

3.1 Constructing Within-View Neighborhood Graph

A neighborhood graph of view **x** is constructed as $G_X = \{X,\ S^X\}$ with a vertex set X and a within-view affinity weight matrix S^X, $S^X = (S_{ij}^X)_{N_x \times N_x}$ is defined as

$$S_{ij}^X = \begin{cases} \exp\left(-||x_i - x_j||^2 / 2\sigma_x^2\right) & x_i \in N_k\left(x_j\right) \vee x_j \in N_k\left(x_i\right) \vee i = j \\ 0 & \text{otherwise} \end{cases}, \tag{11}$$

where $N_k(x_i)$ denotes the k nearest neighbors of x_i. Similarly, a neighborhood graph of view **y** is constructed as $G_Y = \{Y,\ S^Y\}$, $S^Y = (S_{ij}^Y)_{N_y \times N_y}$ is defined as

$$S_{ij}^Y = \begin{cases} \exp\left(-||y_i - y_j||^2 / 2\sigma_y^2\right) & y_i \in N_k\left(y_j\right) \vee y_j \in N_k\left(y_i\right) \vee i = j \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

where $N_k(y_i)$ denotes the k nearest neighbors of y_i.

3.2 Constructing Between-View Neighborhood Graph

Note that views **x** and **y** are heterogeneous, therefore the construction of between-view neighborhood graph seems not so straightforward. Fortunately, the two within-view's neighborhood graph constructed in Sect. 3.1 can be used to guide the construction of between-view neighborhood graph.

Similar to [4], we assume that the neighbors of x_i in view **x** and the neighbors of the y_i in view **y** are between-view similar, where x_i and y_i are paired samples from the two views. Under the assumption, the between-view neighborhood graph can be constructed through leveraging the within-view neighborhood graphs of individual views and paired data and is defined as a bipartite graph $G_{XY} = \{X \cup Y,\ S^{XY}\}$ with a vertex set $X \cup Y$ and a between-view affinity weight matrix $S^{XY} \in R^{N_x \times N_y}$ which describes the affinity weights between samples in different views.

Then, we need to consider how to encode S^{XY} by utilizing both the within-view neighborhood relationships of views **x** and view **y** (i.e. S^X and S^Y) and paired information between different views. Facing such a semi-paired scenario, we adopt a method named SMD [4] to generate S^{XY}, since it can incorporate the within-view neighborhood relationships and the between-view paired information without artificially equating or relating them naturally. According to the definition of SMD, S^{XY} is defined as

$$S_{ij}^{XY} = S_{ij}^{Smd} = \sum_{h=1}^{p} S_{ih}^{X} \times S_{hj}^{Y}, \tag{13}$$

where p is the number of paired data, $i \in \{1, 2, \ldots, N_X\}$ and $j \in \{1, 2, \ldots, N_Y\}$. Intuitively, the term within the sum will be closer to one when x_i is close to x_h (i.e. S_{ih}^X is close to one) in view **x** and y_j is close to y_h (i.e. S_{hj}^Y is close to one) in view **y**. Thus, if x_i and y_j share many paired neighbors, S_{ij}^{XY} will be large. Then we can construct a full bipartite affinity matrix S^{XY} between views using Eq. (13) where h sums over only the p paired data $\{(x_1, y_1), \ldots, (x_p, y_p)\}$. Now S^{XY} can be more compactly written as a matrix form

$$S^{XY} = S^{Smd} = \tilde{S}^X \times \left(\tilde{S}^Y\right)^T, \tag{14}$$

where $\tilde{S}^X = S^X(:, 1 : p) \in R^{N_x \times p}$, $\tilde{S}^Y = S^Y(:, 1 : p) \in R^{N_y \times p}$. Figure 2 shows the graphical view of the encoding of S^{SMD}.

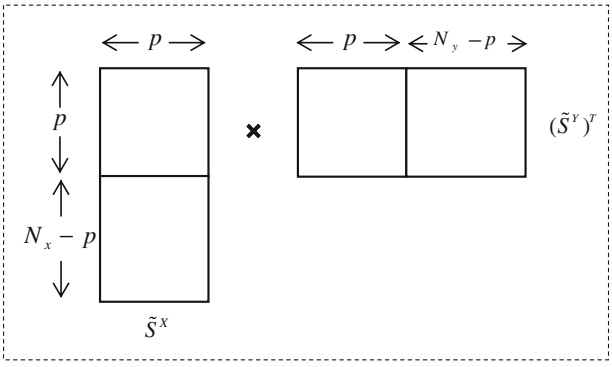

Fig. 2 A graphical *view* of the matrix multiplication required to compute S^{Smd}, when there are p paired samples in both views, $(N_x - p)$ samples only in view **x** and $(N_y - p)$ samples only in view **y** [4]

3.3 Formulation of NeCA

NeCA gets developed through incorporating between-view affinity weight matrix $S^{XY}(= S^{Smd})$ into the CCAs object function (4). Its object function is defined as

$$\min_{w_x,\, w_y} \sum_{i,\, j=1}^{N_x,\, N_y} \left(w_x^T x_i - w_y^T y_j\right)^2 S_{ij}^{Smd}. \tag{15}$$

