2012 IEEE 12th International Conference on Data Mining

Geodesic based semi-supervised multi-manifold feature extraction

Mingyu Fan[1], Xiaoqin Zhang[1,*], Zhouchen Lin[2], Zhongfei Zhang[3] and Hujun Bao[4]
[1]*Institute of Intelligent System and Decision, Wenzhou University, China*
Email: fanmingyu@wzu.edu.cn, zhangxiaoqinnan@gmail.com
[2]*Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing, China*
Email: zlin@pku.edu.cn
[3]*State University of New York, Binghamton, NY 13902, USA*
Email: zhongfei@cs.binghamton.edu
[4]*Department of Computer Science, Zhejiang University, Zhejiang, China*
Email: bao@cad.zju.edu.cn

***Abstract*—Manifold learning is an important feature extraction approach in data mining. This paper presents a new semi-supervised manifold learning algorithm, called Multi-Manifold Discriminative Analysis (Multi-MDA). The proposed method is designed to explore the discriminative information hidden in geodesic distances. The main contributions of the proposed method are: 1) we propose a semi-supervised graph construction method which can effectively capture the multiple manifolds structure of the data; 2) each data point is replaced with an associated feature vector whose elements are the graph distances from it to the other data points. Information of the nonlinear structure is contained in the feature vectors which are helpful for classification; 3) we propose a new semi-supervised linear dimension reduction method for feature vectors which introduces the class information into the manifold learning process and establishes an explicit dimension reduction mapping. Experiments on benchmark data sets are conducted to show the effectiveness of the proposed method.**

***Keywords*-Feature extraction; manifold learning; geodesic distance;**

I. Introduction

Feature extraction plays an important role in data analysis tasks. Classical linear methods, including Principal Component Analysis (PCA) [2], Linear Discriminant Analysis (LDA) [1] and Maximum Marginal Criterion (MMC) [3], are computationally efficient, globally optimal and in addition, converge asymptotically.

However, linear dimension reduction methods cannot discover the nonlinear structure hidden in the high dimensional data. As a new approach to nonlinear feature extraction, manifold learning becomes a hot topic. Two manifold learning algorithms, Isometric Feature Mapping (Isomap) [4] and Locally Linear Embedding (LLE) [5], were introduced in the same issue of *SCIENCE* in 2000. Since then, many new manifold learning algorithms have been proposed based on different motivations, such as Laplacian Eigenmaps (LE) [6], Hessian LLE [7], and Local Tangent Space Alignment (LTSA) [8]. To provide an explicit mapping from the input manifold to output embedding, many linear projection based algorithms have been proposed for manifold learning by assuming that there exists a linear dimension reduction projection. Linear manifold learning algorithms include the Locally Preserving Projections (LPP) [9], Orthogonal Neighborhood Preserving Projections (ONPP) [10], Discriminative Orthogonal Neighborhood-Preserving Projections (SDONPP) [11], and Graph embedding [12].

Geodesic, as an essential measurement for data distances, has been successfully used in manifold learning [4]. However, there are still limitations for most of the existing geodesic based manifold learning algorithms in classification. First, class information is rarely used in computing the geodesic distances between data points on manifolds. They are less effective when the data set is partially labeled or distributes on multiple manifolds, as is common in classification. Second, little efforts have been made to build an explicit dimension reduction mapping for extracting the discriminative information hidden in geodesic distances.

In view of this, we propose a new manifold learning algorithm for image classification, which has three new features.

1. A semi-supervised neighborhood graph construction method is introduced for data distributed on multiple manifolds.
2. We replace each data point with a feature vector whose elements are graph distances from the data point to the remaining data points.
3. To build robust explicit dimension reduction mappings, we propose a new semi-supervised linear dimension reduction method for the feature vectors.

Combining these three features, we propose a semi-supervised manifold learning algorithm, called Multi-Manifold Discriminative Analysis (Multi-MDA).

The rest of the paper is structured as follows. In Section II, the related feature extraction algorithms are reviewed. Our algorithm is described in Section III. In Section IV, experiments are reported on real world data sets to show the effectiveness of the proposed Multi-MDA algorithm. Finally, in Section V, we provide concluding remarks and suggestions for the future work.

*Corresponding author.

