A Supervised Method to Chart Multiple Manifolds

Dan Zhang, Yangqiu Song, Qifeng Qiao, Zhenwei Shi, Changshui Zhang
State Key Laboratory on Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Automation,Tsinghua University, Beijing 100084, China
Email: {dan-zhang05, songyq99, qqf05}@mails.tsinghua.edu.cn, {shizhenwei, zcs}@mail.tsinghua.edu.cn

***Abstract*—The discovery of the manifolds has long been a hot topic in computer vision. In many practical problems, high-dimensional data poses a great obstacle to the researchers. But these data points are often sampled from several low-dimensional sub-manifolds. Therefore, charting the sub-manifolds in one coordinate system will help visualize them simultaneously. However, algorithms developed so far all have their own limitations in solving this problem. In this paper, we propose a new supervised method to capture multiple sub-manifolds.**

I. Introduction

Dimensionality Reduction is an interesting topic in the computer vision community. One of the main reasons is that real-world problems are often confronted with high-dimensional data points. In most cases, there exist low-dimensional structures, i.e. sub-manifolds, underlying these high dimensional data points. It is beneficial to design efficient dimensionality reduction algorithms to find these intrinsic manifolds. The main goal of our paper is to design such an algorithm. Unlike the classification task, whose main goal is to classify unlabeled data points accurately, we focus on how to arrange the labeled data in a space, whose dimension does not exceed three, to help the user visualize these sub-manifolds conveniently.

Most of the dimensionality reduction algorithms can be categorized into two groups: unsupervised and supervised. As for the unsupervised group, some of the previous works, such as Principle Component Analysis (PCA) [1], Locally Linear Embedding (LLE) [9], Isomap [10], Locality Preserving Projections(LPP) [7] and Laplacian Eigenmaps (LE) [6] have been developed to discover the intrinsic data structures, regardless of the labels of these data points. For the supervised ones, Linear Discriminant Analysis (LDA) [1], and its variants, such as Kernel LDA [3], null space LDA [4] and uncorrelated LDA [5], utilize the available discriminant information of the data points, and have shown a great success.

From another point of view, these dimensionality reduction algorithms can also be divided into linear and nonlinear ones. PCA, LPP and LDA, seeking linear maps from high-dimensional feature space to a lower one, are among the linear ones, while Kernel PCA, Kernel LDA, Laplacian Eigenmaps do not restrict this map to be linear, and therefore are categorized into the nonlinear group.

The authors in [14] present a nonlinear algorithm. However, they presume that all the data points lie on the same manifold, and do not take into account the discriminant information. Authors in [15] propose Supervised LLE, and can handle the situation when multiple-manifolds exist. But their main focus is on classification, rather than charting the sub-manifolds. As illustrated by the authors, their algorithm tends to lose the within-class structure, and is not suitable to treat the charting task. The same problem also exists in Kernel LDA. In [11], the authors propose a Supervised Isomap to perform this task. In their paper, the discriminant information is used to redefine the distances between data points. They aim to reduce the distances between data points sharing the same labels, as well as enlarge the distances between data points with different labels. Their motivation is quite reasonable. However, the redefinition of the distance function is too empirical.

In most cases, supervised algorithms can utilize more information than unsupervised methods and nonlinear algorithms have fewer mapping restrictions than the linear ones. Therefore, charting the manifolds through supervised nonlinear dimensionality reduction is a good choice. In this paper, we propose a new Supervised Nonlinear Dimensionality Reduction (SNDR) algorithm to discover the sub-manifolds for each category. SNDR is a nonlinear method, and does not restrict the mapping be linear. Unlike the unsupervised algorithms, SNDR utilizes the class labels of the input data points to guide the dimensionality reduction work.

The rest of the paper is organized as follows: In Section II, we will formulate the multiple mainfolds charting problem. A related method will be given in Section III. We will elaborate our proposed algorithm in Section IV. In Section V, the experimental results are presented. In the end, conclusions will be drawn in Section VI.

