Neural Process Lett (2009) 30:103–112
DOI 10.1007/s11063-009-9112-6

A New Discriminant Principal Component Analysis Method with Partial Supervision

Dan Sun · Daoqiang Zhang

Published online: 10 July 2009

Abstract Principal component analysis (PCA) is one of the most widely used unsupervised dimensionality reduction methods in pattern recognition. It preserves the global covariance structure of data when labels of data are not available. However, in many practical applications, besides the large amount of unlabeled data, it is also possible to obtain partial supervision such as a few labeled data and pairwise constraints, which contain much more valuable information for discrimination than unlabeled data. Unfortunately, PCA cannot utilize that useful discriminant information effectively. On the other hand, traditional supervised dimensionality reduction methods such as linear discriminant analysis perform on only labeled data. When labeled data are insufficient, their performances will deteriorate. In this paper, we propose a novel discriminant PCA (DPCA) model to boost the discriminant power of PCA when both unlabeled and labeled data as well as pairwise constraints are available. The derived DPCA algorithm is efficient and has a closed form solution. Experimental results on several UCI and face data sets show that DPCA is superior to several established dimensionality reduction methods.

Keywords Principal component analysis (PCA) · Discriminant PCA · Dimensionality reduction · Semi-supervised dimensionality reduction · Partial supervision

1 Introduction

With the rapid accumulation of high-dimensional data such as digital images, web documents and gene expression microarrays, dimensionality reduction has been a fundamental tool for many pattern recognition tasks. According to whether supervised information is available

D. Sun · D. Zhang (✉)
Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, 210016 Nanjing, China
e-mail: dqzhang@nuaa.edu.cn

D. Sun
e-mail: dansun@nuaa.edu.cn

or not, existing dimensionality reduction methods can be roughly categorized into supervised ones and unsupervised ones. Linear discriminant analysis (LDA) [1] and principal component analysis (PCA) [2] may be the most well-known supervised and unsupervised dimensionality reduction methods respectively. The former extracts the optimal discriminant vectors when class labels are available, while the latter seeks projective vectors to preserve the global covariance structure when class labels are not available. In this paper, we consider the following interesting problem, i.e. when both labeled and unlabeled data are available, how should we perform dimensionality reduction? That problem arises naturally in many practical pattern recognition applications, where unlabeled training data are readily available but labeled ones are fairly expensive to obtain [3,4]. That is, we are often confronted with problems with large amount of unlabeled data but only a few labeled data. Typically, those labeled data contain much more valuable information for discrimination than unlabeled data.

Unfortunately, neither traditional unsupervised dimensionality reduction methods such as PCA nor supervised dimensionality reduction methods such as LDA can well deal with the above dimensionality reduction problems. On one hand, PCA is unsupervised, and it can not use the useful discriminant information in those labeled data. On the other hand, LDA performs on only labeled data. When labeled data are sufficient enough, LDA will nearly always outperform PCA. In contrast, when the number of labeled data per class is so small that labeled data can not reflect the underlying distribution, the generalization performances of LDA on unseen samples will not be guaranteed and PCA might outperform LDA. To overcome the disadvantages of both PCA and LDA, a natural idea is to simultaneously use both unlabeled data and discriminant information in labeled data for dimensionality reduction. More specifically, we can either introduce unlabeled data into LDA, or introduce discriminant information in labeled data into PCA. In this paper, we focus on the latter case.

In this paper, we propose the discriminant PCA model (DPCA), which exploits both labeled and unlabeled data for dimensionality reduction. DPCA inherits from PCA the characteristic of structure preserving on unlabeled data, and has the new discriminant power by using the discriminant information in labeled data. The derived DPCA algorithm is efficient and has a closed form solution. Moreover, DPCA algorithm has the capability to use external knowledge provided by the user, such as pairwise constraints which specify whether a pair of instances belong to the same class (*must-link* constraint) or different classes (*cannot-link* constraint) [5,6]. Experimental results on several UCI and face data sets show that DPCA outperforms several established dimensionality reduction methods. The rest of this paper is organized as follows: Sect. 2 presents some related work in semi-supervised dimensionality reduction. The detailed DPCA algorithm is introduced in Sect. 3. Section 4 reports on the experimental results. Finally, Sect. 5 concludes this paper with some future work.

