Neighborhood Preserving Projections (NPP): A Novel Linear Dimension Reduction Method

Yanwei Pang[1], Lei Zhang[2], Zhengkai Liu[1], Nenghai Yu[1], and Houqiang Li[1]

[1] Information Processing Center, University of Science and Technology of China, Hefei 230027, China
{pyw, zhengkai, ynh, lihq}@ustc.edu.cn
[2] Microsoft Research Asia, Beijing 100080, China
leizhang@microsoft.com

Abstract. Dimension reduction is a crucial step for pattern recognition and information retrieval tasks to overcome the curse of dimensionality. In this paper a novel unsupervised linear dimension reduction method, *Neighborhood Preserving Projections* (NPP), is proposed. In contrast to traditional linear dimension reduction method, such as principal component analysis (PCA), the proposed method has good neighborhood-preserving property. The main idea of NPP is to approximate the classical locally linear embedding (i.e. LLE) by introducing a linear transform matrix. The transform matrix is obtained by optimizing a certain objective function. Preliminary experimental results on known manifold data show the effectiveness of the proposed method.

1 Introduction

To deal with tasks such as pattern recognition and information retrieval, one is often confronted with the curse of dimensionality [1]. The dimensionality problem arises from the fact that there are usually few samples compared to the sample dimension. Due to the curse of dimensionality, a robust classifier is hard to be built and the computational cost is prohibitive. Dimension reduction is such a technique that attempts to overcome the curse of the dimensionality and to extract relevant features. For example, although the original dimensionality of the space of all images of the same subject may be quite large, its intrinsic dimensionality is usually very small [2].

Many dimension reduction methods have been proposed and can be categorized into linear (e.g. PCA, MDS and LDA) and non-linear (e.g. LLE, ISOMAP, Laplacian Eigenmap, KPCA and KDA) methods. The differences between these methods lie in their different motivations and objective functions. Principal component analysis (PCA) [3] may be the most frequently used dimension reduction method. PCA seeks a subspace that best represents the data in a least-squares sense. Multidimensional scaling (MDS) [3] finds an embedding that preserves the interpoint distances, and is equivalent to PCA when those distances are Euclidean. Linear discriminant analysis (LDA), a supervised learning algorithm, selects a transformation matrix in such a way that the ratio of the between-class

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part I, LNCS 3644, pp. 117–125, 2005.

scatter and the within-class scatter is maximized [4]. By nonlinearly mapping the input space to a high-dimensional feature space, PCA and LDA can be evolved into KPCA (kernel PCA) [5] and KDA (kernel discriminant analysis) [6]. Though, compared to their linear forms PCA and LDA, KPCA and KLDA can deal with nonlinear problem to some extent, it is difficult to determine the optimal kernels.

Recently, several nonlinear manifold-embedding-based approaches were proposed such as locally linear embedding (LLE) [7], isometric feature mapping (Isomap) [8] and Laplacian Eigenmaps [9]. They all utilize local neighborhood relation to learn the global structure of nonlinear manifolds. But they have quite different motivations and derivations. Limitations of such approaches include their demanding for sufficiently dense sampling and heavy computational burden. Moreover, the original LLE, Isomap and Laplcacian Eigenmaps can not directly deal with the out-of-sample problem [10] . Out-of-sample problem states that only the low dimensional embedding map of training samples can be computed but the samples out of the training set (i.e. testing samples) cannot be calculated directly, analytically or even cannot be calculated at all.

Soon after the aforementioned nonlinear manifold embedding approaches were developed, much endeavor is made to improve and extend them. More recently, locality preserving projections (LPP) [11] was proposed based on Laplacian Eigenmaps. When applied to face recognition, this method is called Laplacianfaces [12]. LPP is a linear dimension reduction method which is derived by finding the optimal linear approximations to the eigenfunctions of the Laplace Beltrami operator on the manifold. Besides its capacity to resolve the out-of-sample problem, LPP shares the locality preserving property. The locality preserving property makes LPP distinct from conventional PCA, MDS and LDA. Motivated by LPP, in this paper, we propose novel dimension reduction method which we call *Neighborhood Preserving Projections* (NPP). While LPP is derived from Laplacian Eigenmaps, ours is derived from LLE. Since the proposed method is a linear form of the original nonlinear LLE, NPP inherits LLE's neighborhood property naturally.

