Pattern Recognition 45 (2012) 1205–1210

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Graph optimization for dimensionality reduction with sparsity constraints

Limei Zhang [a,b], Songcan Chen [a,*], Lishan Qiao [b]

[a] Department of Computer Science and Engineering, Nanjing University of Aeronautics & Astronautics, 210016 Nanjing, PR China
[b] Department of Mathematics Science, Liaocheng University, 252000 Liaocheng, PR China

ARTICLE INFO

Article history:
Received 1 February 2011
Received in revised form
30 July 2011
Accepted 13 August 2011
Available online 23 August 2011

Keywords:
Dimensionality reduction
Graph construction
Sparse representation
Face recognition

ABSTRACT

Graph-based dimensionality reduction (DR) methods play an increasingly important role in many machine learning and pattern recognition applications. In this paper, we propose a novel graph-based learning scheme to conduct **G**raph **O**ptimization for **D**imensionality **R**eduction with **S**parsity **C**onstraints (GODRSC). Different from most of graph-based DR methods where graphs are generally constructed in advance, GODRSC aims to simultaneously seek a graph and a projection matrix preserving such a graph in one unified framework, resulting in an automatically updated graph. Moreover, by applying an l_1 regularizer, a sparse graph is achieved, which models the "locality" structure of data and contains natural discriminating information. Finally, extensive experiments on several publicly available UCI and face databases verify the feasibility and effectiveness of the proposed method.

1. Introduction

It is well known that dimensionality reduction (DR) has generally been used as a principled way to understand the high-dimensional data such as image, text and video sequence. Recently, graph-based DR methods become more and more popular in pattern recognition and machine learning fields, due to the fact that graph is a powerful tool to catch the structure information hidden in objects or data. In fact, recent research [1] claimed that most existing DR methods can fall into a graph embedding framework. The representatives have ISOMAP [2], LLE [3], Laplacian eigenmap [4] and locality preserving projections (LPP) [5], just to name a few. Under such a framework, one first constructs a graph from data in terms of some prior knowledge available, and then based on the constructed graph learns a projection matrix, which transforms the original high-dimension data into a lower dimensional space. Among them, graph construction is crucial since the performances of these algorithms depend heavily on how well the graph models the original data structure.

As a consequence, the methods for graph construction have been widely studied in recent years, although building a high-quality graph is still an open problem [6]. In general, most of the graph construction processes can be decomposed into two steps. Firstly, one constructs an adjacency graph by considering the samples as nodes and linking some of them with edges according to given rules such as k-nearest neighbors, ε-ball neighborhood and b-matching [4,5,7]. Secondly, a weight is assigned for each edge. The often-used weight assignment ways include Heat Kernel [4], Inverse Euclidean Distance [8] and Local Linear Reconstruction [3], etc. All these graph construction methods are quite flexible and can in principle be used for any graph-based learning algorithms including DR, spectral clustering and semi-supervised learning [1,7,9,10]. However, as pointed out in [10], there is potential need that graph should be appropriate for the subsequent learning task.

To establish an "appropriate" graph, Zhang et al. recently presented an algorithm called Graph-optimized Locality Preserving Projections (GoLPP) [11] for DR task, which optimizes graph and projections simultaneously in one single objective function. To the best of our knowledge, this is the first attempt to perform graph optimization during a specific DR process, rather than pre-define graph before DR as done in most of graph-based algorithms [1]. Despite GoLPP obtains empirical superiority to traditional LPP on some datasets, the graph resulted from GoLPP usually loses traditional sparsity even though a sparse initial graph is given in its iterative optimization.

To address this problem, in this paper, we propose a novel strategy to conduct **G**raph **O**ptimization for **D**imensionality **R**eduction with **S**parsity **C**onstraints (GODRSC). The proposed method not only shares the advantages of GoLPP with automatically adjustable graph, but also has some additional desirable characteristics:

1) The sparsity of graph is held by replacing the entropy regularizer in GoLPP with an l_1 norm minimization. As pointed out in [7], sparsification is important to graph since it can bring higher efficiency, better accuracy and robustness to noise.
2) Interestingly, with adjustable graph, GODRSC essentially provides an extension to the sparsity preserving projections (SPP) [12], a recently developed DR algorithm based on sparse

[*] Corresponding author. Tel.: +86 25 84896481x12221; fax: +86 25 84892400.
E-mail address: s.chen@nuaa.edu.cn (S. Chen).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.08.015

representation (see Section 2.2 for more details). This establishes a natural link between GODRSC and SPP, which helps to give an intuitive explanation why and how the former might work well [12–14].