Following some simple algebraic steps, we see that

$$\begin{aligned}
&\sum_i^{N_x}\sum_j^{N_y}\left(w_x^T x_i - w_y^T y_j\right)^2 S_{ij}^{Smd} \\
&= \sum_i^{N_x}\sum_j^{N_y}\left(w_x^T x_i S_{ij}^{Smd} x_i^T w_x + w_x^T y_j S_{ij}^{Smd} y_j^T w_x - 2w_x^T x_i S_{ij}^{Smd} y_j^T w_y\right) \\
&= \sum_i^{N_x} w_x^T x_i D_{ii}^{Row} x_i^T w_x + \sum_j^{N_y} w_y^T y_j D_{jj}^{Col} y_j^T w_y - 2\sum_i^{N_x}\sum_j^{N_y} w_x^T x_i S_{ij}^{Smd} y_j^T w_y \\
&= w_x^T X D^{Row} X^T w_x + w_y^T Y D^{Col} Y^T w_y - 2w_x^T X S^{Smd} Y^T w_y
\end{aligned}$$

where D^{Row} and D^{Col} are diagonal matrices and their entries are row and column sums of S^{Smd} respectively, i.e. $D_{ii}^{Row} = \sum_j^{N_y} S_{ij}^{Smd}$ and $D_{jj}^{Col} = \sum_i^{N_x} S_{ij}^{Smd}$. Therefore, the minimization problem (15) can be reduced to the following optimization problem:

$$\begin{aligned}
&\max_{w_x,\, w_y} \; w_x^T X S^{Smd} Y^T w_y \\
&s.t. \quad w_x^T X D^{Row} X^T w_x = 1, \quad w_y^T Y D^{Col} Y^T w_y = 1.
\end{aligned} \tag{16}$$

Finally, using Lagrange multiplier method as CCA, we can recast the optimization problem (16) as the following GEP (see Appendix):

$$\begin{bmatrix} 0 & XS^{Smd}Y^T \\ Y(S^{Smd})^T X^T & 0 \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda \begin{bmatrix} XD^{Row}X^T & 0 \\ 0 & YD^{Col}Y^T \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix}, \tag{17}$$

where λ is the generalized eigenvalue corresponding to the generalized eigenvector $(w_x,\ w_y)$. Consequently, NeCA, which inherits merit of CCA, can be easily solved by computing the GEP in the same decoupled way as CCA and SemiLRCCA. Now taking the top $r \leq \min(d_x, d_y)$ generalized eigenvectors, we obtain r-dimensional mappings ($W_x \in R^{d_x \times r}$, $W_y \in R^{d_y \times r}$) with respect to the two views. From Eq. (16) and Fig. 1, we can know that CCA is a special case of NeCA when it only considers the correlations between paired samples (i.e. X, Y and S^{Smd} are replaced by X_P, Y_P and a $p \times p$ identity matrix respectively in Eq. (16)).

3.4 Algorithm of NeCA

Based on the above formulation and solution, we summarize the specific algorithm of NeCA in Table 1.

Table 1 The algorithm of NeCA

Input: Semi-paired data:
$X = [x_1, \ldots, x_p, x_{p+1}, \ldots, x_{N_X}] \in R^{d_x \times N_x}$ and $Y = [y_1, \ldots, y_p, y_{p+1}, \ldots, y_{N_Y}] \in R^{d_y \times N_y}$, where $X_P = [x_1, \ldots, x_p]$ and $Y_P = [y_1, \ldots, y_p]$ are p paired data;
Output: Projection matrices: W_x, W_y

Procedure:

Step 1: Construct within-view affinity weight matrices S^X and S^X using Eqs. (11) and (12) respectively;

Step 2: Using Eq. (13) to construct between-view affinity weight matrix $S^{XY} (= S^{Smd})$ based on S^X, S^Y and p paired data;

Step 3: Obtain projection matrices w_x and w_y through solving generalized eigenvalue problem Eq. (17).

Step 4: Taking the top $r \leq \min(d_x, d_y)$ generalized eigenvectors, we obtain r-dimensional mappings ($W_x \in R^{d_x \times r}$, $W_y \in R^{d_y \times r}$).

4 Generalization of NeCA

NeCA utilizes those unpaired samples in the way of two-view learning differently from both SemiLRKCCA and SemiCCA in the way of single-view learning and has no any regularization in its optimization problem. In fact, NeCA can further be boosted in performance by the regularization using individual views' structure information just as both SemiLRCCA and SemiCCA do for CCA. As a result, two extensions of NeCA are presented: LRNeCA and PRNeCA. Their optimization problems are defined in Eqs. (18) and (19) respectively.

(1) *LRNeCA*

$$\begin{aligned} \max_{w_x,\, w_y} \quad & w_x^T X S^{Smd} Y^T w_y \\ s.t. \quad & w_x^T \left(X D^{Row} X^T + \gamma_x X L_x X^T \right) w_x = 1, \\ & w_y^T \left(Y D^{Col} Y^T + \gamma_y Y L_y Y^T \right) w_y = 1. \end{aligned} \tag{18}$$

(2) *PRNeCA*

$$\begin{aligned} \max_{w_x,\, w_y} \quad & w_x^T X S^{Smd} Y^T w_y + \eta \left(\frac{1}{2} w_x^T C_{xx} w_x + \frac{1}{2} w_y^T C_{yy} w_y \right) \\ s.t. \quad & w_x^T X D^{Row} X^T w_x + w_y^T Y D^{Col} Y^T w_y + \eta \left(w_x^T w_x + w_y^T w_y \right) = 1. \end{aligned} \tag{19}$$

From Eqs. (18) and (19), we can see that LRNeCA and PRNeCA utilize those unpaired samples in single-view and two-view learning ways simultaneously. As a result, they can take advantage of the between-view's neighborhood correlation information and each view's local or global structure information effectively. Additionally, LRNeCA and PRNeCA naturally degenerate to SemiLRCCA and SemiCCA respectively when only considering correlations between paired samples.