The rest of this paper is organized as follows: Section 2 gives an overview of the proposed method, NPP. Section 3 provides a brief description of LLE. In section 4, the motivation and justification of NPP is presented. Preliminary experimental results are shown in Section 5. Finally, conclusions are drawn in section 6.

2 Overview of the Proposed Method: NPP

2.1 Dimension Reduction Problem

Given N points $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$ in D dimensional space, dimension reduction is conducted such that these points are mapped to be new points $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]$ in d dimensional space where $d << D$. Dimension reduction can be performed either in linear way or in non-linear way. Original LLE is a non-linear

dimension reduction technique while our proposed method NPP is a linear one. For linear method, a linear transformation matrix is determined so that

$$\mathbf{y}_i = \mathbf{A}^T \mathbf{x}_i. \tag{1}$$

The transformation matrix is not computed in an arbitrary way, it is obtained, instead, according to a certain objective function. It is the objective function that makes our proposed linear dimension reduction algorithm, NPP, distinct itself from other algorithms. Before presenting a detailed derivation of NPP algorithm, we will give an overview of it in next subsection.

2.2 Overview

The first two steps of NPP algorithm are the same as those of LLE. Our main contribution lies in third step. The details will be given in section 3 and section 4.

Step 1. Assign neighbors to each data point $\mathbf{x}_i$ (for example by using the K nearest neighbors)

Step 2. Compute the weights W_{ij} that best linearly reconstruct $\mathbf{x}_i$ from its neighbors, solving the constrained least-squares problem in equation (3).

Step 3. Compute the linear transform matrix $\mathbf{A}$ by solving the generalized eigenvalue problem:

$$\mathbf{L}\mathbf{A}^T = \lambda \mathbf{C}\mathbf{A}^T. \tag{2}$$

Where

$$\mathbf{L} = \mathbf{X}\mathbf{M}\mathbf{X}^T$$

$$\mathbf{C} = \mathbf{X}\mathbf{X}^T$$

$$\mathbf{M} = (\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W})^T.$$

Note that we will explain step 3 in detail in section 4.

Step 4. Dimension reduction is performed simply by

$$\mathbf{Y} = \mathbf{A}^\mathbf{T}\mathbf{X}.$$

Because the proposed method is closely related to LLE algorithm, we will give a breif introduction of LLE before the detailed derivation of NPP.

3 Locally Linear Embedding (LLE)

To begin, suppose the data consist of N real-valued vectors $\mathbf{x}_i$, each of dimensionality D, sampled from a smooth underlying manifold. Provided the manifold is well-sampled, it is expected that each data point and its neighbors lie on or close to a locally linear patch of the manifold. We characterize the local geometry of these patches by linear coefficients W_{ij} that reconstruct each data point $\mathbf{x}_i$ from its K neighbors $\mathbf{x}_j$. Choose W_{ij} to minimize a cost function of squared reconstruction errors:

$$J_1(\mathbf{W}) = \sum_{i=1}^{N} ||\mathbf{x}_i - \sum_{j=1}^{K} W_{ij}\mathbf{x}_j||^2. \tag{3}$$

The reconstruction error can be minimized analytically using a Lagrange multiplier to enforce the constraint that (see [13] for details).