II. Problem Statement

We are given a set of n data points $\{\mathbf{x_i}, i = 1, 2, ...n\}$, which are sampled from k categories, as well as their corresponding labels $l(\mathbf{x_i}) \in \{1, 2, ..., k\}$. In most cases, for each category, a sub-manifold exists. The goal is to help chart, in a coordinate system, these sub-manifolds as faithfully as possible. In the context, we will use the term 'observed set' to denote the set of these data points and hence the 'observed data points' refers to the data points in the 'observed set'. The feature space after dimensionality reduction will be referred to as 'reduced feature space'.

The work was supported by the National Science Foundation of China (60475001, 60605002)

III. Linear Transformation

Exploring data structures from a global way, such as PCA, often gives undesirable results. So, some recent algorithms, such as [13], [12] and [8], have considered solving it from a local way.

For each data point $\mathbf{x_i}$, we first find its K nearest neighbors. Then these K nearest neighbors are split into two sets: the data points with the same label as $l(\mathbf{x_i})$ ($l(\mathbf{x_i})$ denotes the label of $\mathbf{x_i}$), i.e. $N_w(\mathbf{x_i})$ and the data points with labels different from $l(\mathbf{x_i})$, i.e. $N_b(\mathbf{x_i})$. Specifically,

$$\begin{array}{l} N_w(\mathbf{x_i}) = \{\mathbf{x_i^m} | l(\mathbf{x_i^m}) = l(\mathbf{x_i}), 1 \leq m \leq K\} \\ N_b(\mathbf{x_i}) = \{\mathbf{x_i^m} | l(\mathbf{x_i^m}) \neq l(\mathbf{x_i}), 1 \leq m \leq K\} \end{array}$$

Then, two graphs, the within-class graph G_w and the between-class graph G_b, are constructed, with each node representing a data point and the adjacency relationship between two data points representing an edge. The corresponding adjacency matrices, $\mathbf{W_w}$ and $\mathbf{W_b}$, are determined as follows:

$$W_{w,mn} = \begin{cases} 1, & if\, \mathbf{x_m} \in N_w(\mathbf{x_n})\, or\, \mathbf{x_n} \in N_w(\mathbf{x_m}) \\ 0, & otherwise \end{cases}$$

$$W_{b,mn} = \begin{cases} 1, & if\, \mathbf{x_m} \in N_b(\mathbf{x_n})\, or\, \mathbf{x_n} \in N_b(\mathbf{x_m}) \\ 0, & otherwise \end{cases}$$

Let $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ be a one-dimensional representation for the data points in the observed set. Suppose $\mathbf{a}$ is a projection vector. i.e. $\mathbf{y}^T = \mathbf{a}^T\mathbf{X}$, where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$. Then, the objective is to seek a projection vector $\mathbf{a}$ to minimize $\sum_{ij}(y_i - y_j)^2 W_{w,ij}$ and maximize $\sum_{ij}(y_i - y_j)^2 W_{b,ij}$, simultaneously. In this way, the local margins between different categories can be maximized. We find that,

$$\begin{array}{l} \frac{1}{2}\sum\limits_{ij}(y_i - y_j)^2 W_{w,ij} \\ = \frac{1}{2}\sum\limits_{ij}(\mathbf{a}^T\mathbf{x_i} - \mathbf{a}^T\mathbf{x_j})^2 W_{w,ij} \\ = \mathbf{a}^T\mathbf{X}\mathbf{D}_w\mathbf{X}^T\mathbf{a} - \mathbf{a}^T\mathbf{X}\mathbf{W}_w\mathbf{X}^T\mathbf{a} \end{array} \tag{1}$$

$$\begin{array}{l} \frac{1}{2}\sum\limits_{ij}(y_i - y_j)^2 W_{b,ij} \\ = \frac{1}{2}\sum\limits_{ij}(\mathbf{a}^T\mathbf{x_i} - \mathbf{a}^T\mathbf{x_j})^2 W_{b,ij} \\ = \mathbf{a}^T\mathbf{X}(\mathbf{D}_b - \mathbf{W}_b)\mathbf{X}^T\mathbf{a} \\ = \mathbf{a}^T\mathbf{X}\mathbf{L}_b\mathbf{X}^T\mathbf{a} \end{array} \tag{2}$$