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.99

II. Related works

There are a lot of successful supervised feature extraction algorithms. LDA [1] is designed to find a linear projection A which maximizes the distances among the means of the classes and minimizes the distances among the points in the same class using the Fisher's criterion:

$$A^* = \arg\max_{A\in\mathbb{R}^{N\times d}} \frac{\text{tr}\left(A^T S_b A\right)}{\text{tr}\left(A^T S_w A\right)}, \tag{1}$$

where S_w, S_b denote the within-class scatter matrix and the between-class scatter matrix, respectively. Despite the success of LDA [1], it has been found to have intrinsic problems [13]: singularity of within-class scatter matrices and limited available projection directions. MMC [3] is based on the same intuition as LDA. The algorithm takes the following approach to find a dimension reduction mapping:

$$A^* = \arg\max_{A\in\mathbb{R}^{N\times d}, A^TA=I} tr\left(A^T(S_b - S_w)A\right). \tag{2}$$

Cai et al. [14] proposed a Semi-Supervised Discriminant Analysis (SSDA), which includes a manifold term which preserves local information of the unlabeled data points to improve the performance of the classification. Let $X = [x_1, \cdots, x_N]$ be the data matrix, L be the graph Laplacian matrix and $S_t = S_b + S_w$ be the total scatter matrix, The objective function for the algorithm is presented as

$$A^* = \arg\max_{A\in\mathbb{R}^{N\times d}} \frac{\text{tr}\left(A^T S_b A\right)}{\text{tr}\left(A^T(S_t + \alpha X L X^T)A\right)},$$

where $\alpha > 0$ is a given parameter. Independently, Song et al. [15] proposed a similar semi-supervised dimension reduction framework based on the LDA and the MMC algorithms.

By introducing the local metrics of semi-Reimannian manifold to describe the structures of classes, Wang et al. [16] proposed the Semi-Riemannian Discriminant Analysis (SRDA) algorithm for supervised dimension reduction. The Discriminative Multi-Manifold Analysis (DMMA) [17] for face recognition first segments each image into non-overlapping patches and then considers the patches of an image as a data manifold. Discriminant Analysis then is implemented on the data manifolds for feature extraction.

E-Isomap [20] is another supervised manifold learning algorithm which has three steps. The first and second steps of E-Isomap are the same as the classical I-somap algorithm. In the third step, a feature vector f_i is used to represent the original data point x_i, where $f_i = (d_G(x_i, x_1), \cdots, d_G(x_i, x_N))^T$, for $i = 1, \cdots, N$, where $d_G(x_i, x_j)$ denotes the graph distances between data points on the adjacent graph. The classical LDA [1] method is then applied to reduce the dimension of the extracted feature vectors $\{f_1, \cdots, f_N\}$.

There are some related multi-manifold analysis methods, such as [18], [19]. The projective mappings proposed in these methods work either on the original data points or on the kernel vectors. None of them utilized the discriminative information hidden in geodesic distances.

Table I
Notation

$\mathbb{R}^D$	The input space, D-dimensional Euclidean space
$\mathbb{R}^d$	The output space, d-dimensional Euclidean space
$\mathcal{X}$	$\mathcal{X} = \{x_1, \cdots, x_l, \cdots, x_{u+l}\}$ with $x_i \in \mathbb{R}^D$, the total data set. $\{x_i\}_{i=1}^{l}$ are labeled points, and $\{x_i\}_{i=l+1}^{u+l}$ are unlabeled points
N	$N = u + l$, the total number of data points in $\mathcal{X}$
$\mathcal{Y}$	$\mathcal{Y} = \{y_1, \cdots, y_l\}$ is the label set. $y_i \in \{1, 2, \cdots, C\}$ is the label of data point x_i
$\mathcal{F}$	$\mathcal{F} = \{f_1, \cdots, f_N\}$. $f_i = (d_G(x_i, x_1), \cdots, d_G(x_i, x_N))^T$ is the feature vector of x_i. If x_i is labeled as y_i, f_i will be labeled as y_i.
$\mathcal{X}^m$	$\mathcal{X}^m = \{x_1^m, \cdots, x_{N_m}^m\}$ with $x_j^m \in \mathcal{X}$. Data points of the m-th class, for $m = 1, \cdots, C$
N_m	The number of data points in the m-th class
X	$X = [x_1, \cdots, x_N] \in \mathbb{R}^{D\times N}$ is the input data matrix
Z	$Z = [z_1, z_2, \cdots, z_N] \in \mathbb{R}^{d\times N}$ is the output matrix. $z_i \in \mathbb{R}^d$ is the low-dimensional representation of data point f_i after dimension reduction
C	The number of classes that the data points belong to
$x_{n(i)}^m$	$(x_{n(i)}^m, x_{m(i)}^n)$ is the i-th shortest edge between $\mathcal{X}^m$ and $\mathcal{X}^n$, and $x_{n(i)}^m$ is an ending vertex from $\mathcal{X}^m$

III. Multi-manifold discriminant analysis

In order to avoid confusion, we give a list of the main notations used in this paper, as shown in Table I. Throughout this paper, all data points and the corresponding feature vectors are in the form of column vectors and denoted by lowercase. All sets are represented by capital curlicue letters. Matrices are denoted by normal capital letters.