2 Related Works

In fact, the idea of using both labeled and unlabeled data for learning is not novel in machine learning. There has appeared a new branch in machine learning called semi-supervised learning whose main concern is to learn from a combination of both labeled and unlabeled data [3–5,7]. Because of its success in many practical applications such as text categorization [3], semi-supervised learning has received much attention in recent years. Current researches on semi-supervised learning could be roughly categorized into three classes, i.e. semi-supervised classification [3], semi-supervised regression [4] and semi-supervised clustering [5]. Research advances of semi-supervised learning can be found in an excellent recent survey [7].

Recently, some research works which utilize both labeled and unlabeled data for semi-supervised dimensionality reduction have appeared. For example, Yu et al. [8] proposed a supervised probabilistic PCA model and a semi-supervised probabilistic PCA model, and the latter can incorporate both labeled and unlabeled data for dimensionality reduction. However, their method is based on probabilistic PCA which is a generative model. Also, their algorithm needs iteration and has no closed form solution. Lu et al. [9] proposed a novel hybrid dimension reduction scheme to merge LDA and PCA in a unified framework. In addition, many subspace learning algorithms such as spectral regression discriminant analysis method [10,11] and semi-supervised discriminant analysis method [12] have been proposed. Specifically, Cai et al. [12] proposed the semi-supervised discriminant analysis method called SDA which utilized local neighborhood information of labeled data for dimensionality reduction. However, the number of neighborhood in SDA is still hard to set. Besides SDA, $\text{SSDA}_{\text{CCCP}}$ is a diverse semi-supervised discriminant analysis algorithm proposed by Zhang et al. [13]. It uses the constrained concave–convex procedure (cccp) to maximize an optimality criterion of LDA which leads to estimation of the class labels for the unlabeled data. In one of our recent work [14], we proposed the semi-supervised dimensionality reduction model which uses the pairwise constraints together with unlabeled data for dimensionality reduction. However, in that paper, we didn't discuss using both labeled and unlabeled data for dimensionality reduction.

3 Discriminant Principal Component Analysis

PCA only preserves the global covariance structure of unlabeled data which can not utilize discriminant information in labeled data. In this section, we present the DPCA algorithm which introduces a new discriminant criterion into the original objective function of PCA.

3.1 The DPCA Algorithm

Given a set of n D-dimensional data samples $X = \{x_1, x_2, ..., x_n\}$, suppose that there exist l labeled data $L = \left\{x_{i_1}, x_{i_2}, ..., x_{i_l}\right\} \subseteq X$, $i_r|_{r=1}^{l} \in \{1, 2, ..., n\}$, with the corresponding labels $y_{i_r} \in \{1, 2, ..., c\}$, our task is to find a set of projective vectors $W = [w_1, w_2, ..., w_d]$, such that the transformed low-dimensional representations $z_i = W^T x_i$, not only can preserve the structure of $\boldsymbol{X}$ but also can reflect the discriminant information in L.

The objective function of PCA is defined as maximizing

$$J_{\text{PCA}} = \frac{1}{n} \sum_{i=1}^{n} \left(w^T x_i - w^T m\right)^2 = w^T S_T w \tag{1}$$

where $m = \frac{1}{n} \sum_{i=1}^{n} x_i$, $S_T = \frac{1}{n} \sum_{i=1}^{n} (x_i - m)(x_i - m)^T$ is the covariance matrix and also called as the normalized total scatter matrix. For the convenience of discussion, one-dimensional case is considered here but it is not difficult to extend to high-dimensions.