$\mathbf{D_w}$ and $\mathbf{D_b}$ are diagonal matrices, with entries $D_{w,ii} = \sum_j W_{w,ij}$ and $D_{b,ii} = \sum_j W_{b,ij}$. The Laplacian matrix for the between-class graph G_b is $\mathbf{L_b} = \mathbf{D_b} - \mathbf{W_b}$. By restricting $\mathbf{a^T X D_w X^T a} = 1$, Eq. (1) turns to $1 - \mathbf{a^T X W_w X^T a}$. Also taking Eq. (2) into account, maximizing the local margins is equivalent to solving the following optimization problem:

$$\mathop{\arg\max}_{\substack{a \\ \mathbf{a^T X D_w X^T a}=1}} \quad \mathbf{a^T X}(\eta\mathbf{L_b} + (1-\eta)\mathbf{W_w})\mathbf{X^T a} \tag{3}$$

η is a trade-off parameter ($0 \leq \eta \leq 1$). This optimization problem can be solved efficiently by the generalized eigenvalue problem.

The objective function Eq. (3) explores the geometry of the data manifolds, characterizes both the geometrical and discriminant structures by utilizing the within-class and between-class graph. Experiments on face recognition give us an impressive result [8]. However, this algorithm restricts that the map should be linear and may sacrifice some manifold information when the dimension of the reduced feature space is very low. Therefore, this is not suitable for the visualization.

In this paper, we seek for a nonlinear map to visualize the sub-manifolds underlying the high dimensionality data points.

IV. The Proposed Algorithm

A. The Construction of The Adjacency Matrix

In SNDR, two graphs are constructed, i.e. the within-class graph $G_w^{'}$ and between-class graph $G_b^{'}$. However, the method how we construct these two graphs is different. For each data point $\mathbf{x_i}$ ($1 \leq i \leq n$), k_w nearest neighbors with labels $l(\mathbf{x_i})$, as well as k_b nearest neighbors with labels different from $l(\mathbf{x_i})$ are selected. k_w and k_b are two parameters that are determined beforehand. Then, the adjacency matrices $\mathbf{W_w^{'}}$ and $\mathbf{W_b^{'}}$ for $G_w^{'}$ and $G_b^{'}$ can be formulated as follows:

$$W_{w,mn}^{'} = \begin{cases} 1, & if\, \mathbf{x_m} \in N_w^{'}(\mathbf{x_n})\, or\, \mathbf{x_n} \in N_w^{'}(\mathbf{x_m}) \\ 0, & otherwise \end{cases} \tag{4}$$

$$W_{b,mn}^{'} = \begin{cases} 1, & if\, \mathbf{x_m} \in N_b^{'}(\mathbf{x_n})\, or\, \mathbf{x_n} \in N_b^{'}(\mathbf{x_m}) \\ 0, & otherwise \end{cases} \tag{5}$$

Here, $N_w^{'}(\mathbf{x_i})$ denotes the k_w nearest neighbors with the same label as $l(\mathbf{x_i})$, while $N_b^{'}(\mathbf{x_i})$ refers to the k_b nearest neighbors with labels different from $l(\mathbf{x_i})$. The between-class Laplacian matrix for the graph $G_b^{'}$ can be defined as $\mathbf{L_b^{'}} = \mathbf{D_b^{'}} - \mathbf{W_b^{'}}$. $\mathbf{D_b^{'}}$ and $\mathbf{D_w^{'}}$ are diagonal matrices, with diagonal entries $D_{b,ii}^{'} = \sum_j W_{b,ij}^{'}$, $D_{w,ii}^{'} = \sum_j W_{w,ij}^{'}$. Remind that in Section III, for $\mathbf{x_i}$, $N_w(\mathbf{x_i})$ and $N_b(\mathbf{x_i})$ are partitioned within its K nearest neighbors. By finding for each $\mathbf{x_i}$ the k_w nearest neighbors with labels $l(\mathbf{x_i})$, $D_{w,ii}^{'}$ will always be nonzero, and $\mathbf{D_w^{'}}$ can be nonsingular. This modification is reasonable, since it still reflects the local margin information between different categories, but from a different perspective. The reason why we require $\mathbf{D_w^{'}}$ to be nonsingular will be elaborated in the section IV-B.