A basic idea behind LLE is that the same weights W_{ij} that reconstruct the ith data in D dimensions should also reconstruct its embedded manifold coordinates in d dimensions. Hence, each high-dimensional data $\mathbf{x}_i$ can be mapped to a low-dimensional vector $\mathbf{y}_i$ by minimizing the embedding cost function:

$$J_2(\mathbf{Y}) = \sum_{i=1}^{N} ||\mathbf{y}_i - \sum_{j=1}^{K} W_{ij}\mathbf{y}_j||^2. \tag{4}$$

$$= ||\mathbf{Y}(\mathbf{I} - \mathbf{W})||^2$$

$$= trace(\mathbf{Y}(\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W})^T\mathbf{Y}^T)$$

$$= trace(\mathbf{Y}\mathbf{M}\mathbf{Y}^T.$$

where

$$\mathbf{M} = (\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W})^T. \tag{5}$$

$$\mathbf{W} = \left[\mathbf{w}_1\ \mathbf{w}_2 \cdots \mathbf{w}_N\right].$$

$\mathbf{I}$ represents an identity matrix.

To make the optimization problem well posed, two constrains can be imposed to remove the translational and rotational degree of freedom:

$$\sum_{i=1}^{N} \mathbf{y}_i = 0 \quad or \quad \mathbf{Y1} = 0. \tag{6}$$

$$\frac{1}{N-1}\sum_{i=1}^{N} \mathbf{y}_i\mathbf{y}_i^T = \mathbf{I} \quad or \quad \frac{1}{N-1}\mathbf{Y}\mathbf{Y}^T = \mathbf{I}. \tag{7}$$

where $\mathbf{1}$ stands for a summing vector: $\mathbf{1}$=[1,1,...,1]^T

The constrained minimization can then be done using the method of Lagrange multipliers:

$$L(\mathbf{Y}) = \mathbf{Y}\mathbf{M}\mathbf{Y}^T + \lambda((N-1)\mathbf{I} - \mathbf{Y}\mathbf{Y}^T). \tag{8}$$

Setting the gradients with respect to $\mathbf{Y}$ to zero

$$\frac{\partial L}{\partial Y} = 0 \Rightarrow \quad 2\mathbf{M}\mathbf{Y}^T - 2\lambda\mathbf{Y}^T = 0. \tag{9}$$

leads to a symmetric eigenvalue problem:

$$\mathbf{M}\mathbf{Y}^T = \lambda\mathbf{Y}^T. \tag{10}$$

We can impose the first constraint above (for zero mean) by discarding the eigenvectors associated with eigenvalue 0 (free translation), and keeping the eigenvectors, $\mathbf{u}_i$, associated with the bottom d nonzero eigenvalues. These produce the d rows of the d-by-N output matrix $\mathbf{Y}$ [15]:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \ \mathbf{y}_2 \cdots \mathbf{y}_N \end{bmatrix}_{d \times N} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_d \end{bmatrix}_{d \times N} . \tag{11}$$

4 The Proposed Method (NPP)

4.1 Motivation

Though LLE possesses some favorable properties [13], its computational cost is expensive than most linear dimension reduction methods. Moreover, it cannot map a new testing point directly, which is referred to as out-of-sample problem as stated in section 1. This problem arises from the fact that the embedding of $\mathbf{y}_i$ is obtained in a way that does not explicitly involve the input point $\mathbf{x}_i$. The cost function J_2 in equation (4) depends merely on the weights W_{ij}. To establish a bridge across this gap, we plug equation (1) into the cost function J_2 and the resultant cost function is optimized. The process of NPP has been presented in section 2. In the next subsection its justification will be given. Because the first two steps of NPP are the same as LLE, only justification related to step 3 is presented.

