ORTHOGONAL SELF-GUIDED SIMILARITY PRESERVING PROJECTIONS

Xiaozhao Fang[1], *Yong Xu*[1*], *Zheng Zhang*[1], *Zhihui Lai*[2], *Linlin Shen*[2]

1 Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology
2 College of Computer Science and Software Engineering, Shenzhen University
{xzhfang168@126.com; yongxu@ymail.com; darrenzz219@gmail.com;
lai_zhi_hui@163.com; llshen@szu.edu.cn}

ABSTRACT

In this paper, we propose a novel unsupervised dimensionality reduction (DR) method called orthogonal self-guided similarity preserving projections (OSSPP), which seamlessly integrates the procedures of an adjacency graph learning and DR into a one step. Specifically, OSSPP projects the data into a low-dimensional subspace and simultaneously performs similarity preserving learning by using the similarity preserving regularization term in which the reconstruction coefficients of the projected data are used to encode the similarity structure information. An interesting finding is that the problem to determine the reconstruction coefficients can be converted into a weighted non-negative sparse coding problem without any explicit sparsity constraint. Thus the projections obtained by OSSPP contain natural discriminating information. Experimental results demonstrate that OSSPP outperforms state-of-the-art methods in DR.

Index Terms— similarity preserving , dimensionality reduction, sparse coding

1. INTRODUCTION

Vision data such as face images, video frames and web documents are often high-dimensional. The high dimensionality of data not only increases the computation cost and memory requirements, but also adversely affects algorithmic performance [1]. It is therefore necessary to transform the original high-dimensional data into lower dimensional but more informative subspace [2]. In the literature, many dimensionality reduction (DR) methods have been proposed for such purpose, such as principle component analysis (PCA) [3] [4], Laplacian eigenmap (LE) [5], locality preserving projections (LPP) [6], local learning projections (LLP) [7] and linear discriminant analysis (LDA) [8]. Yan et al. further reformulated some dimensionality reduction methods, including unsupervised and supervised methods into a unified graph embedding framework [9].

Graph based learning methods attract special attention owing to its computation efficiency and excellent clustering capability [10] [11] [12]. Graph has been successfully applied in characterizing pairwise data relationship and manifold exploration. A number of graph construction methods have been proposed, including ℓ_1 graph [13], sparse probability graph (SPG) [14] and low-rank representation graph (LRR graph) [15]. Although these methods obtain empirical success, there are still some disadvantages. For example, almost all these methods construct the graph structure on the original high-dimensional features space, which is unnecessary to be best for characterizing the pairwise data relationship due to the fact that some unfavorable features may exist in the original data. This drawback can greatly reduce the performance of the designed algorithms. Intuitively, DR can address this problem since DR may remove the unfavorable features.

Inspired by above insights, we propose a novel orthogonal self-guided similarity preserving projections (OSSPP) method, in which the DR and the similarity matrix learning are simultaneously conducted so that the similarity matrix is constructed on the derived optimal low-dimensional subspace. Specifically, the similarity structure information of the data is encoded by the reconstruction coefficients of the projected data, and the projected data are required to respect the similarity structure by the similarity preserving regularization term. Thus OSSPP provides us an method for DR by the learned projections. OSSPP is the first work which uses the reconstruction coefficients of the projected data to encode the similarity structure information of data, and at the same time the projected data are required to respect the similarity structure during the procedure of DR. Although OSSPP is an unsupervised dimensionality reduction method, the projections learned by OSSPP contain natural discriminating information since the problem to determine the reconstruction coefficients can be converted into a weighted non-negative sparse coding problem without any explicit sparsity constraint.

The remainder of this paper is organized as follows. Section 2 introduces the details of OSSPP for dimensionality reduction. Experimental results are presented in Section 3. Finally, Section 4 concludes our paper.

*Corresponding author: yongxu@ymail.com

2. ORTHOGONAL SELF-GUIDED SIMILARITY PRESERVING PROJECTIONS

2.1. Motivation of Our Method

Unlike previous DR methods which firstly encode the similarity structure information of data as the graph relationship and then enforce the projected data to respect the graph structure, OSSPP uses the reconstruction coefficients of the projected data to encode the similarity structure information and simultaneously requires the projected data to respect the similarity structure during the procedure of DR. Moreover, the graph based learning methods empirically formulate the graph structure on the original high-dimensional features space, which can greatly reduce the performance of designed algorithm due to some unfavorable features. However, the performance of designed algorithm can be improved by employing DR on account of removing some unfavorable and redundant features. Therefore, it is necessary to integrate the DR and graph construction into a unified framework to learn an optimal projections.