B. Charting The Sub-manifolds on The Observed Data Set

Still assume that $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ is a one-dimensional representation for the observed data points. Our aim is to find a nonlinear projection that can faithfully preserve the pairwise relationship in the low dimensional space by minimizing the within-class scatter $\sum\limits_{ij}(y_i - y_j)^2 W_{w,ij}^{'}$ and maximizing the between-class scatter $\sum\limits_{ij}(y_i - y_j)^2 W_{b,ij}^{'}$, simultaneously. Inspired by Eq. (3), the optimization problem can be formulated as:

$$\mathop{\arg\max}_{\substack{\mathbf{y} \\ \mathbf{y D_w^{'} y^T}=1}} \quad \mathbf{y}(\eta\mathbf{L_b^{'}} + (1-\eta)\mathbf{W_w^{'}})\mathbf{y^T} \tag{6}$$

Here, we do not restrict $\mathbf{y}$ be a linear transformation $\mathbf{y} = \mathbf{a}^T\mathbf{X}$. Conversely, we directly estimate it from the optimization problem, which results in a non-linear feature space. This amounts to solving the generalized eigenvalue problem:

$$(\eta\mathbf{L}_{\mathbf{b}}^{'} + (1-\eta)\mathbf{W}_{\mathbf{w}}^{'})\mathbf{y^T} = \lambda\mathbf{D}_{\mathbf{w}}^{'}\mathbf{y^T} \tag{7}$$

η is the trade-off parameter, $0 \leq \eta \leq 1$. Still note that why we require $\mathbf{D}_{\mathbf{w}}^{'}$ to be nonsingular in Section 4.1. That's because if $\mathbf{D}_{\mathbf{w}}^{'}$ is singular, it would deteriorate the solution of Eq. (7). Let the column vector $\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_d}$ denote the solutions of equation (7), ordered according to eigenvalues $\lambda_1 > ... > \lambda_d$. the embedding will be given by:

$$\mathbf{Y} = (\mathbf{y_1}, \mathbf{y_2}, ...\mathbf{y_d})^T \tag{8}$$

C. The Whole Algorithm

For the observed set, the visualization procedure can be shown in Table I.

V. Experiments

In this section, we present experimental results on the MNIST handwritten digit test set [1]. This data set contains 10,000 28×28 pixels images, with 1000 for each category and 10 categories in total.

In our experiment, for each category, 100 images are randomly selected as the observed set to chart the manifolds. k_b and k_w are both set to 10, and the trade-off η is set to 0.1. The result of SNDR is shown in Fig. 1. As for comparison, the charting results of some representative algorithms such as LDA, Kernel LDA, Local Sensitive Discriminative Analysis(LSDA) [8] and Laplacian Eigenmaps are shown in Fig. 2. It can be seen that SNDR can separate different manifolds very well. From the several magnified manifolds, we can see that the multiple manifolds for these categories are well retained. For example, the lean degrees of the digit 1 and 7 increase with the x-coordinate.

LDA (Fig. 2(a))doesn't provide a good charting result because its underlying assumption is that the data distribution of each category is gaussian and LDA is itself a linear method. Although Kernel LDA provides a nonlinear supervised map, it maps the data points with the same labels onto the same point in the reduced feature space, as can be shown in Fig. 2(b), and therefore loses the inner sub-manifold struture for each category. LSDA(Fig. 2(c)) restricts the map to be linear and may lose some manifold information in such low dimensions. Therefore, it can not give a good charting result. Laplacian Eigenmaps (Fig. 2(d))is an unsupervised method and can not utilize the discriminative information, so it can not separate the manifold of each category well.