4.2 Justification

Here we rewrite equation (1)

$$\mathbf{y}_i = \mathbf{A}^T \mathbf{x}_i \quad or \quad \mathbf{Y} = \mathbf{A}^T \mathbf{X}. \tag{12}$$

where

$$\mathbf{A} = [\mathbf{a}_0, \mathbf{a}_1, \cdots, \mathbf{a}_d]$$

We plug equation (12) into the cost function J_2:

$$J_2(\mathbf{Y}) = \sum_{i=1}^{N} ||\mathbf{y}_i - \sum_{j=1}^{K} W_{ij} \mathbf{y}_j||. \tag{13}$$

$$= trace(\mathbf{Y}\mathbf{M}\mathbf{Y}^T)$$

$$= trace((\mathbf{A}^T\mathbf{X})\mathbf{M}(\mathbf{A}^T\mathbf{X})^T)$$

$$= trace(\mathbf{A}^T(\mathbf{X}\mathbf{M}\mathbf{X}^T)\mathbf{A}).$$

The two constrains of equation (6) and (7) now becomes:

$$\mathbf{Y1} = 0 \Rightarrow \quad (\mathbf{A}^T\mathbf{X})\mathbf{1} = 0. \tag{14}$$

$$\frac{1}{N-1}\mathbf{Y}\mathbf{Y}^T = \mathbf{I} \quad \Rightarrow \frac{1}{N-1}\mathbf{A}^T\mathbf{X}(\mathbf{A}^T\mathbf{X})^T = \frac{1}{N-1}\mathbf{A}^T(\mathbf{X}\mathbf{X}^T)\mathbf{A} = \mathbf{I}. \tag{15}$$

The constrained minimization can then be done using the method of Lagrange multipliers:

$$\mathcal{L}(\mathbf{A}) = \mathbf{A}^T(\mathbf{X}\mathbf{M}\mathbf{X}^T)\mathbf{A} + \lambda((N-1)\mathbf{I} - \mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A}). \tag{16}$$

Setting the gradients with respect to $\mathbf{A}$ to zero we have

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}} = 0 \Rightarrow 2(\mathbf{X}\mathbf{M}\mathbf{X}^T)\mathbf{A}^T - 2\lambda\mathbf{X}\mathbf{X}^T\mathbf{A}^T.$$

By defining

$$\mathbf{L} = \mathbf{X}\mathbf{M}\mathbf{X}^T. \tag{17}$$

$$\mathbf{C} = \mathbf{X}\mathbf{X}^T. \tag{18}$$

we can rewrite equation (17) in the form of a generalized eigenvalue problem:

$$\mathbf{L}\mathbf{A}^T = \lambda\mathbf{C}\mathbf{A}^T. \tag{19}$$

If $\mathbf{C}$ is invertible, equation (20) can be transformed to a standard eigenvlaue problem:

$$(\mathbf{C}^{-1}\mathbf{L})\mathbf{A}^T = \lambda\mathbf{A}^T. \tag{20}$$

Once $\mathbf{A}$ is obtained by solving equation (20) or (21), $\mathbf{X}$ can be mapped to a low dimensional space by

$$\mathbf{Y} = \mathbf{A}^\mathbf{T}\mathbf{X}.$$

The constraint (14) can be imposed on by subtracting the mean vector of training set from a training vector or testing vector:

$$\mathbf{y}_i = \mathbf{A}^\mathbf{T}(\mathbf{x} - \overline{\mathbf{x}}). \tag{21}$$

where

$$\overline{\mathbf{x}} = \sum_{i=1}^{N} \mathbf{x}_i \tag{22}$$

5 Experimental Results

To demonstrate the effectiveness of the proposed method, NPP, experiments were conducted on data of the famous "swiss roll" and "s-curve" to compare with PCA.

The data set of 2000 points which are randomly chosen from the "swiss roll" (Fig.1 (a)) and "s-curve" (Fig.2 (a)) are shown in Fig. 1(b) and Fig. 2(b) respectively, which are used as training data. PCA seeks a direction onto which projected data has the maximum variance. Therefore, by PCA, data is projected onto a plane perpendicular to the paper plane and parallel to the vertical margin of the paper for our "swiss roll" and "s-curve" experiments. Examining Fig.1(d) and Fig.2 (d), one can find that projected points by PCA are blended. For example, in Fig.1(d) red points overlap largely with blue points and green points. In Fig.2(d) blue points overlap largely with yellow points.