5 Experiments

In this section, we perform experiments for comparing the NeCA family (NeCA and its variants: LRNeCA and PRNeCA) with CCA family (CCA and its variants: SemiLRCCA and

SemiCCA). Their performance is evaluated by two-view classification experiments on different types of widely used multi-view datasets including WebKB, Internet Advertisements, MFD and Yale. Specifically, WebKB and Internet Advertisements both are two classes web page datasets (the former is two-view and the latter is multi-view); MFD is a multi-class handwritten digits dataset and Yale is a multi-class face dataset. In the following experiments, we mainly perform two kind of experimental comparisons: (1) NeCA versus CCA, where NeCA and CCA only utilize the between-view's correlation information; (2) LRNeCA versus SemiLRCCA and PRNeCA versus SemiCCA, where NeCA variants (LRNeCA and PRNeCA) and CCA variants (SemiLRCCA and SemiCCA) all utilize not only the between-view's correlation information but also the single-view's local or global structure information.

5.1 Evaluation Metric

For any test sample x_i in view $\mathbf{x}$ (resp. y_j in view $\mathbf{y}$), we firstly extract features $W_x^T x_i$ (resp. $W_y^T y_j$), then perform its classification based on $W_y^T Y_P$ (resp. $W_x^T X_P$), where Y_P are paired training data in view $\mathbf{y}$ (resp. X_P are paired training data in view $\mathbf{x}$). In this paper, the nearest neighbor classifier is employed to estimate the classification accuracies of different methods.

5.2 Parameter Selection

In both NeCA family and SemiLRCCA, the neighborhood size k is searched from $\{1, 2, \ldots, l\}$(l is smaller than the number of each class's training samples in each view), where l is empirically set as 20, 30, 20 and 5 for Internet Advertisements, WebKB, MFD and Yale respectively; the heat kernel width σ is set as $c \times \sigma_0$ (σ_0 is the mean norm of each view's training samples), where c is simply set as 1 for Internet Advertisements and MFD and c is searched from $\{2^{-4}, 2^{-3}, \ldots, 2^4\}$ for WebKB and Yale. In LRNeCA and SemiLRCCA, parameter $\gamma_x (= \gamma_y)$ is searched from $\{2^{-20}, 2^{-18}, \ldots, 2^{20}\}$. In PRNeCA and SemiCCA, parameter η is also searched from $\{2^{-20}, 2^{-18}, \ldots, 2^{20}\}$. We perform five-fold cross-validation to select the optimal parameters. The parameters sought corresponding to the best results in the validation are used in testing. In our experiments, Tikhonov regularization is used for all the algorithms involved (CCA, SemiLRKCCA, SemiCCA and our NCA).

5.3 Database Description

(1) *Internet Advertisements Dataset*. This dataset[1] is selected from UCI machine learning repository, which is composed of 3,279 web images (459 Ads. and 2,820 Non-ads.) with 1,558 attributes. All attributes, except four missing value, can be split into five sets covering urls and text descriptions. They are: (1) 472 attributes from ancurl terms, i.e. urls provided by images (Ancurl); (2) 111 attributes from alt terms, i.e. alter native text descriptions when some errors occur (Alt); (3) 19 attributes from caption terms, i.e. caption texts of images (Cap); (4) 495 attributes from origurl terms, i.e. original or source urls of images (Origurl); (5) 457 attributes from url terms, i.e. urls of web pages where the images are placed (Url). (2) *WebKB*. The WebKB course dataset[2] has been frequently used in the empirical study of multi-view learning since it was first introduced by Blum et al. [30]. The dataset contains 1,051 web pages collected from computer science departments of four universities. The pages are manually classified into two categories: course (230) and non-course (821). The dataset has two views which are the textual content of a web page (*page view*) and the

[1] The datasets are available from http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements.

[2] The datasets are available from http://www.cs.cmu.edu/afs/cs/project/theo-11/www/wwkb/.

Table 2 The average classification accuracies across 20 runs on Internet advertisements dataset and the corresponding standard deviations

	CCA and its variants			NeCA and its variants		
	CCA	SemiLRCCA	SemiCCA	NeCA	LRNeCA	PRNeCA
1						
Ancurl	69.65 ± 8.5	72.45 ± 2.5	**82.05 ± 9.3**	75.25 ± 7.5	73.65 ± 8.1	80.25 ± 7.3
Altcap	61.15 ± 7.6	65.10 ± 2.3	63.75 ± 7.5	64.25 ± 5.7	**65.70 ± 2.8**	65.50 ± 6.1
2						
Ancurl	77.90 ± 6.4	79.10 ± 4.1	86.00 ± 3.9	81.05 ± 6.7	80.30 ± 4.7	**86.60 ± 4.8**
Origurl	65.15 ± 2.6	66.90 ± 5.2	63.40 ± 5.0	69.00 ± 5.3	68.70 ± 2.7	**69.05 ± 4.9**
3						
Ancurl	79.80 ± 5.2	80.25 ± 4.7	85.35 ± 4.1	81.20 ± 7.2	79.40 ± 4.1	**86.25 ± 4.7**
Url	72.10 ± 8.5	75.30 ± 7.0	70.90 ± 8.0	74.60 ± 7.1	76.95 ± 6.9	**77.80 ± 7.8**
4						
Url	69.90 ± 8.4	71.55 ± 8.0	74.45 ± 8.0	71.00 ± 8.2	72.55 ± 7.4	**75.65 ± 5.9**
Origurl	59.75 ± 5.1	66.75 ± 6.3	60.00 ± 5.6	62.85 ± 7.4	**67.05 ± 6.8**	64.15 ± 4.8
5						
Url	66.30 ± 8.0	67.45 ± 7.6	67.75 ± 8.4	66.15 ± 8.5	67.85 ± 8.2	**69.65 ± 7.8**
Altcap	59.75 ± 6.6	**66.40 ± 2.6**	64.25 ± 1.7	62.00 ± 7.5	65.65 ± 4.9	65.30 ± 3.8
6						
Altcap	64.10 ± 6.4	65.20 ± 4.6	**65.80 ± 2.5**	59.70 ± 6.8	63.30 ± 5.1	64.55 ± 5.7
Origurl	56.85 ± 2.7	60.00 ± 5.4	58.70 ± 5.5	59.45 ± 6.2	60.50 ± 6.3	**61.80 ± 6.5**

