

Metric Learning: A General Dimension Reduction Framework for Classification and Visualization

Chunyuan Lu^{1*}, Guocan Feng¹, Jianmin Jiang² and Patrick Wang³

^{1*,1}*School of Mathematics and Computational Science, Sun Yat-Sen University, China, 510275*

²*School of Informatics, University of Bradford, BD7 1DP, UK*

³*College of Computer and Information Science, Northeastern University, Boston, MA, USA, 02115*

^{1*}luchuny@mail2.sysu.edu.cn, ¹mcsfgc@mail.sysu.edu.cn,

²j.jiang1@bradford.ac.uk, ³pwang@ccs.neu.edu

Abstract

A new general dimension reduction framework based on similar and dissimilar metric learning is proposed in this paper which allows us to exploit the geometry of data to reduce the data dimension for classification and visualization. The general formulation can unify the existing dimension reduction algorithms within a common framework. Furthermore, this metric learning framework can be used as a general platform for developing new dimension reduction algorithms. By utilizing this framework as a tool, we propose a novel supervised dimension reduction algorithm named Sub-Manifold Preserving Analysis (SMPA) in which the intrinsic sub-manifold structure will be preserved while the margin of interclass will be separated. Experimental evidences show that performance of our proposed SMPA algorithm is better than other algorithms.

1. Introduction

Dimension reduction arises often in machine learning and pattern recognition. Traditional techniques such as Principal Component Analysis (PCA) [1], multidimensional scaling (MDS) [2] and Linear Discriminant Analysis (LDA) [3] can only find the linear structure of high-dimensional data. Recently, ISOMAP [4], locally linear embedding (LLE) [5], Laplacian Eigenmap (LE) [6] and other manifold learning methods try to cope with nonlinear structure and related application problems when data lie on or around a lower dimensional manifold.

In this paper, there are two contributions to dimension reduction and applied tasks. First, we present a general framework that offers a unified view for understanding and explaining dimension reduction algorithms such as PCA, LDA, ISOMAP, LLE, Laplacian Eigenmap, Locality Preserving

Projection (LPP) [7], Neighborhood Preserving Embedding (NPE) [8], and Marginal Fisher Analysis (MFA) [9].

Second, this framework can be used as a general platform for developing new dimension reduction algorithms. We accomplish this task by designing similar matrix and dissimilar matrix according to specific motivations. We develop a novel dimension reduction algorithm, Sub-Manifold Preserving Analysis (SMPA) for classification and visualization which aims at preserving the sub-manifold structure of the intraclass data and separating the interclass data. Remarkably, SMPA can achieve classification and visualization both better than other methods.

The rest of this paper is organized as follows: the general dimension reduction framework is introduced in section 2. In section 3, we present our SMPA algorithm with linear form and kernel form. Section 4 reports the experimental results of our proposal. Finally, we give the conclusions in Section 5.

2. Metric Learning: A General Framework for Dimension Reduction

Let $\{x_1, x_2, \dots, x_N; x_i \in \mathbf{X}\}$ be a data set of M -dimension and N_i denote the number of samples belonging to the i th class, $\sum_{i=1}^C N_i = N$. There are C classes and i th class denote by π_i . Dimension reduction $Y=f(X)$ maps $X=[x_1, \dots, x_N]$ into $Y=[y_1, \dots, y_N]$ where $m \ll M$. If f is a linear transform, then $Y=[\omega^T]_{m \times M} X$. $[\omega]_{M \times m}$ is a projection matrix. In [9] and [10], different dimension reduction algorithms have been reformulated within a graph embedding framework. In this section, we present a novel unifying framework of metric learning to provide a common perspective in understanding the

relationship of these algorithms and designing new algorithms. Our metric learning framework is more intuitive and general.

2.1. Metric Learning General Framework for Dimensionality Reduction

Good metric is very important for real problems. We set up a general dimension reduction framework based on metric learning. First, we establish our framework from three lemmas.