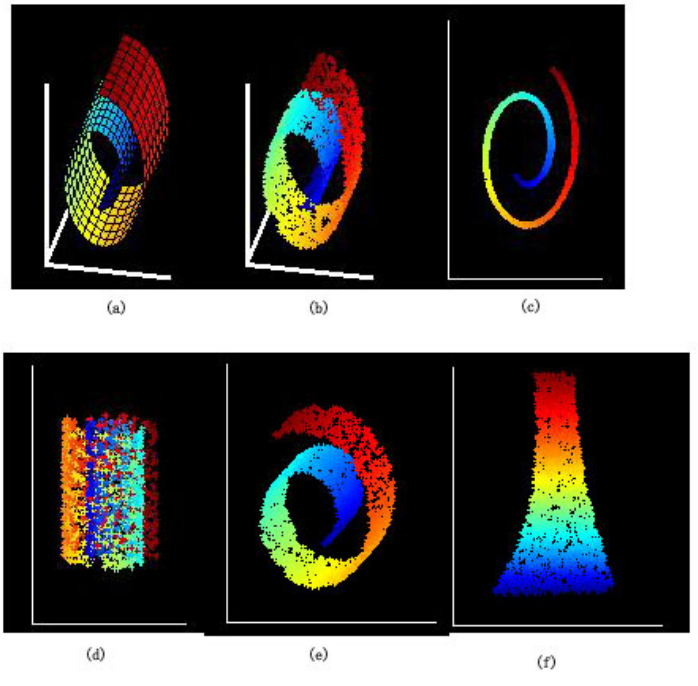

Fig. 1. (a) 3-D "swiss roll"; (b) 2000 points sampled from (a); (c) NPP representation; (d) PCA representationo

In contrast to PCA, our proposed NPP is able to search a direction projected onto which neighborhood relations are preserved along the curve of the manifold as possible. Therefore, by NPP, data are projected onto a plane parallel to the paper plane. Consequently, the projected data is show in fig 1(c) and fig 2(c).

Fig. 1(e) and fig. 2(e) show the results of LPP. From fig. 1(e), it is observed that LPP performs better than PCA. However, in fig. 1(e) blue points nearly connect to red points which is unfavorable. Fig. 2(e) is the result of LPP on "S-

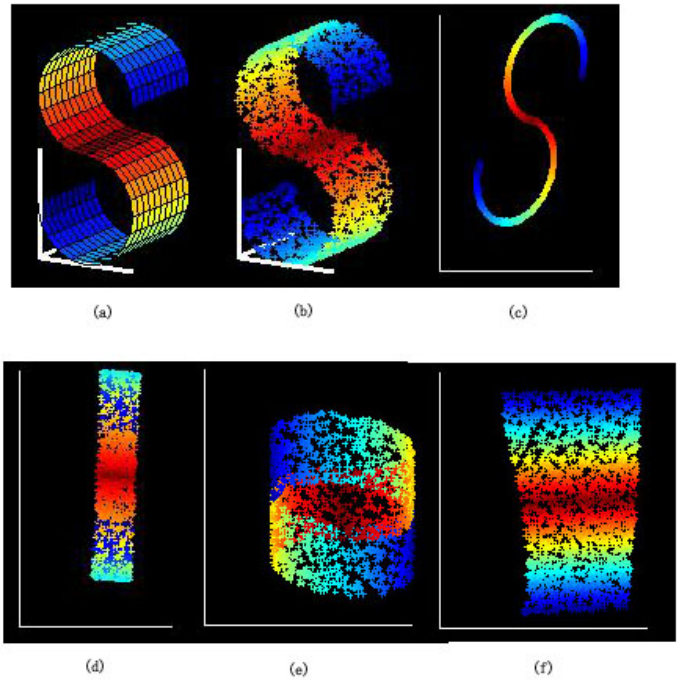

Fig. 2. (a) 3-D "s-curve"; (b) 2000 points sampled from (a); (c) NPP representation; (d) PCA representation

curve" data. We find that red points overlap with both blue and yellow points. Therefore, it is concluded that NPP outperforms LPP.