2.2. Orthogonal Self-guided Similarity Preserving Projections (OSSPP)

In this subsection, we introduce our orthogonal self-guided similarity preserving projections (OSSPP) method which can be used to perform DR. Assume that $X = [x_1, x_2, ..., x_n] \in \Re^{m\times n}$ be a collection of n training samples $\{x_i \in \Re^m\}_{i=1}^n$ and each sample x_i can be well represented by a linear combination of the samples from the same class as x_i in X. That is,

$$x_i = Xz_i, \quad z_{ii} = 0, \quad \|z_i\|_0 \leq d_\ell \tag{1}$$

which can be rewritten as

$$min\|Z\|_1, \quad s.t.\ X = XZ, \quad diag(Z) = 0 \tag{2}$$

where $\| \bullet \|_1$ is the ℓ_1-norm regularization, z_i is the reconstruction coefficient vector and $Z = [z_1, z_2, \ldots, z_n]$.

Let $P \in \Re^{t\times m}$ be a projection matrix that projects the training samples from the original high-dimensional feature space $\Re^m$ into an output space of dimensionality t. Based on the above insights in 2.1, we propose the following objective function for OSSPP.

$$[P^*, Z^*] = \min_{P,Z} F(P, Z) \tag{3}$$

$$s.t.\ P^TP = I,\ \ diag(Z) = 0,\ \ Z \geq 0,\ \ \forall i$$

where

$$F(P, Z) = \|P^TX - P^TXZ\|_F^2 + \alpha\|X - PP^TX\|_F^2 + \beta\sum_{i=1}^{n}\sum_{j=1}^{n}\|P^Tx_i - P^Tx_j\|^2 Z_{ij}.$$

where the reconstruction coefficient matrix Z is required to be non-negative so that it can be directly used as graph weights and the column of the projections P is required to be orthogonal, which can prevent the solution from becoming degenerate and lead to a computationally efficient scheme for optimization [16]. The goal of the first term of F is to insure the reconstruction of data in the reduced space. The second term is a PCA-like regularization term, which ensures that the projections can hold the main energy of data. The last term in function F is the similarity preserving regularization term which aims to require the projected data to respect the similarity structure during the procedure of DR. α and β are the tradeoff parameters. In addition, although we do not explicitly impose the sparsity constraint on the reconstruction coefficients, the problem to determine the reconstruction coefficients matrix is naturally converted to a weighted non-negative sparse coding problem. In this way, OSSPP allows the reconstruction coefficients matrix to have sparsity and neighborhood adaptive.

2.3. Optimization

In this section, we propose an iterative update rule to solve the problem (3) of OSSPP. Specifically, the first step of the optimization algorithm is to solve P by fixing Z and the second step is to solve Z by fixing P.

Solve P by fixing Z: If variables Z is fixed, the optimization problem defined in (3) is written as

$$P^* = \arg\min_P \|P^TX - P^TXZ\|_F^2 + \alpha\|X - PP^TX\|_F^2 + \beta Tr(P^TXLX^TP) \tag{4}$$

$$s.t.\ P^TP = I$$

where $L = D - Z$ is graph Laplacian, in which D is a diagonal matrix with $D_{jj} = \sum_k Z_{jk}$. $Tr(\cdot)$ is the trace operation of matrix.

Considering the constraint $P^TP = I$, (4) can be further transformed into

$$P^* = \arg\min_P Tr(P^T(X - XZ)(X - XZ)^TP) + \alpha Tr(X^TX - P^TXX^TP) + \beta Tr(P^TXLX^TP) \tag{5}$$

which can be written as

$$P^* = \arg\min_P Tr(P^T(M - \alpha XX^T + \beta XLX^T)P) \tag{6}$$

$$s.t.\ P^TP = I$$

where $M = (X - XZ)(X - XZ)^T$. The solution of (6) can be obtained by solving the minimum eigenvalues problem:

$$(M - \alpha XX^T + \beta XLX^T)p_i = \lambda p_i \tag{7}$$

Let $P = [p_1, ..., p_d]$ be the solution of (7). These column vectors $p_i s$ $(i = 1, ..., d)$ correspond to the eigenvectors associated with the smallest d eigenvalues.