From Eq. 1, PCA does not use the discriminant information in labeled data set $\boldsymbol{L}$ at all. To make PCA have the discriminant power, without losing its data representation character, we propose the following objective function

$$J_{\text{DPCA}} = J_{\text{D}} + \lambda J_{\text{PCA}} \tag{2}$$

Here, J_{PCA} which is defined in Eq. 1, is the criterion of PCA, and J_{D} denotes some discriminant criterion on labeled data set $\boldsymbol{L}$. In Eq. 2, λ is a regularized coefficient balancing the contributions of two terms. In this paper, we adopt the following criterion as maximizing J_{D}

$$J_{\mathrm{D}} = w^T \left(S_B^L - \eta S_W^L \right) w \tag{3}$$

Where $\mathbf{S}_B^L$ and $\mathbf{S}_W^L$ are respectively defined in the following Eqs. 4 and 5, and η is a regularized coefficient balancing the contributions of two terms.

$$\mathbf{S}_B^L = \frac{1}{|\Omega_B|} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \Omega_B} \left(\boldsymbol{x}_i - \boldsymbol{x}_j \right) \left(\boldsymbol{x}_i - \boldsymbol{x}_j \right)^T \tag{4}$$

$$\mathbf{S}_W^L = \frac{1}{|\Omega_W|} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \Omega_W} \left(\boldsymbol{x}_i - \boldsymbol{x}_j \right) \left(\boldsymbol{x}_i - \boldsymbol{x}_j \right)^T \tag{5}$$

Where $|A|$ denotes the cardinality of a set A, and Ω_B and Ω_W is respectively defined by Eqs. 6 and 7 as follows

$$\Omega_B = \left\{ \left(x_i, x_j \right) \middle| \, x_i, x_j \in L \quad \text{and} \quad y_i \neq y_j \right\} \tag{6}$$

$$\Omega_W = \left\{ \left(x_i, x_j \right) \middle| \, x_i, x_j \in L \quad \text{and} \quad y_i = y_j \right\} \tag{7}$$

We call $\mathbf{S}_B^L$ and $\mathbf{S}_W^L$as generalized between-class scatter matrix and generalized within-class scatter matrix respectively. The intuition between Eq. 3 is to let the average distance in the transformed low-dimensional space between data examples in different classes as large as possible, while distance between data examples with the same class as small as possible.

Substituting Eqs. 1 and 3 into Eq. 2, we obtain the objective function of DPCA as maximizing J_{DPCA} w.r.t.$w^T w = 1$, where

$$J'_{\mathrm{DPCA}} = w^T \left(S_B^L - \eta S_W^L + \lambda S_T \right) w \tag{8}$$

Clearly, Eq. 8 is a typical eigen-problem, which has a closed form solution by computing the eigen vectors of $S_B^L - \eta S_W^L + \lambda S_T$ corresponding to the largest eigen values. The whole procedure of the proposed DPCA algorithm is summarized in Algorithm 1 as below.

Algorithm 1: DPCA

Input: Data set $X = [x_1, x_2, ..., x_n]$, labeled data set $L = \left[x_{i_1}, x_{i_2}, ..., x_{i_l} \right] \subseteq X$ and corresponding class labels $y_{i_r} \in \{1, 2, ..., c\}$, $i_r|_{r=1}^{l} \in \{1, 2, ..., n\}$; parameters η,λ,d.

Output: Projective matrix $W = [w_1, w_2, ..., w_d]$.

Step 1: Construct the sets Ω_B and Ω_W from labeled data set $\boldsymbol{L}$ according to Eqs. 6 and 7 respectively.

Step 2: Compute $\mathbf{S}_B^L$ and $\mathbf{S}_W^L$ using Eqs. 4 and 5 respectively.

Step 3: Compute $S_T = \frac{1}{n} \sum_{i=1}^{n} (x_i - m) (x_i - m)^T$, $m = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Step 4: Compute the d eigenvectors W of $\mathbf{S}_B^L - \eta \mathbf{S}_W^L + \lambda \mathbf{S}_T$ corresponding to the largest d eigenvalues.