words that occur in the hyperlinks of other web pages pointing to that web page (*link view*). We borrowed a processed WebKB course dataset from Sindhwani et al. [31] and used it in our experiment. For the page representation, 3,000 features were selected according to information gain. For the link representation, 1,840 features were generated with no feature selection. The first 200 samples are selected for each class in a balanced manner. Each view's dimension is preprocessed by PCA [20,21] to 100. (3) *MFD.* The MFD[3] is picked out from UCI machine learning repository. It is composed of six feature sets of handwritten digits from 0 to 9. Each class contains 200 samples. The six sets are flourier coefficient (Fou), contour correlation characteristics (Fac), Karhunen-Loève expansion coefficient (Kar), pixel average (Pix), Zernike moment (Zer) and morphological characteristics (Mor), and their dimension are 76, 216, 64, 240, 47 and 6 respectively. In this section, we evaluate the effectiveness of the proposed NeCA. (4) *Yale.* This database[4] [32] contains 165 face images of 15 individuals. There are 11 images per subject, and these 11 images are respectively under the following different facial expression or configuration: center-light, wearing glasses, happy, left-light, wearing no glasses, normal, right-light, sad, sleepy, surprised, and wink. Furthermore, in the following experiments, we use the centered data based on all the training samples for each dataset.

[3] The datasets are available from http://www.ics.uci.edu/~mlearning/MLSummary.html.

[4] The datasets are available from http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

5.4 Internet Advertisements Dataset

According to the five feature sets of Internet Advertisements Dataset, we construct four views: (1) Ancurl; (2) AltCap (which is composed of Alt and Cap due to that they both belong to the text descriptions of images); (3) Origurl; (4) Url. We choose any two views from the four views as views **x** and **y** respectively, and thus there are six two-view combinations in total. The first 200 samples are selected for each class in a balanced manner. Fifty percent of the samples of each class are randomly selected for training and the rest for testing. Further, we randomly select 10 % in the training set as the paired data and the rest as the unpaired. Table 2 reports the accuracies averaged over 20 independent trials, for the subspace dimensionality that provided the highest accuracy.

From the Table 2 (in which the best performances are bolded), we can obtain several attractive observations:

(1) NeCA versus CCA. NeCA achieves better classification accuracy on 10 out of the 12 cases (view combinations 6 × 2 cases/combination) and improves over 3 % on 6 cases.
(2) LRNeCA versus SemiLRCCA: LRNeCA wins nine cases. PRNeCA versus SemiCCA: PRNeCA wins 10 cases and improves over 3 % on 4 cases.
(3) NeCA family versus CCA family: NeCA family wins nine cases.

5.5 WebKB Course Dataset

In this experiment, 50 % web pages of each category are randomly selected for training and the rest for testing. Further, three groups of experiments are performed, where 5 %, 10 % and 15 % of the training set are respectively selected as the paired data and the rest as the unpaired data. Table 3 reports the accuracies averaged over 20 independent trials, for the subspace dimensionality that provided the highest accuracy.

From the Table 3 (in which the best performances are bolded), we can obtain several attractive observations:

(1) NeCA versus CCA: NeCA achieves better classification accuracy on all cases and improves over 4 % on four cases.

Table 3 The average classification accuracies across 20 runs on WebKB and the corresponding standard deviations

	CCA and its variants			NeCA and its variants		
	CCA	SemiLRCCA	SemiCCA	NeCA	LRNeCA	PRNeCA
5 %						
Page	76.80 ± 6.7	79.20 ± 4.1	77.10 ± 6.7	77.15 ± 6.6	80.00 ± 5.2	**84.60 ± 6.9**
Link	76.45 ± 8.9	77.00 ± 8.1	79.00 ± 6.8	80.35 ± 8.6	**82.55 ± 4.4**	82.35 ± 3.8
10 %						
Page	77.15 ± 7.9	81.70 ± 2.8	83.65 ± 4.7	81.10 ± 5.1	83.70 ± 5.3	**90.15 ± 1.4**
Link	77.95 ± 8.0	81.05 ± 7.6	85.60 ± 2.6	82.70 ± 3.6	83.75 ± 3.1	**86.45 ± 3.3**
15 %						
Page	81.95 ± 5.7	82.15 ± 6.5	83.50 ± 5.2	83.80 ± 3.9	84.60 ± 3.6	**91.45 ± 3.1**
Link	79.35 ± 8.6	83.05 ± 8.8	85.75 ± 2.3	84.00 ± 3.0	84.20 ± 2.0	**87.00 ± 2.2**

(2) LRNeCA versus SemiLRCCA: LRNeCA wins all cases. PRNeCA versus SemiCCA: PRNeCA wins all cases and improves over 6 % on nine cases.
(3) NeCA family versus CCA family: NeCA family wins all cases.