Solve Z by fixing P: If variables P is fixed, the optimization problem defined in (3) is written as

$$\min_{Z} \|P^T X - P^T X Z\|_F^2 + \beta \sum_{i=1}^{n} \sum_{j=1}^{n} \|P^T x_i - P^T x_j\|^2 Z_{ij} \tag{8}$$

$$s.t.\ diag(Z) = 0,\ \ Z \geq 0$$

which can be rewritten as

$$\min_{Z} \|H - HZ\|_F^2 + \beta Tr(\Theta(R \odot Z)) \tag{9}$$

$$s.t.\ diag(Z) = 0,\ \ Z \geq 0,\ \ \forall i$$

where $H = P^T X = [h_1, ..., h_n] \in \Re^{d \times n}$, $R_{ij} = \|P^T x_i - P^T x_j\|^2 (R = [r_1, ..., r_n] \in \Re^{n \times n})$ and $\Theta \in \Re^{n \times n}$ is a matrix with all elements as 1. The optimization problem in (9) can be decomposed into n independent sub-problems for each coding coefficient $z_i (i = 1, ..., n)$ corresponding to $h_i (i = 1, ..., n)$ and each sub-problem is a weighted non-negative sparse coding problem:

$$\min_{z_i} \sum_{k}^{n} r_i^k z_i^k + \beta \|h_i - H z_i\|^2 \tag{10}$$

$$s.t.\ \ z_i \geq 0,\ \ z_i^i = 0,\ \ \forall i$$

where z_i^k and r_i^k are the kth elements of the vectors z_i and r_i, respectively. Many algorithms can be used to solve (10), such as basis pursuit (SP)[17] and fast iterative shrinkage and thresholding (FISTA)[18]. Here, the alternating direction method (ADM) [19] [20] is used to solve the optimization problem (10).

The above two steps are iteratively conducted to obtain the solution for (3). The overall algorithm of OSSPP is described in detail in Algorithm 1.

Algorithm 1 : OSSPP

Input: Training samples matrix X; Parameters α, β; Dimensionality of low-dimensional feature space d;
Initialization: Initializing Z as a similarity matrix by k nearest neighbor graph;
while not converged **do**
1. Update P by solving (6)
2. Update Z by solving (9)

end while
Output: Projections P

3. EXPERIMENTS AND ANALYSIS

In this section, we apply OSSPP for dimensionality reduction along with showing our experimental results. In summary, let P^* be the solution of (3), then we use the obtained P^* to perform the dimensionality reduction.

3.1. Experiment Settings

Four public datasets are selected for our experiments: USPS digit image data set [21], COIL20 data set and two face image data sets, i.e. YaleB [22] and CMU PIE (PIE) datasets [23]. In our experiments, the USPS data set contains 9298 handwritten digit images with 16×16 pixels. COIL20 data set consists of images of 20 objects, and each object has 72 images with 32×32 pixels captured from varying angles at intervals of five degrees. The YaleB dataset has 38 individuals, each subject has around 64 near frontal images with 32×32 pixels under different illuminations. The images of the PIE data set used are from the frontal pose (C27) and each subject has around 49 images with 64×64 pixels from varying illuminations and facial expressions. For the sake of computational efficiency, PCA is used as a preprocessing step to preserve 98% energy of data for the USPS data set, and 95% energy of data for the YaleB, COIL20 and PIE face data sets, respectively.

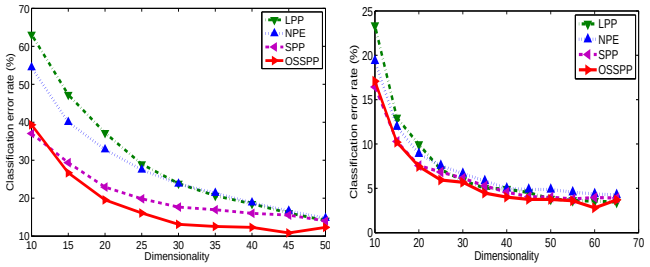
3.2. Experimental Results on Dimensionality Reduction

Unsupervised dimensionality reduction is a fundamental step in pattern recognition. In our OSSPP method, the projection P^* is useful for the similarity preserving. After learning P^* from the training set, it is straightforward to use P^* to map both of the training samples and test samples into the desired low-dimensional subspace, and then utilize nearest neighbor (NN) classifier to predict the labels of test samples. In the experiments, we only use the NN classifier (based on Euclidean distance) to perform classification due to space limit. For each data set, we randomly select different training samples from per subject for training and rest for testing and all experiments are run 10 times (unless otherwise stated) and then the mean classification accuracy and standard deviation are reported.