3.2 DPCA with Pairwise Constraints

In general, domain knowledge can be expressed in diverse forms, such as class labels, pairwise constraints or other prior information [14]. Pairwise constraints arise naturally in many tasks such as image retrieval. In those applications, considering the pairwise constraints is more practical than trying to obtain class labels, because the true labels may not be known a priori, while it could be easier for a user to specify whether some pairs of instances belong

to the same class or not. Moreover, the pairwise constraints can be derived from labeled data but not vice versa. Furthermore, unlike class labels, the pairwise constraints can sometimes be automatically obtained without human intervention [6]. Fortunately, our DPCA algorithm can easily utilize both pairwise constraints and labeled data.

Given some supervision information in the form of must-link constraint set $M = \{(x_i, x_j) | x_i$ and x_j belongs to the same class$\}$ and cannot-link constraint set $C = \{(x_i, x_j) | x_i$ and x_j belongs to the different classes$\}$, we can define the new generalized between-class scatter matrix $\mathbf{S}_B^{L'}$ and generalized with-class scatter matrix $\mathbf{S}_W^{L'}$ using both pairwise constraints sets M, C and the labeled data set L as follows

$$\mathbf{S}_B^{L'} = \frac{1}{|\Omega_B \cup C|} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \Omega_B \cup C} \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right) \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^T \tag{9}$$

$$\mathbf{S}_W^{L'} = \frac{1}{|\Omega_W \cup M|} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \Omega_W \cup M} \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right) \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^T \tag{10}$$

Then we can obtain the new objective function of DPCA as maximizing J'_{DPCA} w.r.t.$w^T w = 1$, where

$$J'_{\mathrm{DPCA}} = w^T \left(S_B^{L'} - \eta S_W^{L'} + \lambda S_T\right) w. \tag{11}$$

4 Experiments

In this section, we evaluate the performance of our proposed DPCA algorithm on several UCI data sets [15] including *Dermatology*, *Horse*, *Iris*, *Lymph*, *Sonar*, *Soybean*, *Vowel* and *Wine*, and on one face database: YaleB [16]. Table 1 gives the statistics of the 8 UCI data sets. For each UCI data set, we choose the first half of samples from each class as the training data, and the remaining for testing data. Then we randomly select a few data samples from the training data as the labeled data. The process is repeated for 100 runs and the averaged results are recorded.

The performances of all algorithms are measured by the classification accuracy on testing data. In all experiments, the nearest neighborhood (1-NN) classifier is employed for classification, after dimensionality reduction with the above algorithms. For DPCA, we choose the values for parameters η and λ from the set $\{0.1, 1, 10\}$. More specifically, For *Horse*, *Iris*,

Table 1 Statistics of the UCI data sets

Data sets	Size	Dimension	# Of classes
Dermatology	366	33	6
Horse	368	27	2
Iris	150	4	3
Lymph	148	18	4
Sonar	208	60	2
Soybean	47	35	4
Vowel	528	10	11
Wine	178	13	3

Sonar, and *Vowel,* we set $\eta = 1$, while for *Dermatology*, *Lymph*, *Soybean* and *Wine*, we set $\eta = 10$. For all 8 UCI data sets, we set $\lambda = 1$. For YaleB, we set $\eta = 10$ and $\lambda = 0.1$.

4.1 Results on UCI Datasets

Figure 1 shows the plots for accuracy under desired number of reduced dimensions versus different numbers of labeled data on 8 UCI data sets. Here we also plot the error bars representing standard deviations for PCA-p, LDA-p, SDA and DPCA. The desired number of reduced dimensions is the number of the classes of labeled data. For each data set, we compare DPCA with PCA and LDA. PCA performs on the full training data without using labels, while LDA performs on the full training data using all labels. In other words, PCA is fully unsupervised while LDA is fully supervised. For comparison, we also implement two variants on PCA and LDA, i.e. PCA-p and LDA-p, which denote PCA and LDA performing on the partial labeled data set respectively. In addition, we also compare our DPCA with SDA which is proposed in [12].