3) By solving the trace ratio problem directly, GODRSC avoids the nonuniqueness of the solutions involved in GoLPP. See Section 3 for more details.

The rest of the paper is organized as follows: Section 2 reviews three related DR algorithms, GoLPP, SPP and its orthogonalized extension. Section 3 introduces the GODRSC model and algorithm. In Section 4, some experimental results are presented. Finally, conclusions are drawn in Section 5.

2. Related works

2.1. Graph-optimized locality preserving projections (GoLPP)

Given a set of sample points $X=[x_1,x_2,\ldots,x_n]$, where $x_i \in R^D$, $i=1.2,\ldots,n$, we firstly review the objective function of LPP [5], which GoLPP is based on

$$\min_W \frac{\sum_{i,j=1}^n \|W^T x_i - W^T x_j\|^2 P_{ij}}{\sum_{i=1}^n D_{ii}\|W^T x_i\|^2}$$

where $W \in R^{D\times d}(d<D)$ is the projection matrix, $D_{ii}=\sum_{j=1}^n P_{ij}$, and $P=(P_{ij})_{n\times n}$ is the edge weight matrix of a neighbor graph, which has been specified before learning W. In contrast to LPP with such a pre-defined graph, GoLPP simultaneously completes graph optimization and projection learning within a unified objective function [11] below:

$$\min_{W,S_{ij}} \frac{\sum_{i,j=1}^n \|W^T x_i - W^T x_j\|^2 S_{ij}}{\sum_{i=1}^n \|W^T x_i\|^2} + \eta \sum_{i,j=1}^n S_{ij}\ln S_{ij}$$
$$\text{s.t.} \sum_{j=1}^n S_{ij}=1, i=1,\ldots,n$$
$$S_{ij}\ge 0, i,j=1,\ldots,n$$

which can in turn be rewritten as the following *trace ratio* form

$$\min_{W,S_{ij}} \frac{tr(W^T XLX^T W)}{tr(W^T XX^T W)} + \eta \sum_{i,j=1}^n S_{ij}\ln S_{ij}$$
$$\text{s.t.} \sum_{j=1}^n S_{ij}=1, i=1,\ldots,n$$
$$S_{ij}\ge 0, i,j=1,\ldots,n \tag{1}$$

where $S=(S_{ij})_{n\times n}$ is an unknown affinity weight matrix of graph, L is the graph Laplacian; $\sum_{i,j=1}^n S_{ij}\ln S_{ij}$ is an entropy regularization term with sum-to-one constraint $\sum_{j=1}^n S_{ij}=1$ and non-negative constraint $S_{ij}\ge 0$ for avoiding degenerate solution as well as endowing S_{ij} with probability meaning; η is a tradeoff parameter. According to Zhang et al. [11], the GoLPP model can be solved by alternating iteration and the iteration process is theoretically proved convergence. Finally its performance empirically outperforms LPP for visualization and classification tasks on a number of often-used public datasets, benefiting from the automatically optimized graph.

2.2. Sparsity preserving projections (SPP) and its orthogonalization

SPP [12] is an unsupervised DR algorithm based on graph construction by sparse representation. In particular, SPP firstly constructs a graph by representing each sample point x_i using as few sample points in $X\backslash\{x_i\}$ as possible. With different assumptions to noise, it can be cast into different l_1-minimization problems such as the following one:

$$\min_{S_i} \|S_i\|_1$$
$$\text{s.t.} \|x_i - XS_i\|^2 < \varepsilon$$
$$\sum_{j=1}^n S_{ij}=1 \tag{2}$$

where S_i is a column vector consisting of the representative coefficient of sample x_i,[1] and minimizing the l_1 norm aims to obtain a sparse solution; $\|x_i - XS_i\|^2$ is the error for reconstructing x_i. Naturally, the jth element S_{ij} in coefficient vector S_i can be used as the affinity weights between samples x_i and x_j, and thus SPP builds a graph $G=(X,(S_{ij})_{n\times n})$, which describes the sparse reconstructive relationship among the original samples.