We compare OSSPP with some popular unsupervised dimensionality reduction methods including PCA, LPP [6], NPE [24] and sparsity preserving projection (SPP)[25]. Table 1 shows that classification results on these data sets, in which $\#Tr$ and d denote the optimal number of training samples selected from each subject of the data set and the optimal dimensionality. From Table 1, one can see that PCA generally gets much worse performance than LPP, NPE, SPP and OSSPP. Moreover, LPP and NPE generally outperform PCA with lower dimensionalities. This indicates that by preserving the local structure of the data, the classification accuracy can be improved. That is, when NN classifier (nearest neighbor search) is used, local structure seems to be important than global structure. In addition, OSSPP consistently outperforms all the compared methods with NN classifier. This suggests that the orthogonal projections learned by OSSPP contain more discriminating information than those of the compared methods, which is benefit from the weighted non-negative sparse coding for the solution of reconstruction coefficients matrix Z. Furthermore, unlike SPP, we do not impose explicitly sparsity constraint on the reconstruction

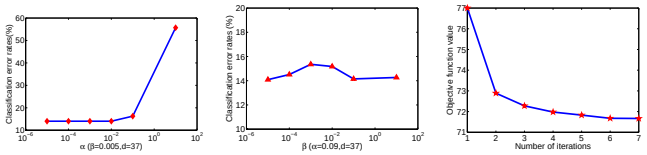
Data set ($\#Tr$)	PCA (d)	LPP (d)	NPE (d)	SPP (d)	OSSPP (d)
USPS (10)	27.09±1.73 (49)	27.00±1.86 (20)	26.07±2.30 (20)	19.30±1.28 (47)	**15.59±1.36 (48)**
USPS (20)	21.25±1.78 (46)	19.60±0.71 (26)	17.86±2.05 (20)	13.22±0.78 (39)	**10.60±0.67 (47)**
USPS (30)	17.95±0.78 (50)	16.13±0.88 (32)	14.27±0.68 (40)	11.27±0.54 (38)	**8.76±0.66 (48)**
COIL20 (3)	39.70±2.19 (48)	-	-	22.67±2.07 (20)	**20.68±2.77 (83)**
COIL20 (5)	33.61±1.69 (48)	38.62±3.96 (24)	40.52±1.93 (35)	**15.16±1.83 (30)**	**15.30±1.96 (30)**
COIL20 (7)	28.01±1.78 (44)	25.52±1.82 (37)	24.07±2.19 (54)	12.23±1.45 (58)	**11.89±2.06 (28)**
YaleB (20)	35.42±1.59 (318)	17.19±0.34 (61)	16.65±0.96 (61)	16.08±0.68 (61)	**13.89±1.12 (37)**
YaleB (30)	26.75±1.56 (325)	14.44±0.94 (61)	14.21±0.93 (61)	12.51±0.92 (61)	**9.59±2.01 (30)**
YaleB (40)	21.42±1.14 (450)	13.33±0.81 (61)	12.56±0.89 (61)	11.59±1.08 (61)	**8.86±0.95 (30)**
PIE (15)	30.52±0.91 (410)	8.32±0.74 (68)	5.83±0.75 (63)	4.91±0.43 (68)	**3.06±0.31 (58)**
PIE (20)	25.45±1.17 (290)	6.31±0.65 (68)	4.03±0.73 (55)	3.52±0.31 (53)	**3.02±0.34 (60)**
PIE (25)	22.73±1.07 (280)	4.93±0.59 (68)	3.34±0.52 (60)	3.00±0.46 (68)	**2.54±0.42 (60)**

Table 1: Classification error rates (mean classification error rates ± standard deviation %) of different algorithms with NN classifier under different number of training samples. The bold numbers are the lowest error rates.



(a) (b)

Fig. 1: Classification error rates (%) of different algorithms with NN classifier versus dimensionalities on the (a) YaleB and (b) PIE face data sets. For the YaleB and PIE data sets, we randomly select 30 and 20 samples per subject for training and use the remaining for testing, respectively.



(a) (b) (c)

Fig. 2: Parameters sensitivity and convergence: (a) and (b) show the performance of OSSPP vs. the parameters α and β, respectively. (c) shows the convergence curve of OSSPP. We randomly select 20 images per subject for training and use the remaining for testing on the YaleB data set.

coefficient matrix.