Figure 1 indicates that, in most cases, DPCA achieves the best performances among the 6 algorithms except LDA. It can be also seen from Fig. 1 that, on most data sets, as the number of labeled data increases, the accuracies of LDA-p, SDA and DPCA also increase. It verifies that labeled data are very useful for discrimination. Moreover, the error bars show that DPCA almost has smallest standard deviations when varying the number of labeled data. It indicates that DPCA is more stable than other algorithms. On the other hand, a closer study on Fig. 1 reveals that, generally, the accuracy of DPCA increases fast in the beginning (with a few labeled data) and slows down at the end (with relatively more labeled data). It implies that too many labeled data won't help too much to further boost the accuracy, and only a few labeled data are sufficient in DPCA. In contrast, LDA-p or SDA typically requires relatively more labeled data to obtain a satisfying accuracy as shown in Fig. 1. We conjecture that the reason may be DPCA uses the unlabeled data, which makes it more stable. The averaged accuracy under different numbers of labeled data on 8 UCI data sets is summarized in Table 2. It is impressive to see that DPCA is nearly always superior to SDA on eight datasets.

Furthermore, we discuss the DPCA after introducing supervision in the form of pairwise constraints. Figure 2 shows the plots for accuracy versus different numbers of labeled data and different levels of constraints on *Lymph* and *Wine* data sets. Table 3 summarizes the averaged accuracy under different numbers of labeled data with different numbers of constraints. There are 3 levels of constraints, i.e. 10 constraints, 30 constraints and 50 constraints. From Fig. 2 and Table 3 we notice the same tendency on two datasets. That is, increasing the number of constraints improves the accuracy, which is more apparent when fewer labeled data are used.

4.2 Results on YaleB Face Database

Finally in this subsection, we use DPCA for 2D visualization. We choose the first 5 subjects from YaleB face database and get totally 320 face samples. Figure 3 shows an illustration for one person. As shown from the figure, the variability between images of the same person is mainly due to different lighting conditions. These factors make the variability among images belonging to the same subject greater than the variability among images of different subjects. Figure 4 gives the 2D visualization results when different numbers of labeled data are used. Note that when 0 labeled data is used, DPCA is equivalent to original PCA. It can be seen from Fig. 4 that, PCA cannot correctly indicate the intrinsic structure of the dataset in 2-dimensional space. However, with a few labeled data, DPCA can find the intrinsic structure of face images and correctly represent them in 2-dimensional space.