Then, SPP seeks a projection matrix W best preserving the sparse graph above. Similar to NPE [15], a linear version of LLE [3], SPP does this by the following objective function:

$$\min_W \frac{\sum_{i=1}^n \|W^T(x_i - XS_i)\|^2}{\sum_{i=1}^n \|W^T x_i\|^2} \tag{3}$$

which is equivalent to the trace ratio problem:

$$\max_W \frac{tr(W^T XS_\beta X^T W)}{tr(W^T XX^T W)}$$

where$S_\beta = S+S^T - SS^T$. Similar to most trace ratio models [1,5,15], it can be approximately solved by generalized eigenvalue decomposition.

Furthermore, one can generalize SPP by introducing some priors or constraints to its model as in many linear DR algorithms such as LPP. In order to better discuss and validate the proposed GODRSC method later, here, we give an orthogonalized extension of SPP, and call it OSPP simply, which can be modeled just by imposing orthogonal constraint on projection matrix W. With the same notations as in Eq. (3), the model of OSPP is established by minimizing the objective defined as follows:

$$\min_W \frac{\sum_{i=1}^n \|W^T(x_i - XS_i)\|^2}{\sum_{i=1}^n \|W^T x_i\|^2}$$
$$\text{s.t. } W^T W = I$$

Similarly, we have its corresponding trace ratio form

$$\max_W \frac{tr(W^T XS_\beta X^T W)}{tr(W^T XX^T W)}$$
$$\text{s.t. } W^T W = I$$

With the orthogonal constraint, the OSPP model above can be solved *exactly* by many recently proposed algorithms [16–18], rather than *approximated* by the generalized eigenvalue problem as original SPP.

3. Graph optimization for dimensionality reduction with sparsity constraints

3.1. Motivations

Note that, in the GoLPP model (1), the maximum entropy term makes the edge weights of graph as uniform as possible, consequently incurring the loss of sparsity, which is a basic common merit in typical graph construction using k-NN and ε-ball, or l_1 regularization, etc. In fact, the resulted graph updating formulation (see Eq. (11) in Appendix) of GoLPP has shown that there exist nonzero edge weights in all the pairs of samples. On the

[1] It is worthwhile to point out that the ith entry in S_i is zero due to removing x_i from sample matrix X.

other hand, according to an often-used strategy, the trace ratio term in GoLPP (1) is transformed into a corresponding ratio trace form, which can be simply solved via generalized eigenvalue decomposition. However, such a solution is not unique and deviates from the original objective function, which further incurs the nonuniqueness of the subsequent learning results (see Appendix where we give particular illumination).

3.2. *Model for GODRSC*

To address the above problems involved in GoLPP, a new graph learning model for DR is given below:

$$\min_{W,S_i} J(W,S) = \frac{\sum_{i=1}^{n} \|W^T(x_i - XS_i)\|^2}{\sum_{i=1}^{n} \|W^T x_i\|^2} + \sum_{i=1}^{n} \lambda_i \|S_i\|_1 \quad \text{s.t. } W^T W = I \tag{4}$$

where $W \in R^{D \times d}(d < D)$ is the projection matrix, $X = [x_1, x_2, \ldots, x_n], x_i \in R^D$, is a set of given samples, $S = (S_{ij})_{n \times n}$ is the affinity weight matrix with zero diagonal entries, S_i is the ith column of S as well as the reconstructive coefficient vector of sample x_i, λ_i is a tradeoff parameter, and I is a $d \times d$ unit matrix. We call the proposed model Graph Optimization for Dimensionality Reduction with Sparsity Constraints (GODRSC) since it deals with both optimization and sparsity of graph for DR process.

As described previously, the l_1 regularizer on S_i in the proposed model (4) aims at obtaining a sparse graph. Furthermore, its numerator (i.e., $\sum_{i=1}^{n} \|W^T(x_i - XS_i)\|^2$) of the first term is different from the one (i.e., $\sum_{i,j=1}^{n} \|W^T x_i - W^T x_j\|^2 S_{ij}$) in GoLPP. This is mainly due to the following two factors: (1) Avoiding degenerate solution. The model would generate a trivial solution if replacing the entropy term with l_1 regularizer in GoLPP model directly. (2) Establishing relationship between the proposed algorithm and SPP. Shortly, we will see that with such a consideration, GODRSC potentially gives an extension to SPP. On the other hand, in model (4) we impose orthogonal constraint on the projection matrix so as to directly solve an orthogonally constrained trace ratio problem for avoiding nonunique solution as in GoLPP.

3.3. *Algorithm*

Despite its non-convexity, problem (4) can be easily solved by alternating iteration scheme.