We provide a comparison result with Supervised Isomap [11] in Fig. 2(e). The observed set in this figure is the same as that in Fig. 1. But it is hard to distinguish the inner structure for each category. This is because, in Supervised Iso-map, the between-class distances for some categorie pairs are much longer than the within-class distances and therefore the within-class structures are concealed. However, the long distance between several categories doesn't mean a good separation for each category pairs, either. In fact, in Fig. 1, digit 2 overlaps digit 5 a little. But in Fig. 2(e), digit 1 and 3, digit 5 and 6 are strongly overlapped. This is because, in Supervised Isomap, the redefinition of the distance between two data points is too empirical, and may contradict the real distribution.

[1] http://yann.lecun.com/exdb/mnist/

VI. Conclusions

In this paper, by utilizing the local information, we propose a new algorithm-SNDR to chart the sub-manifolds of different categories under one coordinate system. Experimental results on Minist have shown its superior performance over several state-of-art algorithms. In the future, we will consider how to use this supervised nonlinear map to help improve the accuracy of classification tasks and how to depict the low-dimensional coordinates of the out-of-sample data points.

References

[1] Richard O. Duda, Peter E. Hart, and David G. Stork: Pattern Analysis. John Wiley and Sons, Inc.(2001)

[2] G. Baudat and F. Anouar: Generalized discriminant analysis using a kernel approach. Neural Computation,12(10): 2385-2404,(2000)

[3] S. Mika, G.Rätsch, B. Schölkopf, A. Smola, J. Weston, and K. R. Müller: Invariant feature extraction and classification in kernel spaces. NIPS, 12, (1999)

[4] Li-Fen Chen, Hong Yuan, Mark Liao etc: A new LDA-based face recognition system which can solve the small sample size problem. PR 33 (2000) 1713–1726

[5] Jin, Zhong; Yang, Jingyu; Hu, Zhongshan; Lou, Zhen: Face recognition based on the uncorrelated discriminant transformation. PR 34, No.7, 1405-1416 (2001).

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

[7] Xiaofei He; Partha Niyogi: Locality Preserving Projections.NIPS (2005)

[8] Cai, D., He, X., Zhou, K., Han, J., Bao, H.: Locality sensitive discriminant analysis. IJCAI (2007) 1713–1726

[9] Roweis, S. T., Lawrance, K. S.: Nonlinear dimensionality reduction by locally linear embedding. Science, 290 (2000), 2323-2326.

[10] Tenenbaum, J., de Silva, V., and Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science, 290(2000):2319-2323.

[11] Xin Geng, De-Chuan Zhan, and Zhi-Hua Zhou: Supervised Nonlinear Dimensionality Reduction for Visualization and Classification. IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, (2005), 35(6): 1098-1107.

[12] Mingrui Wu, Bernhard Schölkopf: Transductive Classification via Local Learning Regularization. AIStatics(2007)

[13] Masashi Sugiyama: Local Fisher discriminant analysis for supervised dimensionality reduction. ICML(2006)

[14] RMatthew Brand: Charting a manifold. NIPS(2003)

[15] D de Ridder, O Kouropteva, etc: Supervised locally linear embedding.ICANN/ICONIP(2003)

Input: data points $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}$, as well as their labels, where n denotes the number of all the observed data points. The desired dimension of the reduced feature space is d ($d \leq 3$); k_b, k_w, and the trade-off parameter η.
1. Construct the within-class matrix $\mathbf{W_w'}$ and the between-class matrix $\mathbf{W_b'}$, using (4) and (5), respectively.
2. Calculate the Laplacian matrix $\mathbf{L_b'} = \mathbf{D_b'} - \mathbf{W_b'}$
3. Solving the optimization problem (6). The optimal embedding is given by Eq. (8)
Ouput: Eq. (8) gives the optimal embedding for the observed set.