5.6 Multiple Feature Handwritten Digit Database (MFD)

According to the first five feature sets of MFD, we construct five views: Fou, Fac, Kar, Pix and Zer. We don't consider the sixth feature set Mor, sine it only has 6 dimensions. We select two views from the five views as views **x** and **y** respectively, and so there are ten view combinations. Fifty samples of each class are selected for training and the rest for testing at random. Table 4 reports the accuracies averaged over 20 independent rounds, for the subspace dimensionality that provided the highest accuracy. During every round, we randomly choose 10 % of the each class's training samples as paired data.

From the Table 4 (in which the best performances are bolded), we can obtain the following observations:

(1) NeCA versus CCA: NeCA achieves better classification accuracy on 16 out of the 20 cases (view combinations 10×2 cases/combination) and NeCA improves over 20 and 10 % on 3 and 6 cases respectively.
(2) LRNeCA versus SemiLRCCA: LRNeCA wins 19 cases and improves over 3 % on 8 cases. PRNeCA versus SemiCCA: PRNeCA wins all cases and improves over 3 % on eight cases.
(3) NeCA family versus CCA family: NeCA family wins 19 cases.

5.7 Yale Database

In this experiment, we use the cropped 32×32 images which can be considered as the first view. To construct multi-view data, we provide another two representations of each image: (1) one from down-sampling to 16×16 pixels as the second view, since images in different resolutions can provide information at different levels; (2) the other from Local Binary Pattern Code (LBPC) [33] as the third view. So there are three view combinations in total for views **x** and **y**. To deal with the small sample size problem, each view is preprocessed by PCA with 98 % energy kept ratio.

The face images are divided into different training and test sets, and the training set is sub-partitioned into paired and unpaired sets. Concretely, eight images of each individual are randomly selected for training and the rest for testing; two groups of experiments are performed, where p ($= 2$ and 3, respectively) samples in the training set are selected as the paired data and the rest as the unpaired. Table 5 reports the accuracies averaged over 20 independent trials, for the subspace dimensionality that provided the highest accuracy.

From the Table 5 (in which the best performances are bolded), we can obtain the following observations:

(1) NeCA versus CCA: NeCA achieves better classification accuracy on all 12 cases and improves over 4 % on 9 cases.
(2) LRNeCA versus SemiLRCCA: LRNeCA wins 10 cases and equals 2 cases. PRNeCA versus SemiCCA: PRNeCA win all cases and improves over 4 % on eight cases.
(3) NeCA family versus CCA family: NeCA family wins 10 cases and equals 2 cases.

Table 4 The average classification accuracies across 20 runs on MFD and the corresponding standard deviations

	CCA and its variants			NeCA and its variants		
	CCA	SemiLRCCA	SemiCCA	NeCA	LRNeCA	PRNeCA
1						
Fac	79.54 ± 2.6	81.90 ± 3.4	79.55 ± 2.7	76.00 ± 3.9	**82.27** ± 3.4	80.93 ± 4.3
Fou	41.17 ± 4.9	64.98 ± 3.1	65.32 ± 2.9	64.85 ± 3.6	**69.39** ± 2.8	68.25 ± 2.6
2						
Fac	79.39 ± 2.8	84.57 ± 1.5	80.01 ± 2.3	80.76 ± 2.9	**84.69** ± 2.9	81.49 ± 3.0
Kar	56.23 ± 3.7	81.87 ± 2.0	80.26 ± 2.3	78.96 ± 3.2	**83.22** ± 2.1	82.99 ± 2.4
3						
Fac	79.15 ± 2.7	83.87 ± 2.7	80.10 ± 2.2	79.97 ± 3.3	**84.19** ± 2.6	82.23 ± 3.6
Pix	72.13 ± 4.3	75.35 ± 3.1	82.21 ± 2.1	69.73 ± 3.1	81.19 ± 2.0	**84.31** ± 2.3
4						
Fac	69.45 ± 3.7	**82.69** ± 2.5	73.94 ± 3.7	76.29 ± 3.1	81.96 ± 3.3	77.37 ± 3.4
Zer	58.99 ± 3.5	67.97 ± 2.2	63.11 ± 2.9	68.47 ± 2.8	**69.24** ± 2.7	68.85 ± 2.7
5						
Fou	62.79 ± 4.1	66.48 ± 2.6	64.15 ± 4.3	64.11 ± 3.3	**69.80** ± 3.1	67.18 ± 3.2
Kar	59.72 ± 4.2	75.81 ± 4.2	77.94 ± 2.7	74.17 ± 2.8	79.56 ± 3.2	**80.10** ± 2.6
6						
Fou	63.43 ± 3.3	64.49 ± 3.6	64.60 ± 2.3	65.80 ± 2.8	**68.61** ± 2.4	67.67 ± 2.9
Pix	67.35 ± 4.6	72.77 ± 2.1	76.01 ± 2.7	63.07 ± 3.0	76.89 ± 2.4	**78.47** ± 3.5
7						
Fou	49.09 ± 5.1	63.69 ± 2.6	62.04 ± 3.3	62.03 ± 3.6	**66.31** ± 2.4	64.72 ± 3.5
Zer	62.60 ± 3.6	66.18 ± 2.6	64.92 ± 1.8	65.61 ± 3.5	**68.27** ± 2.5	66.96 ± 2.1
8						
Kar	79.61 ± 2.0	82.41 ± 1.9	84.37 ± 2.0	81.23 ± 2.5	85.30 ± 2.2	**85.73** ± 2.2
Pix	80.27 ± 2.3	83.05 ± 1.8	83.25 ± 1.5	73.88 ± 2.0	**85.57** ± 2.0	84.17 ± 2.4
9						
Kar	47.87 ± 3.7	77.11 ± 3.7	72.62 ± 3.8	72.39 ± 4.5	**78.67** ± 3.3	74.88 ± 3.5
Zer	63.45 ± 3.9	67.17 ± 3.2	64.54 ± 2.6	67.26 ± 3.4	**69.55** ± 2.7	67.49 ± 3.0
10						
Pix	60.74 ± 4.1	74.03 ± 2.8	73.09 ± 5.0	62.73 ± 4.2	**76.69** ± 2.6	75.30 ± 3.9
Zer	65.35 ± 2.6	67.63 ± 3.2	65.82 ± 2.7	68.81 ± 3.2	**69.29** ± 3.0	68.91 ± 3.4