3.3.1. *Initialization*

We simply initialize $W = I$,[2] and thus Eq. (4) reduces to

$$\min_{S_i} \frac{\sum_{i=1}^{n} \|x_i - XS_i\|^2}{\sum_{i=1}^{n} \|x_i\|^2} + \sum_{i=1}^{n} \lambda_i \|S_i\|_1$$

which can be further simplified into the following form:

$$\min_{S_i} \sum_{i=1}^{n} \|x_i - XS_i\|^2 + \sum_{i=1}^{n} \tilde{\lambda}_i \|S_i\|_1 \tag{5}$$

where $\tilde{\lambda}_i$ is the constant times of λ_i, and, without loss of generality, we continue to use λ_i in the later text. It is easy to see that problem (5) is just a sparse representation problem, which has been studied deeply and can be efficiently solved by many off-the-shelf algorithms such as l_1-magic [19,20]. In this paper, we deal with it through a recently proposed and very popular tool package, SLEP, owing to its high efficiency [21].

Then, the algorithm runs alternately between the following steps:

Step 1. With the obtained graph S by (5), problem (4) becomes an orthogonal SPP (OSPP) problem:

$$\min_{W} \frac{\sum_{i=1}^{n} \|W^T(x_i - XS_i)\|^2}{\sum_{i=1}^{n} \|W^T x_i\|^2} \quad \text{s.t. } W^T W = I$$

or equivalently,

$$\min_{W} \frac{tr[W^T X(I - S - S^T + SS^T)X^T W]}{tr(W^T XX^T W)} \quad \text{s.t. } W^T W = I \tag{6}$$

This is a trace ratio problem with orthogonality constraint whose global optimal solution can be found via several efficient iterative procedures [16–18]. For example, Guo et al. [17] and Wang et al. [16] handled it by solving a series of *trace difference* problems. Here, we choose the decomposed Newton's method (DNM) in [18] to solve optimal W in (6) since it has been empirically proved more efficient [18,22].

Step 2. Fix W and compute optimal $S = (S_{ij})_{n \times n}$ in problem (4), which can be reduced to the following optimization problem:

$$\min_{S_i} \sum_{i=1}^{n} \|W^T(x_i - XS_i)\|^2 + \sum_{i=1}^{n} \lambda_i \|S_i\|_1 \tag{7}$$

By contrast with Eq. (5), Eq. (7) is a sparse representation problem in the *transformed* data space by current W. Similarly, we can get its optimal solution by SLEP.

The GODRSC algorithm is summarized below.

```
Input: $X$— data matrix;
        $\lambda_i$, $i = 1, \ldots, n$— regularized parameter;
        $\varepsilon$— iterative stop threshold ;
Output: $W$— projection matrix.
Procedure:
Initialize $W = I$ and compute $S$ by solving Eq. (5);
Calculate the objective function value of (4): $J_0 \leftarrow J(W,S)$
For $k = 1, 2, \ldots, MaxIter$
        Calculate projection matrix $W$ by solving Eq. (6) shown in Step 1;
        Update weight matrix $S$ using the solution of Eq. (7) shown in Step 2;
        Calculate the objective function value of (4): $J_k \leftarrow J(W,S)$;
        If $|J_k - J_{k-1}| < \varepsilon$
                Break and return $W$;
        EndIf
EndFor
```

Based on the algorithm above, the alternating iterative procedure obtains optimal solution at each step for S and W, respectively. Thus, it can also easily be proven that GODRSC algorithm is convergent according to the block coordinate descent method [23].

3.4. *Comparison with related works*

- **GoLPP**: Both GoLPP and GODRSC attempt to optimize graph and learning projections simultaneously in one single objective function. However, as mentioned before, in GoLPP the graph is fully connected and the solution is not unique. On the contrary, GODRSC obtains a desired sparse graph, which models the "local" structure of data and contains natural discriminant information

[2] Of course, one may initialize W according to other priors or assumptions. Also, one can firstly initialize the graph S, but it is uneasy here due to the scarcity of prior information about such a graph.

though in unsupervised manner [12–14,24]. Furthermore, GODRSC avoids the nonuniqueness of the solution in GoLPP algorithm by addressing the orthogonally constrained trace ratio problem directly.