Lemma 1 Given a symmetric matrix A and a positive definite matrix B , the first m generalized eigenvectors of the generalized eigenvectors problems $Ay = \lambda By$ ($y \in R^m$) is an optimal solution of the constrained maximization problem:

$$\begin{cases} y^* = \arg \max & w(y^T Ay) \\ \text{subject to:} & \text{diag}(y^T By) = d \end{cases}$$

where d is a diagonal matrix. Similarly, the solution of the corresponding minimization problem is the last m generalized eigenvectors of (A, B)

For symmetric pairwise dissimilarity matrix D and similarity matrix S , which D_{ij} and S_{ij} denote the dissimilarity and the similarity of x_i and x_j separately. Define the associated Laplacian matrix L_D and L_S as:

$$(L_D)_{ij} = \begin{cases} \sum_{j \neq i} D_{ij} & i = j \\ -D_{ij} & i \neq j \end{cases}, (L_S)_{ij} = \begin{cases} \sum_{j \neq i} S_{ij} & i = j \\ -S_{ij} & i \neq j \end{cases} \quad (1)$$

Derive from the relationship of D and L_D , we can get lemma 2:

Lemma 2 For the maximization problem $\sum_{i,j} D_{ij} \|y_i - y_j\|^2$ the optimal solution is the first m eigenvectors of the matrix $y^T L_D y$. Similarly, for the minimization problem $\sum_{i,j} S_{ij} \|y_i - y_j\|^2$ the optimal solution is the last m eigenvectors of the matrix $y^T L_S y$.

$$\begin{cases} \arg \max \sum_{i \neq j} D_{ij} \|y_i - y_j\|^2 \\ \text{s.t. : } \text{diag}(y^T L_S y) = I \end{cases} \quad \text{and} \quad \begin{cases} \arg \max \frac{\sum_{i \neq j} D_{ij} \|y_i - y_j\|^2}{\sum_{i \neq j} S_{ij} \|y_i - y_j\|^2} \\ \text{s.t. : } \text{diag}(y^T L_S y) = d \end{cases}$$

are equivalent. According to lemma 1, the solution of this problem is given by the first m generalized eigenvectors of (L_D, L_S) .

Lemma 3 For the maximization problem and the minimization problem:

$$\begin{cases} \arg \max \frac{\sum_{i,j} D_{ij} \|y_i - y_j\|^2}{\sum_{i,j} S_{ij} \|y_i - y_j\|^2} \\ \text{s.t. : } \text{diag}(y^T L_S y) = d \end{cases}, \begin{cases} \arg \min \frac{\sum_{i,j} S_{ij} \|y_i - y_j\|^2}{\sum_{i,j} D_{ij} \|y_i - y_j\|^2} \\ \text{s.t. : } \text{diag}(y^T L_D y) = d \end{cases}$$

the optimal solution is the first m generalized eigenvectors of (L_D, L_S) and the last m generalized

eigenvectors of (L_S, L_D) separately.

So our dimension reduction framework can be given as following figure:

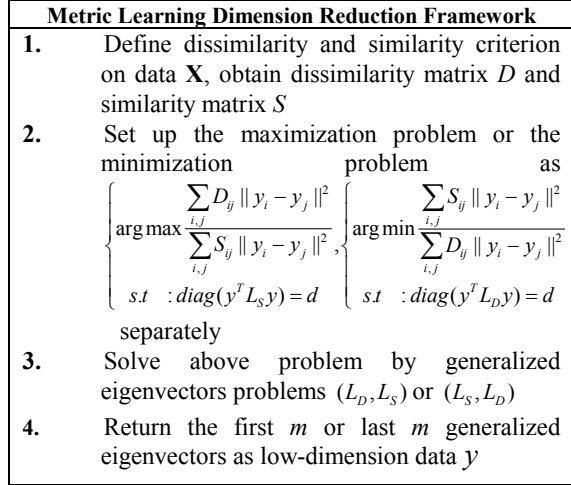


Fig. 1. Metric Learning Dimension Reduction Framework
Our framework founded on optimal problem, generalized eigenvalues decomposition and Metric Learning. It possesses several advantages:

- (i) The general framework can unify the existing dimension reduction algorithms especially manifold learning methods, including LLE, ISOMAP, LPP and MFA etc. in a common metric learning framework.
- (ii) In terms of different task we can add different constrains into the optimal problem, so it is very flexible.