5.8 Impaction of Different Paired Ratio on the Performance of NeCA and CCA

In this subsection, we perform experiments on MFD for analyzing the impact of different paired samples ratio on NeCA and CCA. The classification accuracies averaged over that of the two view (views **x** and **y**) are employed for comprehensive analyzing the performance of the two methods, since they both are two-view learning methods.

Figure 3 shows the comparisons of the classification accuracies of CCA and NeCA with different proportion of paired samples. Here, we mainly consider the following four view combinations: Fac and Fou, Fac and Kar, Fou and Zer, Kar and Zer.

Table 5 The average classification accuracies across 20 runs on yale and the corresponding standard deviations

	CCA and its variants			NeCA and its variants		
	CCA	SemiLRCCA	SemiCCA	NeCA	LRNeCA	PRNeCA
$p = 2$						
1						
32×32	68.22 ± 8.4	74.22 ± 5.9	67.56 ± 8.1	71.78 ± 8.0	$\mathbf{74.44 \pm 5.8}$	72.44 ± 6.9
16×16	68.44 ± 6.9	75.11 ± 6.9	69.78 ± 7.8	74.00 ± 6.7	$\mathbf{76.33 \pm 4.1}$	74.67 ± 6.6
2						
32×32	66.89 ± 8.9	74.44 ± 5.9	70.22 ± 7.1	70.89 ± 6.0	$\mathbf{76.00 \pm 6.3}$	73.33 ± 6.2
LBPC	71.78 ± 9.8	82.00 ± 6.8	67.33 ± 8.8	76.67 ± 9.0	$\mathbf{82.89 \pm 6.3}$	77.33 ± 9.2
3						
16×16	68.44 ± 6.0	$\mathbf{76.22 \pm 6.7}$	72.22 ± 7.9	74.67 ± 6.4	$\mathbf{76.22 \pm 6.3}$	74.44 ± 6.4
LBPC	71.78 ± 9.8	82.00 ± 6.3	67.78 ± 8.5	76.67 ± 9.0	$\mathbf{82.89 \pm 6.3}$	75.56 ± 9.2
$p = 3$						
1						
32×32	71.11 ± 7.3	77.11 ± 7.4	72.67 ± 6.7	75.11 ± 6.9	$\mathbf{77.33 \pm 6.9}$	76.44 ± 6.3
16×16	72.44 ± 7.3	77.56 ± 7.0	71.56 ± 6.5	76.67 ± 6.2	$\mathbf{78.22 \pm 5.6}$	75.78 ± 5.3
2						
32×32	69.56 ± 7.3	77.33 ± 6.6	76.22 ± 9.4	73.33 ± 6.7	$\mathbf{79.33 \pm 7.0}$	77.11 ± 8.7
LBPC	75.78 ± 9.4	83.33 ± 5.9	70.44 ± 6.3	77.11 ± 7.3	$\mathbf{83.56 \pm 5.8}$	77.33 ± 8.3
3						
16×16	72.67 ± 6.7	$\mathbf{79.33 \pm 5.8}$	74.22 ± 9.9	77.33 ± 6.3	$\mathbf{79.33 \pm 5.8}$	76.22 ± 3.3
LBPC	77.11 ± 8.8	82.89 ± 6.3	72.00 ± 8.4	78.22 ± 7.2	$\mathbf{83.11 \pm 6.3}$	76.77 ± 7.4

According to the Fig. 3, we can obtain the following observations:

(1) With the increase of the paired samples ratio, the classification accuracies of CCA and NeCA are also improving generally.
(2) When the paired sample proportion is low, NeCA performs much better than CCA. This indicates that NeCA can effectively mitigate CCAs overfitting problem caused by limited paired data.

5.9 Summary of Experimental Results

According to the experimental results on different types of datasets shown from Sect. 5.4 to 5.8, we have the following general observations:

(1) NeCA performs much better than CCA in general, stating that NeCA can effectively mitigate overfitting caused by limited paired data and has more discriminative power through the utilization of between-view neighborhood relationships. Moreover, this also indicates that NeCA can make more sufficient use of the unpaired data to improve its performance.
(2) NeCAs variants (LRNeCA and PRNeCA) respectively outperform their corresponding CCAs variants (SemiLRCCA and SemiCCA) in most cases, which shows that NeCAs variants really benefit from the exploitation for the unpaired data simultaneously in single-view and two-view learning ways rather than only in the single-view learning way in CCAs variants for the unpaired data.