- **SPP**: GODRSC shares a common point with SPP. Both of them involve graph construction based on sparse representation. However, GODRSC alternately performs sparse graph construction and projection learning in a unified framework. Consequently, it considers the sparse representation of *original* data as well as *transformed* data by *different W*. Different from it, the graph construction in SPP is completed only in the *original* data space and has no direct connection to subsequent DR process. In addition, the obtained projection matrix is orthogonal in GODRSC, while nonorthogonal in SPP.

4. Experiments

In this section, we experimentally evaluate the effectiveness of the proposed GODRSC algorithm in terms of classification accuracy on five UCI and two face datasets, comparing with GoLPP, SPP and OSPP, respectively. Then, to demonstrate why the proposed algorithm might work well, we give the visualization of the resulting graph adjacency matrices via GoLPP, SPP and GODRSC on a face data set.

4.1. Dataset descriptions and experimental sets

The five UCI datasets used in the experiments include *Sonar*, *Wine*, *Letter*, *Soybean* and *Ionosphere*. Their detailed descriptions are shown in Table 1.

The two well-known face datasets used here are *AR* and *Extended Yale B*. For *AR* database, we only use a subset provided by Martinez and Kak [25], which contains 100 persons (50 men and 50 women). Each person has 14 different images taken in two sessions separated by 2 weeks, and each session contains 7 images with different illumination and expressions. The *Extended Yale B* database [26] contains 2414 front-view face images of 38 individuals. Each individual has about 64 pictures taken under various laboratory-controlled lighting conditions. For the two face datasets, we randomly select 7 and 10 images of each class from each dataset as the training data, and the remaining for test. Furthermore, all the images in two datasets are, respectively, cropped to 66×48 and 32×32, and the gray level values are all rescaled to [0.1].

In each experiment, we first learn the subspaces using GoLPP, SPP, OSPP and GODRSC, respectively, based on the training data. To avoid the small sample size problem, these algorithms on *AR* and *Extended Yale B* involve a PCA process preserving 98% and 99% energy of data, respectively. Then the test data are projected onto the learned subspaces by the algorithms above, and the nearest neighbor (1-NN) classifier is performed for classification. All the experiments are performed 20 times based on different random training/test splits.

For impartial comparison, in SPP and OSPP we use the sparse representation (5) in GODRSC to construct their graphs, and empirically set its tradeoff operator $\lambda_i=0.001, i=1,\ldots,n$ in the three algorithms. Moreover, in GODRSC the iteration stop threshold ε is set to 10^{-3}, and the parameters in GoLPP are inferred as [11].

Table 1
The five UCI datasets and their corresponding partitions.

Datasets	Dimensions	Class numbers	Simple sizes	For training	For test
Sonar	60	2	208	100	108
Wine	13	3	178	90	88
Letter (a–m)	16	13	3900	260	3640
Soybean	35	4	47	20	27
Ionosphere	34	2	351	200	151

4.2. Experimental results

The classification results of GoLPP, SPP, OSPP and GODRSC on the five UCI and two face datasets across 20 random training/test splits are shown in Table 2.

From the experimental results, we can obtain several observations:

(1) GODRSC performs better than GoLPP on almost all the datasets, which indicates that the sparsity of graph via l_1 regularization is important to subsequent classification due to its natural discriminating power. This point caters for our motivation and is also consistent with the results based on the sparse graph construction [12,14,24].
(2) On most of the used databases, GODRSC outperforms SPP and OSPP, which illustrates GODRSC may benefit from the automatically optimized graph rather than the pre-constructed one as in SPP and OSPP, since they are all based on the sparse graph construction.
(3) In contrast with SPP, OSPP obtains slightly better results with higher accuracies on four datasets, and lower on the other three datasets. This states that orthogonal projection constraint may benefit to the subsequent classification task, due to its measure invariance [27] and the optimal solution of trace ratio problem.

Furthermore, we take *Extended Yale B* as an example to show the resulting graphs in GoLPP, SPP and GODRSC, respectively. This is expected to intuitively illustrate why the proposed algorithm generally works well. The visualization results of three edge weight matrices obtained by GoLPP, SPP and GODRSC are all demonstrated in Fig. 1, where only a sub-block of each matrix corresponding to the first 30 face images from the first three

Table 2
The best average classification accuracies, corresponding standard deviations and dimensions of GoLPP, SPP, OSPP, GODRSC on the used datasets across 20 splits.