From fig. 1(c) and fig. 2(c), we can see that NPP can not always unfold the manifold as LLE (Fig. 1(f) and Fig. 2(f))can. Furthermore, many neighbors are collapsed into a single point in the low dimensional space.The reason is that NPP is a linear transform instead of nonlinear one like LLE. Nevertheless, the NPP has favorable properties against other linear transform methods such as PCA.

6 Conclusions and Future Work

By introducing a linear transform matrix into LLE algorithm, a novel unsupervised linear dimension reduction method , Neighborhood Preserving Projections (NPP),has been proposed in this paper. The linear transform matrix is obtained by optimizing a certain objective function which is similar to that of LLE. Hence, NPP inherits LLE's neighborhood property naturally. In contrast to traditional linear dimension reduction method, such as principal component analysis (PCA), the proposed method has good neighborhood-preserving property along the direction of the manifold.

Note that equation (20) is similar to equation (2) in [14] in some sense where another linear dimension reduction method, LPP was proposed. We will com-

pare NPP with LPP both in theory and in applications. Moreover, additional experiments will be conducted on real data.

Though NPP as well as LPP, LDA and PCA, because of their linear nature, might not outperform nonlinear LLE, Isomap and Laplacian Eigenmaps, NPP is a novel and useful linear dimension reduction method.

As future work, we will perform NPP in a large high-dimensional space by introducing a kernel [16-17]. It is believed that kernel NPP ,which is a nonlinear dimension reduction method, can outperform NPP.

References

1. Jain,A.K., Duin, R.P.W., Mao,J.: Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, (2000) 4-37
2. Belkin,M., Niyogi,P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation, Vol. 15, No. 6, (2003)1373-1396
3. R. O. Duda,R. O., Hart, P. E., Stork,D.G.: Pattern Classification. Wiley-Interscience, 2000
4. N.B. Peter,N.B., Joao,P.H., David,J.K.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, (1997)711-720
5. Scholkopf,B., Smola,A., Muller,K.R.: Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation, Vol 10, No. 5, (1998) 1299-1319
6. Baudat,G., Anouar,F.: Generalized Discriminant Analysis Using a Kernel Approach. Neural Computation, Vol. 12, (2000) 2385-2404
7. Roweis,S., Saul,L.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science, Vol. 290, No. 5500 , (2000) 2323-2326
8. Joshua,B., Tenenbaum, Langford, J.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science, Vol. 290, No. 5500, (2000) 2319-2323
9. Belkin,M., Niyogi,P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation, Vol. 5, No. 6, (2003)1373-1396
10. Bengio,Y., Paiement,J., Vincent,P., Dellallaeu,O., Roux,N.L, Quimet,M.: Out-of-sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. Neural Information Processing Systems, 2003
11. He. X., Yan,S., Hu,Y., Zhang,H.: Learning a Locality Preserving Subspace for Visual Recognition. In Proc. IEEE International Conference on Computer Vison, 2003
12. He,X., Yan. S, Hu,Y., Niyogi,P., Zhang,H.J.: Face Recognition Using Laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 3, (2005)328-340
13. Saul,L.K., Roweis,S.T.: Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds. Journal of Machine Learning Research, Vol. 4, (2003) 119-155
14. He,X., Niyogi,P.: Locality Preserving Projection. Technical Report TR-2002-09, Department of Computer Science, the University of Chicago
15. Gering,D.: Linear and Nonlinear Data Dimensionality Reduction. Technical Report, the Massachusettes Institute of Technology (2002)
16. John, S.T., Nello, C.: Kernel Methods for Pattern Analysis, Cambridge University Press, 2004
17. Ham, J., Lee, D.D., Mika, S., and Scholkopf, B.: A Kernel View of the Dimensionality Reduction of Manifold. Proc. Int. Conf. Machine Learning, (2004) 369-376