The classification error rates versus dimensionalities on the YaleB and PIE data sets are shown in Fig. 1. We compare the dimensionalities up to 50 and 70 for the YaleB and PIE data sets, respectively. Again, OSSPP performs better than the other methods. We examine the parameter sensitivity of OSSPP to classification error rate. α is to hold the main energy of data, while β is to ensure the similarity preserving on the projections. Fig. 2 shows the parameters sensitivity and convergence of OSSPP. From Fig. 2 (a), we can see that the performance of OSSPP is robust to the parameter α when $\alpha \leq 10^{-2}$. OSSPP is not sensitive to the parameter β in the given wide range (see Fig. 2 (b)). In practice, we first fix α due to its more stronger robustness than β, and then select the optimal value of β from the given set. Fig. 2 (c) shows that the objective function values decreases very fast. After only about 6-7 iterations, the objective value converges, which suggests that our iterative update rule is very effective.

4. CONCLUSION

This paper proposes a novel DR method, called orthogonal self-guided similarity preserving projections (OSSPP) for DR. The core idea of OSSPP is that OSSPP uses the reconstruction coefficients of projected data to encode the similarity structure information and requires the projected data to respect the similarity structure during the procedure of DR. Extensive experiments on DR show the effectiveness of the proposed method.

5. REFERENCES

[1] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. on PAMI*, vol. 35, no. 11, pp. 2765–2781, 2013.

[2] B. Ni, S. Yan, and S. Kassim, "Learning a propagable

graph for semisupervised learning: classification and regression," *IEEE Trans. on Knowledge and Data Engineering*, vol. 24, no. 1, pp. 114–126, 2012.

[3] Z. Fan, Y. Xu, W. Zuo, J. Tan, Z. Lai, and D. Zhang, "Modified principal component analysis: An integration of multiple similarity subspace models," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1538–1552, 2014.

[4] X. Li, Y. Pang, and Y. Yuan, "L1-norm-based 2dpca," *IEEE Trans. Syst., Man, and Cybern., B, Cybern*, vol. 40, no. 4, pp. 1170–1175, 2010.

[5] B. Mikhail and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *NIPS*, vol. 14, pp. 585–591, 2001.

[6] X Niyogi, "Locality preserving projections," *NIPS*, vol. 16, pp. 153, 2004.

[7] M. Wu, K. Yu, S. Yu, and B Scholkopf, "Local learning projections," *ICML*, vol. 16, pp. 1309–1046, 2007.

[8] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," *in IJCAI'07*, 2007.

[9] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yan, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. on PAMI*, vol. 29, no. 1, pp. 40–51, 2007.

[10] X. Zhou, "Semi-supervised learning literature survey," *Technical Report 1530, Department of Computer Sciences, University of Wisconsin, Madison, 2005*, 2005.

[11] B. Nick and X. Zhou, "p-voltages: Laplacian regularization for semi-supervised learning on high-dimensional data," *In Eleventh Workshop on Mining and Learning with Graphs (MLG2013)*, 2013.

[12] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," *in ICCV 2007*, 2007.

[13] S. Yan and H. Wang, "Semi-supervised learning by sparse representation," *In SIAM Intl Conf. on Data Mining, SDM*, pp. 792–801, 2009.

[14] R. He, W. Zheng, B. Hu, and W. Kong, "Non-negative sparse coding for discriminative semi-supervised learning," *In CVPR*, pp. 2849–2856, 2011.

[15] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," *In ICML*, 2010.

[16] V. Patel, H. Nguyen, and R. Vidal, "Latent space sparse subspace clustering," *in ICCV 2013*, 2013.

[17] J. Yang, D. Chou, L. Zhang, Y. Xu, and Yang. J., "Sparse representation classifier steered discriminative projection with applications to face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1023–1035, 2013.

[18] L. Zhang, W. Zhou, and P. Chang, "Kernel sparse representation-based classifier," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1684–1695, 2012.

[19] J. Yang and Y. Zhang, "Alternating direction algorithms for l1-problems in compressive sensing," *SIAM Journal on Scientific Computing*, vol. 33, no. 1, pp. 250–278, 2011.

[20] Y. Zhang, J. Yang, and W. Yin, "Yall1: Your algorithms for l1," *online at http://yall1.blogs.rice.edu/*, 2011.

[21] M. Zheng, J. Bu, C. Chen, and C. Wang, "Graph regularized sparse coding for image representation," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1327–1336, 2011.

[22] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, "Non-negative low rank and sparse graph for semi-supervised learning," *in CVPR 2012*, 2011.

[23] Z. Lai, W. Wong, Z. Jin, J. Yang, and Y. Xu, "Sparse approximation to the eigensubspace for discrimination," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 23, no. 12, pp. 1948–1960, 2012.

[24] X. He, D. Cai, S. Yan, and H Zhang, "Neighborhood preserving embedding," *in ICCV'05*, 2005.

[25] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognition*, vol. 43, no. 1, pp. 331–341, 2010.