Datasets	Methods	Accuracies (%)	Std (%)	Dim.
Sonar	GoLPP	72.78	5.99	59
	SPP	77.59	2.88	57
	OSPP	83.89	4.01	57
	GODRSC	**84.87**	4.65	53
Wine	GoLPP	87.92	2.53	6
	SPP	89.06	5.24	11
	OSPP	87.92	4.34	7
	GODRSC	**91.02**	3.44	4
Letter	GoLPP	76.08	0.81	14
	SPP	78.26	1.33	16
	OSPP	77.25	1.78	12
	GODRSC	**79.98**	2.37	12
Soybean	GoLPP	**83.19**	3.4	24
	SPP	82.04	3.92	35
	OSPP	82.89	2.80	35
	GODRSC	82.22	2.13	23
Ionosphere	GoLPP	92.85	1.27	21
	SPP	91.20	1.35	33
	OSPP	94.04	2.86	34
	GODRSC	**94.94**	1.34	30
AR	GoLPP	85.24	1.89	234
	SPP	89.89	1.8	233
	OSPP	88.93	2.0	234
	GODRSC	**91.10**	1.16	227
Yale B	GoLPP	80.11	2.92	167
	SPP	81.82	3.99	132
	OSPP	84.24	4.11	159
	GODRSC	**85.16**	1.02	134

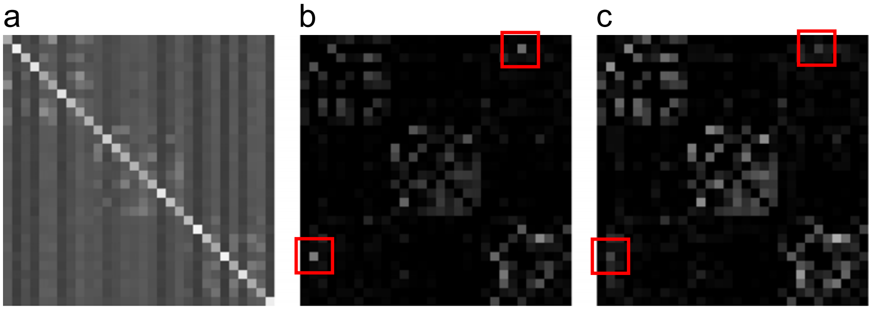

Fig. 1. Visualization of a segment of the resulting graph adjacency matrix in (a) GoLPP, (b) SPP and (c) GODRSC on Extended Yale B, where images from the same individual are arranged together.

persons is shown due to the limitation of page format here. From this figure, we can get the following observations: (1) the graph of GoLPP is dense, although the intra-class edge weights are generally larger (high gray level) than the inter-class ones. In contrast, the graphs in SPP and GODRSC are both sparse and nonzero weights mostly in the same class samples. (2) For the graphs in SPP and GODRSC, despite sharing common sparsity, the GODRSC generally makes the intra-class connections of graph stronger (higher gray level), while the inter-class ones weaker (lower gray level, see the reference points indicated by the red square), which thanks the graph update in GODRSC.

5. Conclusions

In most traditional graph-based DR algorithms, graph construction is independent of DR task. Different from them, in this paper, based on a task-dependent graph construction strategy we propose a new algorithm called Graph Optimization for Dimensionality Reduction with Sparse Constraints (GODRSC). In the proposed algorithm, graph optimization with sparse constraint and projection pursuing with orthogonal constraint are simultaneously realized in one unified framework, resulting in an automatically optimized and sparsity-holding graph. Furthermore, the effectiveness of GODRSC is verified by comparing with several related algorithms including GoLPP, SPP and OSPP on some publicly available datasets, which further illustrates well-constructed graph is important to dimensionality reduction for improving the generalization of subsequent classifier. In the future, we will attempt to provide theoretical analysis for this task-dependent graph construction mode.

Acknowledgment

This work was partly supported by National Natural Science Foundations of China (60973097, 61035003) and Natural Science Foundation of Shandong Province (ZR2010FL011).

Appendix. The solution of GoLPP is not unique

In this section, we will review the solving process of the GoLPP model (1) (see [11] for more details), from which we can illuminate why its projection matrix and weight matrix are not unique.