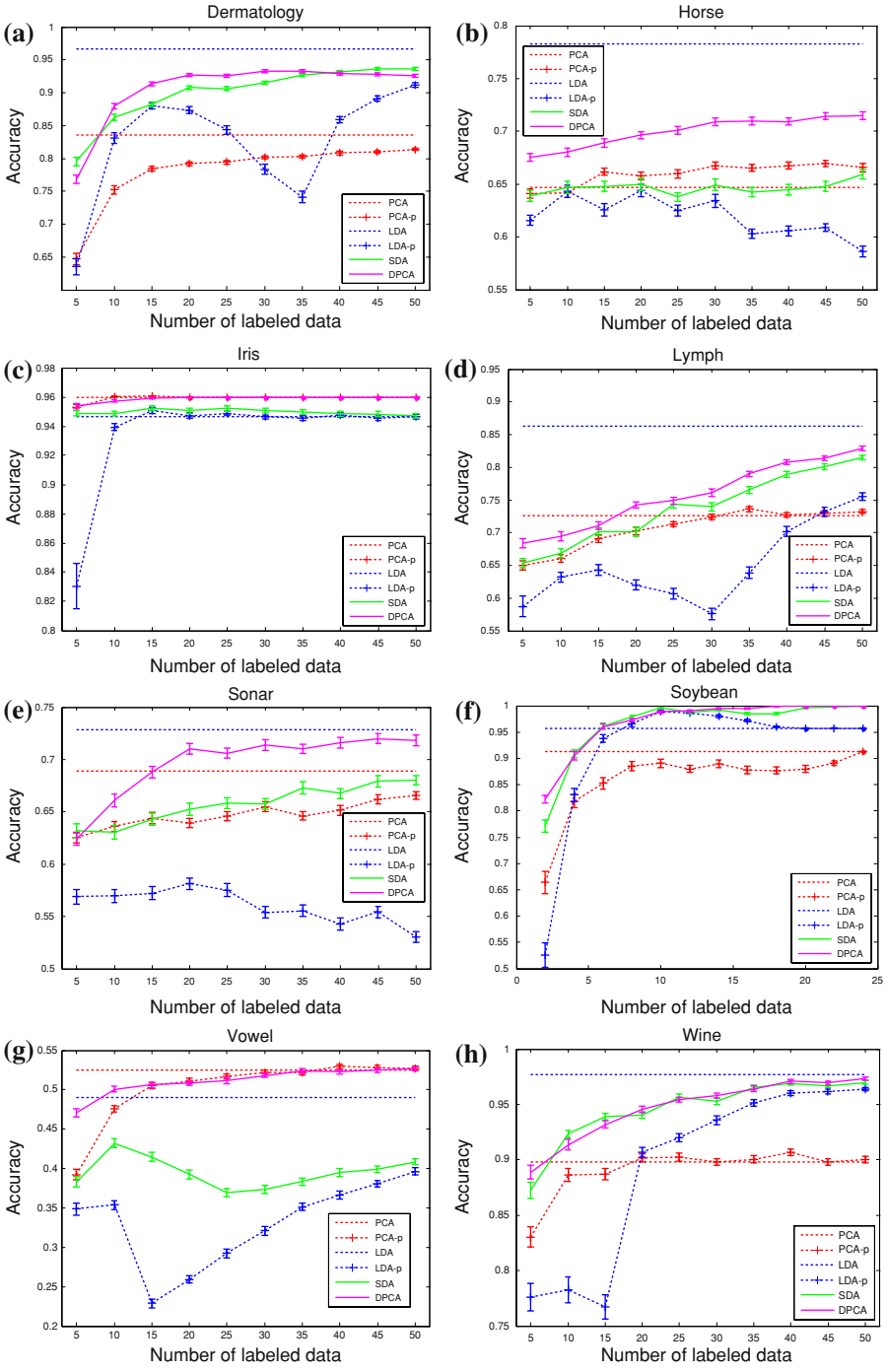

Fig. 1 Accuracy versus different numbers of labeled data on 8 UCI data sets: **a** on *Dermatology*, **b** on *Horse*, **c** on *Iris*, **d** on *Lymph*, **e** on *Sonar*, **f** on *Soybean*, **g** on *Vowel*, **h** on *Wine*

Table 2 Averaged accuracy (%) of different algorithms on UCI data set (the values behind the symbol '±' denote the standard deviation)

Data sets	PCA	LDA	PCA-p	LDA-p	SDA	DPCA
Dermatology	83.5	96.7	78.1 ± 5.2	82.5 ± 8.3	89.9 ± 4.4	90.5 ± 5.1
Horse	64.7	78.3	66.3 ± 1.1	61.9 ± 1.7	64.7 ± 0.7	70.1 ± 1.6
Iris	96.0	94.7	95.9 ± 0.1	93.6 ± 3.3	95.0 ± 0.3	96.0 ± 0.2
Lymph	72.6	86.3	70.9 ± 3.1	64.9 ± 6.5	73.8 ± 5.8	75.8 ± 5.1
Sonar	68.9	72.8	64.5 ± 1.3	56.0 ± 1.6	65.3 ± 2.0	69.7 ± 3.0
Soybean	91.3	95.7	85.9 ± 6.7	91.9 ± 12.6	96.4 ± 6.0	97.2 ± 5.0
Vowel	52.5	49.1	50.1 ± 4.4	33.2 ± 5.1	39.6 ± 1.7	51.2 ± 2.1
Wine	89.8	97.7	88.8 ± 2.4	89.4 ± 8.4	94.5 ± 2.8	94.7 ± 2.9