In this section, we propose the Multi-MDA algorithm with the following three new features for dimension reduction.

1. In Section III-A, a new neighborhood graph construction method is proposed and used in our algorithm.
2. In Section III-A, each data point is replaced with a feature vector built by graph distances from it to the remaining data points.
3. In Section III-B, a new semi-supervised linear dimension reduction method is proposed to provide explicit dimension reduction mappings for feature vectors.

A. Multi-manifold modeling and feature vectors building

In this subsection, we consider the construction of a neighborhood graph for multi-manifold data and the replacement of the original data points with the feature vectors built from the graph distances.

Data lying on multiple manifolds are natural in the real world. For instance, in face recognition each person forms his or her own manifold in the feature space; in object tracking, moving subjects trace different trajectories which are low dimensional manifolds. Traditional graph construction methods, k-NN and ε-NN, cannot guarantee the connectivity of the graph for multi-manifold data. In order to provide a better graph, we propose a new graph construction method,

the k-Connectivity Graph (k-CG) method. The proposed method first builds a k-NN graph over the whole data in the semi-supervised manner, and then connects the adjacent graph components by the k shortest inter-edges. Details of the method are presented as bellow:

The k-CG method:

Step 1. *Construct the k-NN or ε-NN neighborhood graph in a semi-supervised manner.* Given an appropriate neighborhood size, define a graph G with the data points as the vertices by the means of k-NN or ε-NN method. For the training data with the class labels, each data point is connected to its nearest neighbors in the same class; for the training data without the class labels, each data point is connected to its nearest neighbors in the training set. Apparently, the nearest neighbor approach cannot guarantee a connected graph. At this step, several disconnected graph components may be obtained and each graph component can be considered as a data manifold. It is assumed that there are M data manifolds and the m-th data manifold contains $\mathcal{X}^m = \{x_1^m, \cdots, x_{N_m}^m\}$.

Step 2. *Compute the average number of the neighbors.* If the ε-NN method is applied to define the graph G at Step 1, the average number of the neighbors k needs to be computed. Let s_i be the number of the neighbors of x_i. The value of k is set to be the nearest integer to $\sum_{i=1}^{N} s_i/N$.

Step 3. *Connect the k nearest inter-manifold data points among manifolds.* Identify the k nearest inter-manifold data pairs, $\{(x_{n(i)}^m, x_{m(i)}^n), i = 1, \cdots, k\}$, between $\mathcal{X}^m$ and $\mathcal{X}^n$, and connect these data pairs by edges, for $m, n = 1, \cdots, M$. Then the k-CG graph is constructed on $\mathcal{X}$.

Remark 3.1: In Algorithm 3.1, we use the notion of the m-th class $\mathcal{X}^m = \{x_1^m, \cdots, x_{N_m}^m\}$ interchangeably to represent the m-th data manifold, although they are not necessarily the same. The reason for this is that the inconsistency between the discovered data manifolds and the true classes/clusters of the data set does not degrade the performance of the proposed Multi-MDA method.

Afterwards, the lengths of the shortest paths among the data points can be computed by the classical Floyd-Warshall's or Dijkstra's algorithm. Let $d_G(x_i, x_j)$ be the graph distance between data points x_i and x_j in the neighborhood graph and let the feature vector f_i be $f_i = (d_G(x_i, x_1), \cdots, d_G(x_i, x_N))^T$.

We propose an incremental graph construction procedure for the new data points. When a new data point x is obtained, we first identify its k-nearest or ε-nearest neighbors in $\mathcal{X}$, which are assumed to be $\{x_1, \cdots, x_k\}$. Then, we set the edges between x and these neighboring points. In this way, the lengths of the shortest paths of x to the data points in $\mathcal{X}$ can be computed by

$$d_G(x, x_i) = \min_{t=1,\cdots,k} \{\|x - x_t\| + d_G(x_t, x_i)\}, \quad \text{for} \quad i = 1, \cdots, N. \tag{3}$$

This procedure may be less accurate than implementing Floyd-Warshall's or Dijkstra's algorithm on the new data set $\mathcal{X} \cup \{x\}$. But it has a low computational complexity. Only $O((k+1)N)$ computational time is needed to include each new data point. Then the feature vector of x is obtained as $f = (d_G(x, x_1), \cdots, d_G(x, x_N))^T$.

B. Semi-supervised discriminant analysis for feature vectors

In this subsection, we propose the Semi-Supervised Discriminant Analysis (SSDA) method for the feature vectors. As the third step of the Multi-MDA algorithm, we apply it to the set of the feature vectors $\mathcal{F}$ instead of the original data set $\mathcal{X}$.