As described in Section 2.1, the original model of GoLPP is given as follows:

$$\min_{W,S_{ij}} \frac{tr(W^T XLX^T W)}{tr(W^T XX^T W)} + \eta \sum\nolimits_{i,j=1}^{n} S_{ij} \ln S_{ij}$$

$$\text{s.t.} \sum\nolimits_{j=1}^{n} S_{ij} = 1, i = 1,\ldots,n$$

$$S_{ij} \geq 0, i,j = 1,\ldots,n \tag{1}$$

However, for obtaining a closed-form solution the trace ratio term in (1) is transformed into ratio trace form, and we have

$$\min_{W,S_{ij}} J(W,S) = tr[(W^T XX^T W)^{-1} W^T XLX^T W] + \eta \sum\nolimits_{i,j=1}^{n} S_{ij} \ln S_{ij}$$

$$\text{s.t.} \sum\nolimits_{j=1}^{n} S_{ij} = 1, i = 1,\ldots,n$$

$$S_{ij} \geq 0, i,j = 1,\ldots,n \tag{8}$$

which can be solved by the following alternating iteration steps. One is to seek W fixing S, which can be completed by the generalized eigenvalue decomposition; the other is to solve S given W, which may reduce to the following optimization problem:

$$\min_{S_{ij}} J(S) = tr[(W^T XX^T W)^{-1} W^T XLX^T W] + \eta \sum\nolimits_{i,j=1}^{n} S_{ij} \ln S_{ij}$$

$$\text{s.t.} \sum\nolimits_{j=1}^{n} S_{ij} = 1, i = 1,\ldots,n$$

$$S_{ij} \geq 0, i,j = 1,\ldots,n \tag{9}$$

Let $U = (W^T XX^T W)^{-1}$, by SVD we obtain

$$U = V\Lambda V^T = V\Lambda^{1/2}\Lambda^{1/2}V^T \tag{10}$$

since U is positive definite. From (10) and (9), we can get

$$\begin{aligned} J(S) &= tr[UW^T XLX^T W] + \eta \sum\nolimits_{i,j=1}^{n} S_{ij} \ln S_{ij} \\ &= tr[V\Lambda^{1/2}\Lambda^{1/2}V^T W^T XLX^T W] + \eta \sum\nolimits_{i,j=1}^{n} S_{ij} \ln S_{ij} \\ &= tr[\Lambda^{1/2}V^T W^T XLX^T WV\Lambda^{1/2}] + \eta \sum\nolimits_{i,j=1}^{n} S_{ij} \ln S_{ij} \\ &= tr[\tilde{W}^T XLX^T \tilde{W}] + \eta \sum\nolimits_{i,j=1}^{n} S_{ij} \ln S_{ij} \\ &= \sum\nolimits_{i,j=1}^{n} \|\tilde{W}^T x_i - \tilde{W}^T x_j\|^2 S_{ij} + \eta \sum\nolimits_{i,j=1}^{n} S_{ij} \ln S_{ij} \end{aligned}$$

where $\tilde{W} = WV\Lambda^{1/2}$. Then embedding it into problem (9), we have

$$S_{ij} = \frac{\exp(-\|\tilde{W}^T x_i - \tilde{W}^T x_j\|^2/\eta)}{\sum\nolimits_{j=1}^{n} \exp(-\|\tilde{W}^T x_i - \tilde{W}^T x_j\|^2/\eta)} \tag{11}$$

by the Lagrangian multiplier method.

During the above solving process, we focus on decomposition (10) and find that the order of singular values in Λ can be alterable. That is, as long as the order of these singular values and their singular vectors in V is together exchanged, the matrix U is invariant no matter how to be exchanged. However, once the order of these singular values makes some changes, the matrix $\tilde{W} = WV\Lambda^{1/2}$ would accordingly alter, which leads to the weights (11) and the next projection matrix also alters correspondingly. That is, the weight matrix S and the projection matrix W are not unique. Naturally, the last results such as classification would be not stable.

A possible reason of presence of such nonuniqueness of the solutions is that models (8) and (9) are ratio trace form, which is addressed in our proposed model in Section 3.2.

References

[1] S.C. Yan, D. Xu, B.Y. Zhang, H.J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1) (2007) 40–51.

[2] J.B. Tenenbaum, V. Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.

[3] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.

[4] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Computation 15 (6) (2003) 1373–1396.

[5] X.F. He, P. Niyogi, Locality preserving projections, in: Neural Information Processing Systems (NIPS), 2003.

[6] W. Liu, S.-F. Chang, Robust multi-class transductive learning with graphs, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[7] T. Jebara, J. Wang, S. Chang, Graph construction and b-matching for semi-supervised learning, in: International Conference on Machine Learning (ICML), 2009.