TABLE I
THE PROCEDURE TO FIND THE OPTIMAL EMBEDDING FOR THE OBSERVED DATA POINTS

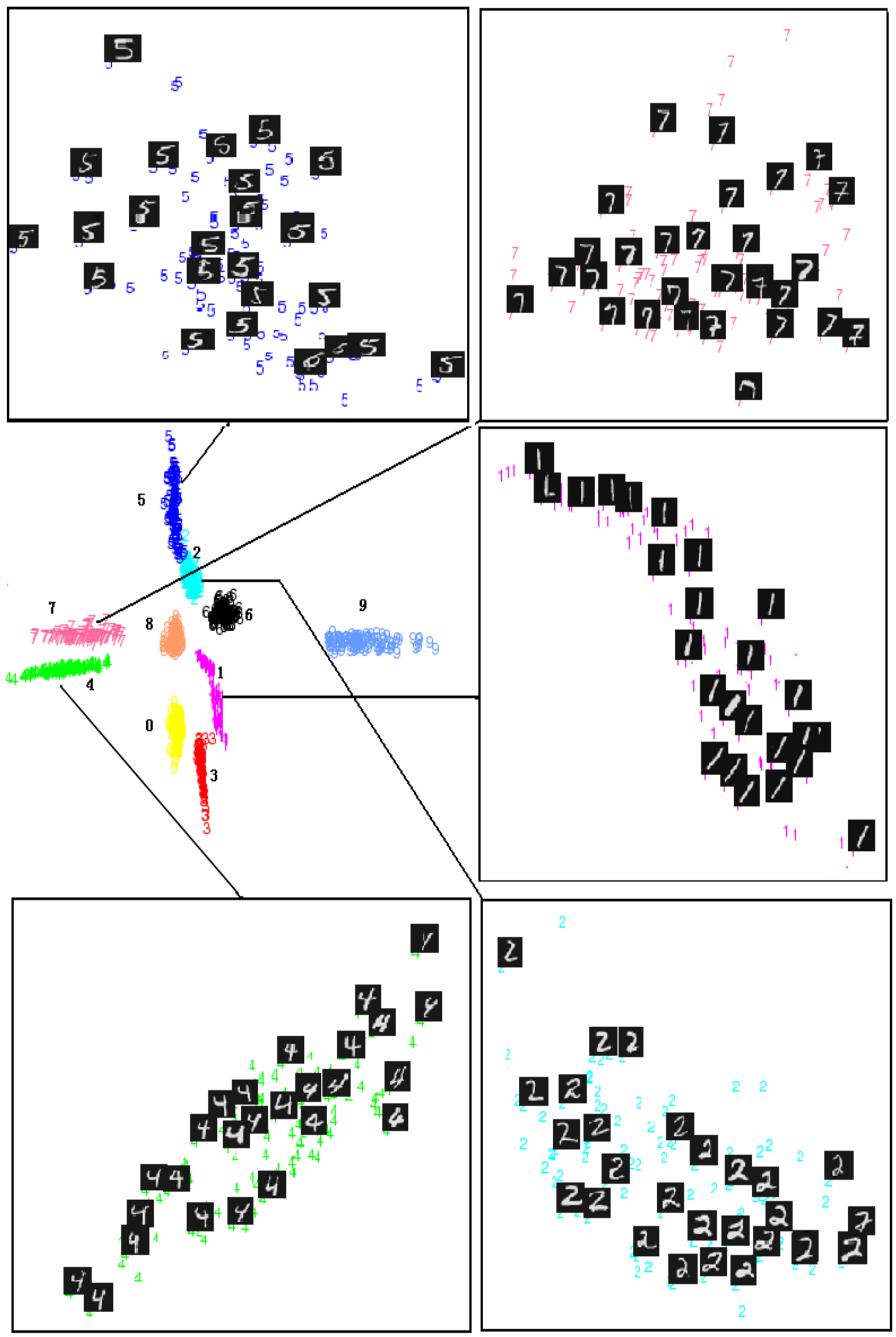
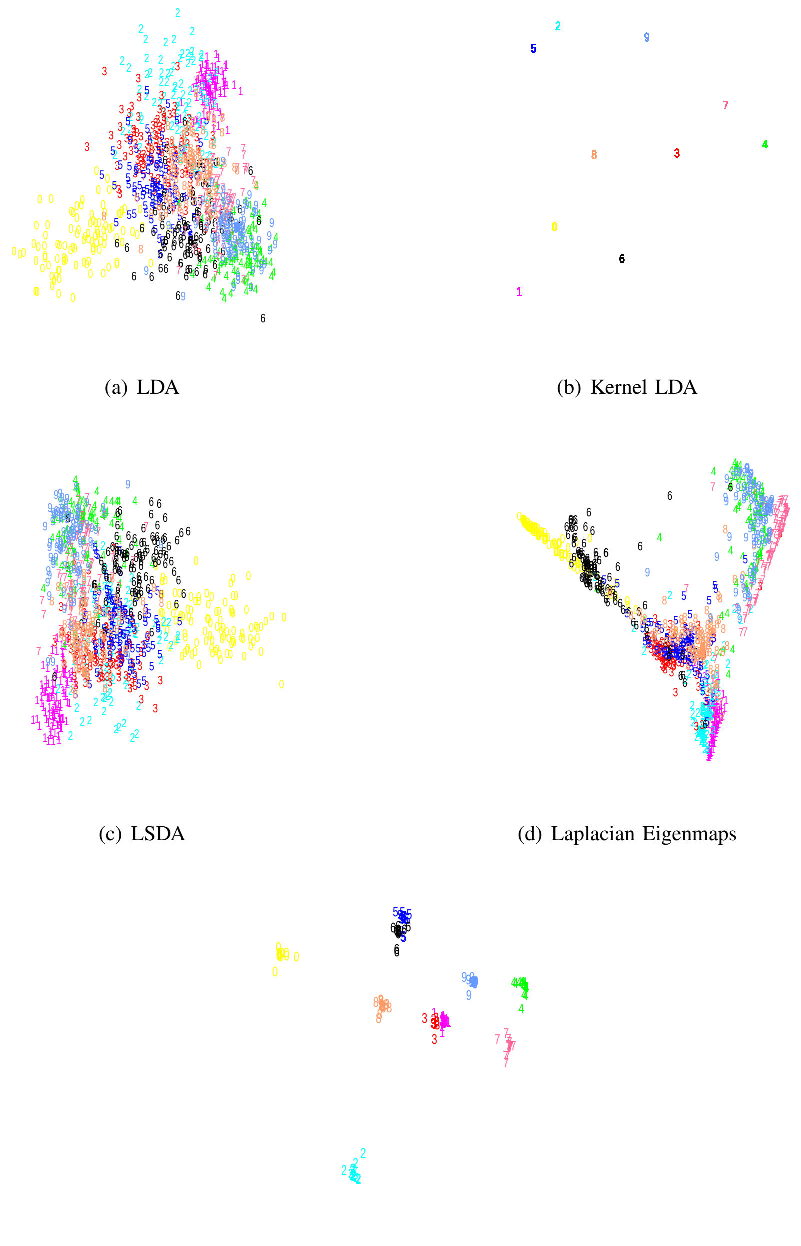


Fig. 1. The charting for the manifolds of all the categories. Each color represents a different category, and the sub-manifolds of several digits are magnified.



(a) LDA

(b) Kernel LDA

(c) LSDA

(d) Laplacian Eigenmaps

(e) Supervised Isomap

Fig. 2. The comparison charting result of LDA, Kernel LDA, LSDA, Laplacian Eigenmaps and Supervised Isomap