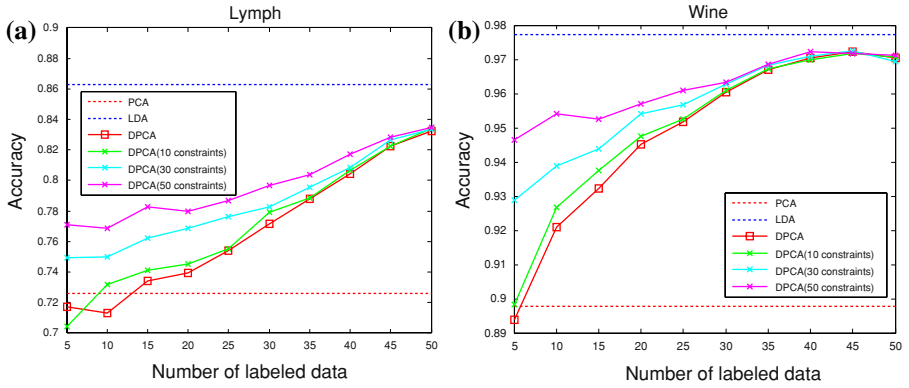

Fig. 2 Accuracy versus different numbers of labeled data and different levels of constraints on 2 UCI data sets: **a** on *Lymph*, **b** on *Wine*

Table 3 Averaged accuracy (%) with different levels of constraints on UCI data set

Data sets	PCA	LDA	DPCA	DPCA(10)	DPCA(30)	DPCA(50)
Lymph	72.6	86.3	76.8 ± 4.3	77.1 ± 4.2	78.5 ± 3.0	79.7 ± 2.4
Wine	90.0	97.7	94.5 ± 2.9	94.9 ± 2.3	95.6 ± 1.4	96.0 ± 0.1