[8] C. Cortes, M. Mohri, On transductive regression, in: Neural Information Processing Systems (NIPS), 2007.

[9] M. Maier, U. Luxburg, Influence of graph construction on graph-based clustering measures, in: Neural Information Processing Systems (NIPS), 2008.

[10] X. Zhu, Semi-supervised Learning Literature Survey, Technical Report, 2008.

[11] L. Zhang, L. Qiao, S. Chen, Graph-optimized locality preserving projections, Pattern Recognition 43 (6) (2010) 1993–2002.

[12] L.S. Qiao, S.C. Chen, X.Y. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognition 43 (1) (2010) 331–341.

[13] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2) (2009) 210–227.

[14] S. Yan, H. Wang, Semi-supervised learning by sparse representation, in: SIAM International Conference on Data Mining (SDM), 2009.

[15] X.F. He, D. Cai, S.C. Yan, H.J. Zhang, Neighborhood preserving embedding, in: IEEE International Conference on Computer Vision (ICCV), 2005.

[16] H. Wang, S.C. Yan, D. Xu, X.O. Tang, T. Huang, Trace ratio vs. ratio trace for dimensionality reduction, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[17] Y. Guo, S. Li, J. Yang, T. Shu, L. Wu, A generalized Foley–Sammon transform based on generalized fisher discrimimant criterion and its application to face recognition, Pattern Recognition Letters 24 (1–3) (2003) 147–158.

[18] Y. Jia, F. Nie, C. Zhang, Trace ratio problem revisited, IEEE Transactions on Neural Networks 20 (4) (2009) 729–735.

[19] T. Hesterberg, N.H. Choi, L. Meier, C. Fraley, Least angle and l1 penalized regression: a review, Statistics Surveys 2 (2008) 61–93.

[20] J. Liu, J. Ye, R. Jin, Sparse learning with euclidean projection onto l1 ball, Journal of Machine Learning Research (2009).

[21] J. Liu, S. Ji, J. Ye, SLEP: sparse learning with efficient projections, 2009.

[22] F. Nie, S. Xiang, Y. Jia, C. Zhang, Semi-supervised orthogonal discriminant analysis via label propagation, Pattern Recognition 42 (2009) 2615–2627.

[23] P. Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization, Journal of Optimization Theory and Applications 109 (3) (2001) 475–494.

[24] L.S. Qiao, S.C. Chen, X.Y. Tan, Sparsity preserving discriminant analysis for single training image face recognition, Pattern Recognition Letters 31 (5) (2010) 422–429.

[25] A.M. Martinez, A.C. Kak, PCA versus LDA, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2) (2001) 228–233.

[26] K.C. Lee, J. Ho, D.J. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (5) (2005) 684–698.

[27] D. Cai, X.F. He, J.W. Han, H.J. Zhang, Orthogonal laplacianfaces for face recognition, IEEE Transactions on Image Processing 15 (11) (2006) 3608–3614.

Limei Zhang received his B.S. and M.S. degree in mathematics from Liaocheng University in 2001 and 2007, respectively. Currently she is a Ph.D. Student at the Department of Computer Science & Engineering, Nanjing university of Aeronautics & Astronautics (NUAA). Her research interests focus on Pattern Recognition and Machine Learning.

Songcan Chen received the B.Sc. degree in Mathematics from Hangzhou University (now merged into Zhejiang University) in 1983. In December 1985, he completed the M.Sc. degree in Computer Applications at Shanghai Jiaotong University and then worked at Nanjing university of Aeronautics & Astronautics (NUAA) in January 1986 as an Assistant Lecturer. There he received a Ph.D. degree in Communication and Information Systems in 1997. Since 1998, as a full Professor, he has been with the Department of Computer Science and Engineering at NUAA. His research interests include pattern recognition, machine learning and neural computing. In these fields, he has authored or coauthored over 130 scientific journal papers.

Lishan Qiao received the B.Sc. degree in mathematics from Liaocheng University (LU) in 2001. He completed the M.Sc. degree in applied mathematics from Chengdu University of Technology in 2004, and then worked at LU as an assistant lecturer. In 2010, he received a Ph.D. degree in computer applications from Nanjing university of Aeronautics & Astronautics (NUAA). Currently he is an associate professor in the Department of Mathematics Science, LU. His research interests focus on Image Processing, Pattern Recognition and Machine Learning.