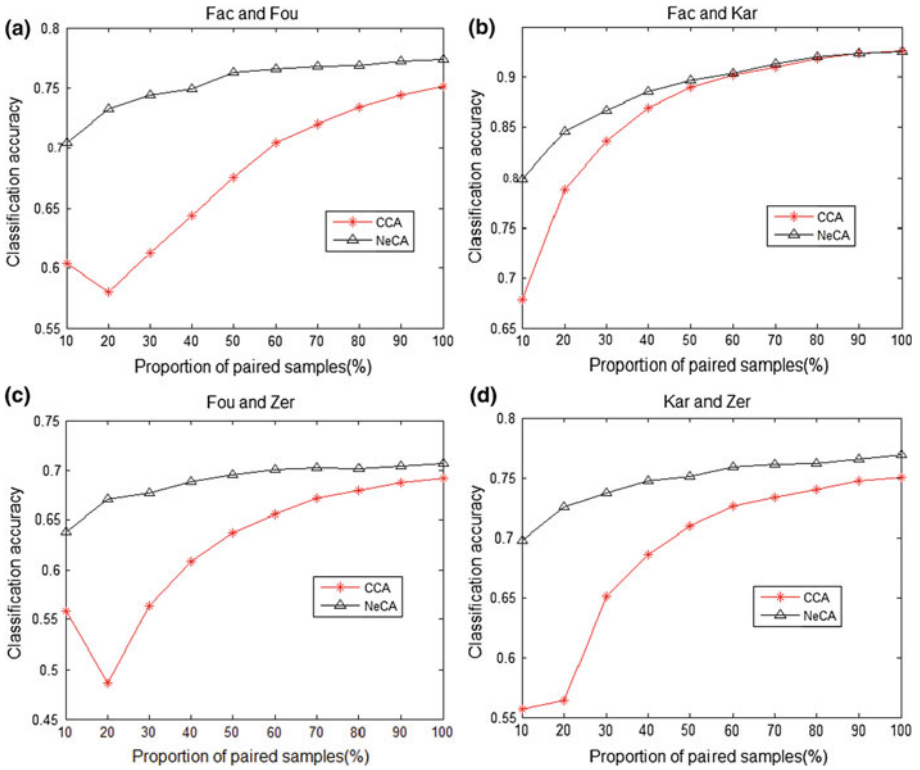

Fig. 3 Comparisons of the classification accuracies of CCA and NeCA with different proportion of paired samples, where the classification accuracies are averaged over the ones of the two view (views **x** and **y**)

(3) When the paired sample proportion is low, NeCA performs much better than CCA which indicates that NeCA can effectively mitigate CCAs overfitting problem caused by limited paired data.

6 Conclusions and Future Work

In this paper, we have proposed a novel semi-paired variant of CCA named NeCA. It can work with semi-paired data and take more sufficient advantage of the unpaired samples by constructing the correlations between samples in one view and their between-view local neighbors in the other view. As a result, NeCA can effectively mitigate CCAs overfitting problem caused by limited paired data. Furthermore, considering that NeCA can use those unpaired samples in a two-view leaning way, thus we further present its two regularization versions (LRNeCA and PRNeCA) for better performance. Experimental results on four different datasets show that NeCA family performs better than CCA family.

Our further works include: (1) Our NeCA is currently an unsupervised dimension reduction method. When supervised information can be available, incorporating class label information into the NeCA for extracting more discriminative features will be considered. (2) Though in this paper, we focus on dimension reduction in semi-paired scenario, when targeted at a classification task, how to design a classifier directly under such a scenario is an

interesting topic. (3) NeCA is currently a linear dimension reduction method. Actually, it can easily be extended to work in a nonlinear feature space by the powerful kernel methods. (4) Though NeCA is now still a two-view learning method, the extension to multi-view scenario is straightforward by the pair-view combination. (5) Under semi-paired scenario, the construction of between-view neighborhood relationships in NeCA can be applied to the work [28] for establishing its between-view neighborhood relationships more faster and perhaps more accurately.

Acknowledgements The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. This work was partly supported by National Natural Science Foundation of China under Grant No: 61170151, Natural Science Foundation of Jiangsu China under Grant No: BK2011728 and Natural Science Research Project of Higher Education of Jiangsu China under Grant No: 12KJB520018.

Appendix: Solving Optimization (15) Through Lagrange Multiplier Method

$$\begin{aligned} \max_{w_x,\, w_y} \;& w_x^T X S^{Smd} Y^T w_y \\ s.t.\;\; & w_x^T X D^{Row} X^T w_x = 1,\; w_y^T Y D^{Col} Y^T w_y = 1. \end{aligned}$$

Defining its Lagrangian function as

$$\begin{aligned} L\left(w_x,\, w_y\right) = w_x^T X S^{Smd} Y^T w_y - \frac{\lambda_x}{2}\left(w_x^T X D^{Row} X^T w_x - 1\right) \\ -\frac{\lambda_y}{2}\left(w_y^T Y D^{Col} Y^T w_y - 1\right). \end{aligned}$$

Taking derivatives with respect to w_x and w_y respectively, we obtain

$$\frac{\partial L\left(w_x,\, w_y\right)}{\partial w_x} = X S^{Smd} Y^T w_y - \lambda_x X D^{Row} X^T w_x = 0 \tag{19.1}$$

$$\frac{\partial L\left(w_x,\, w_y\right)}{\partial w_y} = Y\left(S^{Smd}\right)^T X^T w_x - \lambda_y Y D^{Col} Y^T w_y = 0 \tag{19.2}$$