2.2. Illustration Existing Dimensionality Reduction method in Our Framework

In this section, we show that the previously mentioned dimension reduction algorithms can be reformulated within the presented metric learning framework. The differences between these algorithms lie on the selection of the dissimilarity and similarity criterion.

$\alpha \in R^+$ is a constant. We briefly list the choices of D and S for these algorithms as follows.

PCA: $S_{ij} = \begin{cases} \alpha & i \neq j \\ \text{arbitrary} & i = j \end{cases}, D_{ij} = \begin{cases} \alpha & i \neq j \\ \text{arbitrary} & i = j \end{cases}$

LDA:

$$S_{ij} = \begin{cases} \frac{\alpha}{N_k} & i \neq j, x_i, x_j \in \pi_k \\ \text{arbitrary} & i = j \end{cases}, D_{ij} = \begin{cases} \alpha & i \neq j \\ \text{arbitrary} & i = j \end{cases}$$

ISOMAP:

$$S_{ij} = \begin{cases} \alpha \cdot \tau(d^{\circ})_{ij} & i \neq j \\ \text{arbitrary} & i = j \end{cases}, D_{ij} = \begin{cases} \alpha & i \neq j \\ \text{arbitrary} & i = j \end{cases}$$

where $\tau(d^{\circ})$ is the normalized geodesic distance matrix

LLE/NPE: Let W be a local reconstruction coefficient matrix which is defined as follows:

$W_{ij} = 0$ if $x_j \notin N_k(x_i)$, otherwise W_{ij} obtain

$$\text{by: } \begin{cases} \arg \min \|x_i - \sum_{j \in N_k(x_i)} W_{ij} x_j\|^2 \\ \text{subject to: } \sum_{j \in N_k(x_i)} W_{ij} = 1 \end{cases} . \text{ It is clear that:}$$

$$S_{ij} = \begin{cases} \alpha \cdot (W_{ij} + W_{ji} - \sum_k W_{jk} W_{ki}) & i \neq j \\ \text{arbitrary} & i = j \end{cases}, D_{ij} = \begin{cases} \alpha & i \neq j \\ \text{arbitrary} & i = j \end{cases}$$

LE/LPP:

$$S_{ij} = \begin{cases} \alpha \cdot e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} & i \neq j, j \in N_k(x_i) \text{ or } i \in N_k(x_j) \\ \text{arbitrary} & i = j \end{cases},$$

$$D_{ij} = \begin{cases} \alpha \cdot \frac{\sum_j D_{ij}}{N-1} & i \neq j \\ \text{arbitrary} & i = j \end{cases}$$

MFA: $S_{ij} = \begin{cases} \alpha & x_i, x_j \in \pi_k, j \in N_k(x_i) \text{ or } i \in N_k(x_j) \\ 0 & \text{otherwise} \end{cases},$

$$D_{ij} = \begin{cases} \alpha & x_i \in \pi_k, x_j \notin \pi_k, j \in N_{k2}(x_i) \text{ or } i \in N_{k2}(x_j) \\ 0 & \text{otherwise} \end{cases}$$

It is obvious that previous methods are mainly focus on similarity matrix.

3. Sub-Manifold Preserving Analysis (SMPA)

In order to make the best use of giving data we focus on dissimilarity information and similarity information both. We regard data with same class lie on or around a sub-manifold. Data with different class are lie on different sub-manifolds. So in order to keep the sub-manifold structure while separate interclass data, define SMPA dissimilarity matrix and similarity matrix as follows:

$$D_{ij} = \begin{cases} \lambda_1 d_{ij}^G & x_i, x_j \in \pi_k \\ \lambda_2 / d_{ij}^G & x_i \in \pi_k, x_j \notin \pi_k \end{cases}, k=1, \dots, C \quad (2)$$

$$S_{ij} = \begin{cases} \lambda_3 / d_{ij}^G & x_i, x_j \in \pi_k \\ 0 & x_i \in \pi_k, x_j \notin \pi_k \end{cases}, k=1, \dots, C \quad (3)$$

where d^G is the normalized geodesic distance matrix (d^G is the same as it used in ISOMAP), λ_1, λ_2 and λ_3 are three constants.