Fig. 3 Illustration of a subject in YaleB face database

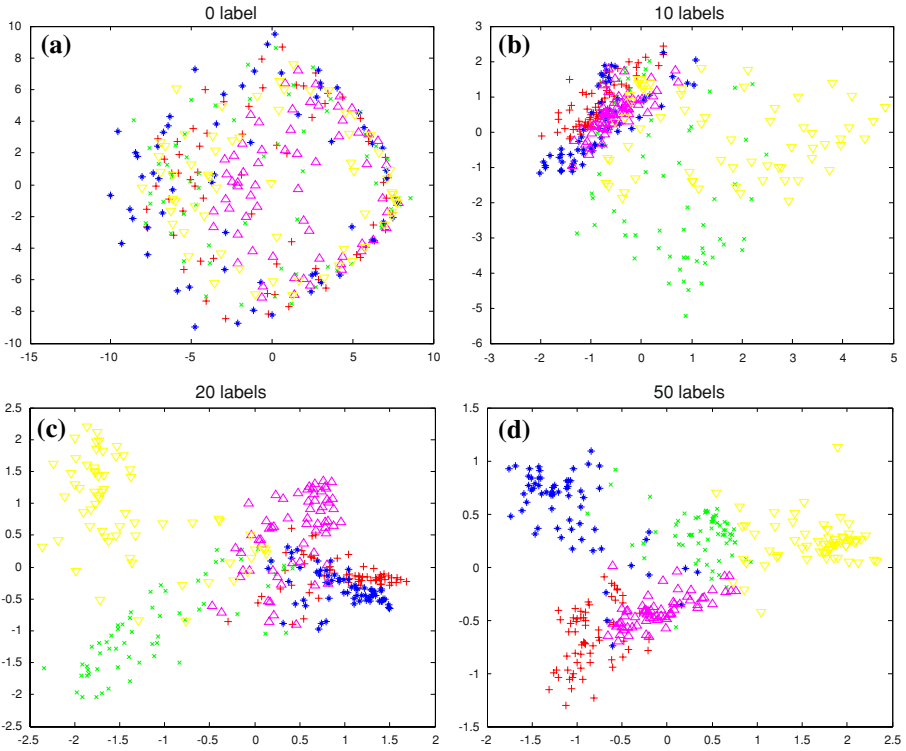

Fig. 4 Two dimensional visualization results of DPCA with different labeled data: **a** 0 label (equivalent to PCA), **b** 10 labels, **c** 20 labels, **d** 50 labels. *Each symbol* (color) represents a different subjects

5 Conclusion

In this paper, we address the problem of dimensionality reduction when a few labeled data, besides the large amount of unlabeled data, are available. We introduce a new discriminant criterion into the original objective function of PCA, and propose the discriminant PCA (DPCA) algorithm to boost its discriminant power. DPCA can effectively use the unlabeled data as well as the discriminant information in the labeled data for dimensionality reduction. Moreover, DPCA can easily utilize the supervision information in the form of pairwise constraints. The derived DPCA algorithm is efficient and has a closed form solution. Experimental results on several UCI and face data sets show that DPCA is superior to several established dimensionality reduction methods.

Acknowledgements This work is partially supported by National Science Foundation of China under Grant No. 60875030.

References

1. Duda R, Hart P, Stork D (2001) Pattern classification. 2nd edn. Wiley, New York
2. Jolliffe I (1986) Principal component analysis. Springer, New York

3. Joachims T (1999) Transductive inference for text classification using support vector machines. In: Proceedings of the International Conference on Machine Learning, Bled, Slovenia, pp 200–209, June
4. Zhou Z, Li M (2005) Semi-supervised regression with co-training. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, pp 908–913
5. Basu S (2005) Semi-supervised clustering: probabilistic models, algorithms and experiments. PhD thesis, Department of Computer Sciences, University of Texas at Austin
6. Bar-Hillel A, Hertz T, Shental N, Weinshall D (2005) Learning a mahalanobis metric from equivalence constraints. J Mach Learn Res 6:937–965
7. Zhu X (2006) Semi-supervised learning literature survey. Tech. Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI. http://www.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf
8. Yu S, Yu K, Tresp V, Kriegel HP, Wu M (2006) Supervised probabilistic principal component analysis. Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining, Philadelphia, pp 464–473
9. Lu Y, Tian Q, Sanchez M, Wang Y (2005) Hybrid PCA and LDA analysis of microarray gene expression data, computational intelligence in bioinformatics and computational biology, 2005. CIBCB '05. Proceedings of the 2005 IEEE Symposium on 14–15, California, pp. 1–6, November
10. Cai D, He X, Han J (2008) SRDA: an efficient algorithm for large scale discriminant analysis. IEEE Trans Knowl Data Eng 20(1):1–12
11. Cai D, He X, Han J (2007) Spectral regression for efficient regularized subspace learning. IEEE International Conference on Computer Vision (ICCV), pp 1–8, Rio de Janeiro, Brazil, October
12. Cai D, He X, Han J (2007) Semi-Supervised discriminant analysis. Proceeding of the IEEE Int'l Conference on Computer Vision Rio de Janeiro, pp 1–7
13. Zhang Y, Yeung D (2008) Semi-supervised discriminant analysis via CCCP. Proceedings of the European conference on Machine Learning, Antwerp, Belgium, pp 644–659
14. Zhang D, Zhou Z, Chen S (2007) Semi-supervised dimensionality reduction. In: Proceedings of the 7th SIAM International Conference on Data Mining, Minneapolis, Minnesota, pp 1–6
15. Blake C, Keogh E, Merz CJ (1998) UCI repository of machine learning databases. [http://www.ics.uci.edu/~mlearn/MLRepository.html], Department of Information and Computer Science, University of California, Irvine
16. Georghiades A, Belhumeur P, Kriegman D (2001) From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell 23:643–660