Subtracting the second equation premultiplied by w_y^T from the first one premultiplied by w_x^T, we have

$$\begin{aligned} & w_x^T X S^{Smd} Y^T w_y - \lambda_x w_x^T X D^{Row} X^T w_x - w_y^T Y\left(S^{Smd}\right)^T X^T w_x + \lambda_y w_y^T Y D^{Col} Y^T w_y \\ & = \lambda_y w_y^T Y D^{Col} Y^T w_y - \lambda_x w_x^T X D^{Row} X^T w_x = 0, \end{aligned}$$

which together with the constraints leads to $\lambda_y - \lambda_x = 0$. Now let $\lambda_y = \lambda_x = \lambda$, we have

$$X S^{Smd} Y^T w_y = \lambda X D^{Row} X^T w_x \tag{20.1}$$

$$Y\left(S^{Smd}\right)^T X^T w_x = \lambda Y D^{Col} Y^T w_y \tag{20.2}$$

As a result, we can recast the optimization problem (15) as the following GEP:

$$\begin{bmatrix} 0 & X S^{Smd} Y^T \\ Y (S^{Smd})^T X^T & 0 \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda \begin{bmatrix} X D^{Row} X^T & 0 \\ 0 & Y D^{Col} Y^T \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix}.$$

References

1. McFee B, Lanckriet G (2011) Learning multi-modal similarity. J Mach Learn Res 12:491–523
2. Hou C, Zhang C, Wu Y, Nei F (2010) Multiple view semi-supervised dimensionality reduction. Pattern Recognit 43(3):720–730
3. Bickel S, Scheffer T (2004) Multi-view clustering. In: International conferrence on data mining (ICDM), pp 19–26
4. de Sa Virginia R, Gallagher Patrick W, Lewis Joshua M, Malave Vicente L (2010) Multi-view kernel construction. Mach Learn 79(1–2):47–71
5. Ando KR, Zhang T (2007) Two-view feature generation model for semi-supervised learning. In: International conference on machine learning (ICML), pp 25–32
6. Li G, Hoi Steven CH, Chang K (2010) Two-View transductive support vector machines. In: SIAM international conference on data mining (SDM), pp 235–244
7. Szedmaka S, Shawe-Taylorc J (2007) Synthesis of maximum margin and multiview learning using unlabeled data. Neurocomputing 70(7–9):1254–1264
8. Hardoon DR, Szedmaka S, Shawe-Taylor J (2004) Canonical correlation analysis: an overview with application to learning method. Neural Comput 16(12):2639–2664
9. Correa NM, Eichele T, Adalı T, Li Y-O et al (2010) Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI. NeuroImage 50(4):1438–1445
10. Chaudhuri K, Kakade SM, Livescu K, Sridharan K (2009) Multi-view clustering via canonical correlation analysis. In: International conference on machine learning (ICML), pp 129–136
11. Suna Q, Zeng S, Liu Y, Heng P et al (2005) A new method of feature fusion and its application in image recognition. Pattern Recognit 38(12):2437–2448
12. Hardoon DR, Shawe-Taylor J (2011) Sparse canonical correlation analysis. Machine Learning 83(3):331–353
13. Hotelling Harold (1936) Relations between two sets of variates. Biometrika 28(3–4):321–377
14. Zhu X (2008) Semi-supervised learning literature survey. Technical Report, Computer Sciences, University of Wisconsin-Madison
15. Chapelle O, Schölkopf B, Zien A (2006) Semi-supervised learning. MIT Press, Cambridge
16. Blaschko MB, Lampert CH, Gretton A (2008) Semi-supervised Laplacian regularization of kernel canonical correlation analysis. In: European conference on machine learning and knowledge discovery in databases (ECML PKDD), pp 133–145
17. Kimura A, Kameoka H, Sugiyama M, Nakano T (2010) SemiCCA: efficient semi-supervised learning of canonical correlations. In: International conference on pattern recognition (ICPR), pp 2933–2936
18. Chen X, Chen S, Xue H, Zhou X (2012) A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data. Pattern Recognit 45(5):2005–2018
19. Melzer T, Reiter M, Bischof H (2003) Appearance models based on kernel canonical correlation analysis. Pattern Recognit 39(9):1961–1971
20. Mackiewicz A, Ratajczak W (1993) Principal components analysis (PCA). Comput Geosci 19:303–342
21. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86
22. Belkin M, Niyogi P, Sindhwani V, Bartlett P (2006) Manifold regularization: a geometric framework for learning from examples. J Mach Learn Res 7:2399–2434
23. Peng Y, Zhang D, Zhang J (2010) A new canonical correlation analysis algorithm with local discrimination. Neural Process Lett 31(1):1–15
24. Aria H, Liang P, Berg-kirkpatrick T, Klein D (2008) Learning bilingual lexicons from monolingual corpora. In: Annual meeting of the Association for Computational Linguistics, pp 771–779
25. Tripathi A, Klami A, Virpioja S (2010) Bilingual sentence matching using kernel CCA. In: IEEE international workshop on machine learning for signal processing (MLSP), pp 130–135
26. Tripathi A, Klami A, Orešič M, Kaski S (2011) Matching samples of multiple views. Data Min Knowl Discov 23(2):300–321
27. Yamada M, Sugiyama M (2011) Cross-domain object matching with model selection. In: International conference on artificial intelligence and statistics (AISTATS)
28. Wang C, Mahadevan S (2009) Manifold alignment without correspondence. In: international jont conference on artifical intelligence, pp 1273–1278
29. Vía J, Santamaría I, Pérez J (2007) A learning algorithm for adaptive canonical correlation analysis of several data sets. Neural Netw 20(1):139–152
30. Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: International conference on learning theory (COLT), pp 92–100

31. Sindhwani V, Niyogi P, Belkin M (2005) Beyond the point cloud: from transductive to semi-supervised learning. In: International conference on machine learning (ICML), pp 824–831
32. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
33. Mäenpää T, Ojala T, Pietikäinen M, Soriano M (2000) Robust texture classification by subsets of local binary patterns. In: International conference on pattern recognition (ICPR), pp 3947–3950