It is clear that we assume interclass data has no similarity for classification. The similarity is bigger than dissimilarity in intraclass data, and interclass data dissimilarity is bigger than intraclass data dissimilarity. Moreover, if interclass samples x_i and x_j are closer, D_{ij} is bigger since they are more likely to be mislabeled.

3.1. Linear SMPA

Linear SMPA is linear approximations to the SMPA algorithm which can find nonlinear sub-manifold structure. We have the Linear SMPA criterion:

$$\begin{cases} \omega^* = \arg \max_{\omega} \frac{\sum_{i,j} D_{ij} \|\omega^T x_i - \omega^T x_j\|^2}{\sum_{i,j} S_{ij} \|\omega^T x_i - \omega^T x_j\|^2} \\ \text{s.t. } \text{diag}(x^T \omega L_S \omega^T x) = d \end{cases} \quad (4)$$

ω is a projection matrix.

It is worthwhile to generalize several aspects of the proposal here:

- (i) Linear SMPA can discover the sub-manifold intrinsic structure of the data.
- (ii) Linear SMPA makes it fast and suitable for practical applications and easy to extend to Kernel SMPA.
- (iii) Linear SMPA is defined everywhere in ambient space rather than just on the training data points. It can solve the out-of-sample problem easily.

3.2. Kernel SMPA

Kernel trick has been used to enhance the nonlinear separability of the linear SMPA in this paper. Suppose that the Euclidean space $X \subset R^D$ is mapped to a Hilbert space H through a nonlinear mapping function $\Phi: X \rightarrow \Phi(X)$. Kernel function is $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle = \Phi^T(x_i) \Phi(x_j)$. Because the kernel eigenvectors are linear combinations of $\Phi(x_1), \dots, \Phi(x_N)$, so $\omega = \sum_{i=1}^N \alpha_i \Phi(x_i)$ is the projection direction in new feature space, then the optimal α can be obtained by:

$$\begin{cases} \alpha^* = \arg \max_{\alpha} \frac{\sum_{i,j} D_{ij} \|\sum_j \alpha_j K(x_i, x_j) - \sum_i \alpha_i K(x_i, x_j)\|^2}{\sum_{i,j} S_{ij} \|\sum_j \alpha_j K(x_i, x_j) - \sum_i \alpha_i K(x_i, x_j)\|^2} \\ \text{s.t. } \text{diag}(\sum_i \alpha_i K(x_i, x_i))' L_S \sum_i \alpha_i K(x_i, x_i) = d \end{cases} \quad (5)$$

The generalized eigenvectors problems (L_D, L_S) , in the Hilbert space can be written as follows: $\Phi^T(X) L_D \Phi(X) = \lambda \Phi^T(X) L_S \Phi(X)$. By simple algebra formulation, we can finally obtain the following eigenvector problem: $KL_D K \alpha = \lambda KL_S K \alpha$. For a new sample x_{new} , the low-dimension projection is $y_{new} = \langle \omega, \Phi(x) \rangle = \sum_i \alpha_i \langle \Phi(x), \Phi(x_i) \rangle = \sum_i \alpha_i K(x, x_i)$.

4. Experiments

In this section, we give a set of experiments to demonstrate our proposed Linear SMPA method and Kernel SMPA in the capability of visualization and classification; including synthesized examples and face recognition. We set scaling factors $\lambda_1 = \lambda_2 = \lambda_3 = 1$ for its simplicity in this paper.

4.1. Synthesized Visualization Problems

One of the most important aspects of exploratory data analysis is data visualization [14]. In this part of

experiment we use two synthesized examples to illustrate the effectiveness of our SMPA method for visualization. As shown in Fig.2. we give some 2-D data, they come from 2 classes. We show the 1-D projection after dimension reduction with LDA, MFA, LPP and SMPA.

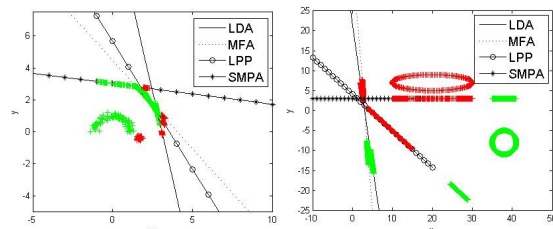


Fig. 2. Recognition rate versus dimension of reduced space. From the slope of each projection line, it is obvious that without ruining their separation, the different structure of the two class data is preserved better in our method.

4.2. Face Recognition Problem

A great amount of face recognition algorithms have been proposed in this area [1, 3, 9, 10, 13, 15] during the past two decades.

In order to test the classification ability of SMPA, second parts of our experiments are performed on two face databases: 1) The Yale database [11]; 2) The YaleB database [12]. In the Yale database, 165 frontal face images cover 15 individuals taken under 11 different conditions. Each individual has different facial expressions, illumination conditions and small occlusion. There are 10 individuals under 64 different lighting conditions for 9 poses in the YaleB Database. Only frontal face images under varying lighting conditions are used in this paper. In all the experiments, images have been cropped into 32×32 pixels.

We average the results over T random splits. A random subset with k images per individual form the training set. In the Yale database we set $T=10$, $k=6$, and $T=5$, $k=40$ in the YaleB database.

In order to show the performance of the proposed SMPA algorithm, we use LDA, LPP and MFA as benchmarks in comparison with our proposed algorithm in the experiment.

Table 1. Recognition rate on Yale and YaleB database.

	Yale (dimensions of reduction space)	YaleB (dimensions of reduction space)
LDA	78.7% (14)	96.8% (9)
LPP	78% (30)	96.8% (30)
MFA	62.2% (30)	95.2% (30)
LSMPA	79.9% (30)	98.2% (30)
KSMPA	81.4% (30)	97.6% (30)

It can be seen in Table.1 that our method is improve over other methods.

5. Conclusions

In this paper, we provide a novel framework of dimension reduction. This framework brings together ideas from the theory of optimal problem, generalized eigenvalues decomposition and metric learning. It gives a unified perspective to explain and understand existing dimension reduction algorithms. Moreover, it can also be used as a platform to develop new algorithm for dimension reduction. And in order to testify our framework we have proposed a novel dimension reduction algorithm, named Sub-Manifold Preserving Analysis (SMPA). This algorithm can preserve the intraclass data structure and separate interclass data. The synthesized problem and the face recognition experiments have illustrated the good performance of our proposed Linear SMPA and its kernel extension.

References

- [1] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71-86, 1991.
- [2] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 1994.
- [3] P.N. Belhumeur, J.P. Hefanpha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on PAMI*, 19(7):711-720, 1997.
- [4] J. B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319-2323, 2000.
- [5] Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(22): 2323-2326, 2000.
- [6] M. Belkin and P. Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. *Advances in NIPS 15*, 2002.
- [7] X. He and Partha Niyogi. Locality Preserving Projections. *Advances in NIPS 16*, 2003.
- [8] X. He, D. Cai, S. Yan, H.-J. Zhang. Neighborhood preserving embedding. In *Proceedings of the Tenth IEEE ICCV*, 1208-1213, 2005.
- [9] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang and S. Lin. Graph embedding and extension: A general framework for dimensionality reduction. *IEEE Transactions on PAMI*, 29(1):40-51, 2007.
- [10] D. Cai, X. He, Y. Hu, J. Han, and T. Huang, "Learning a Spatially Smooth Subspace for Face Recognition", *IEEE Conference on CVPR*, 1-7, 2007.
- [11] Yale University, database available from: < <http://cvc.yale.edu/projects/yalefaces/yalefaces.html> >
- [12] Yale University, database available from: < <http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html> >
- [13] D. Cai, X. He and J. Han. Using Graph Model for Face Analysis. Technical Report, UIUC, 2005.
- [14] Y. Koren and L. Carmel, Visualization of Labeled Data Using Linear Transformations, *Proceedings of IEEE Information Visualization*, 121-128, 2003.
- [15] Xinge You, Dan Zhang, Qihui Chen, Patrick Wang and Yuan Yan Tang. Face Representation By Using Non-tensor Product Wavelets. *ICPR*, 503-506, 2006.