Let the within-class scatter matrix S_w and the new between-class scatter matrix $\bar{S}_b$ be

$$S_w = \sum_{m=1}^{C} \sum_{j=1}^{N_m} (f_j^m - c_m)(f_j^m - c_m)^T, \quad \text{and}$$

$$\bar{S}_b = \sum_{m=1}^{C} N_m (c_m - c)(c_m - c)^T + \sum_{m \neq n = 1}^{C} \sum_{j=1}^{k} (f_{n(j)}^m - f_{m(j)}^n)(f_{n(j)}^m - f_{m(j)}^n)^T,$$

where f_j^m is the feature vector of x_j^m, $f_{n(j)}^m$ is the j-th nearest feature vector of the m-th class to the n-th class, c_m is the mean vector of the feature vectors in the m-th class, and c is the mean vector of all the feature vectors. In comparison with the S_b used in LDA and MMC, the new scatter matrix enhances the robustness by maximizing distances among the margins of classes.

In the proposed SSDA algorithm, we include a term which preserves the locally linear reconstruction coefficients from the input data X to improve the performance of semi-supervised classification. Let the locally linear reconstruction coefficients be computed as

$$M_i = \arg\min \|x_i - \sum_j M_{ij} x_j\|_2^2.$$
$$s.t. \quad \sum_j M_{ij} = 1, \text{ for } i = 1, \cdots, N.$$

It is required that $M_{ij} = 0$ if there is no edge between x_i and x_j in the k-CG graph. The new term is

$$A^* = \arg\min_{A \in \mathbb{R}^{N \times d}} \sum_{i=1}^{N} \|A^T f_i - \sum_{j=1}^{N} M_{ij} A^T f_j\|^2 = \arg\min_{A \in \mathbb{R}^{N \times d}} \operatorname{tr}\left(A^T F L F^T A\right),$$

where $L = (I - M)(I - M)^T$, $F = [f_1, \cdots, f_N]$ is the feature matrix, $M = \{M_{ij}\}$ is the reconstruction coefficient matrix of order N, and I is the identity matrix. Then, let $0 \leq \beta \leq 1$, the SSDA algorithm is given as:

$$A^* = \arg\max_{A^T A = I} \left\{ \mathrm{tr} \left(A^T \left(\bar{S}_b - S_w - \beta F L F^T \right) A \right) \right\}. \quad (4)$$

The optimization problem (4) can be solved by computing the eigenvalues and the eigenvectors of the matrix $\bar{S}_b - S_w - \beta F L F^T$. More precisely, let $\{a_1, \cdots, a_d\}$ be the orthonormal eigenvectors corresponding to the top d eigenvalues. Then the required feature mapping is given by $A = [a_1, \cdots, a_d]$ and the final low-dimensional embedding is given as $Z = A^T F$.

C. The multi-manifold discriminant analysis algorithms

By combining the k-CG graph construction method, the replacement of the original data with the feature vectors, and the new SSDA method, our Multi-MDA algorithm has three steps. It should be noted that the proposed algorithm needs three parameters, β, neighborhood size k or ε, and dimension d.

Algorithm 3.2. (Multi-MDA)

Step 1. *Construct a connected graph.* Construct a neighborhood graph over $\mathcal{X}$ using the k-CG method. A weighted graph $G = \{\mathcal{X}, D\}$ is constructed, where $(D)_{ij} = \|x_i - x_j\|$ if x_i and x_j are connected by an edge and $(D)_{ij} = \infty$ otherwise.

Step 2. *Compute feature vectors.* Compute the lengths of pair-wise shortest paths on the graph by implementing the Floyd-Warshall's or Dijkstra's algorithm, and then replace x_i with the feature vector $f_i = [d_G(x_i, x_1), \cdots, d_G(x_i, x_N)]^T$, for $i = 1, \cdots, N$. The class label of f_i is set to be y_i, for $i = 1, \cdots, l$.

Step 3. *Compute d-dimensional embedding.* Apply the SSDA algorithm on the feature vectors. Let the computed feature mapping be A. Each data point x_i is represented by a low-dimensional vector $z_i = A^T f_i$.

Algorithm 3.2 presents the multi-MDA algorithm, which trains an explicit dimension reduction mapping for feature vectors of both the labeled and unlabeled data. When the low-dimensional representations of data points are obtained, one can train a classifier using the labeled low-dimensional representations.

In following, we propose the incremental multi-MDA for test data. Given an unlabeled sample x, incremental multi-MDA first maps it to low-dimensional space, and then applies the trained classifier to its low-dimensional representation.

Algorithm 3.3. (Online Multi-MDA)

Step 1. Compute pair-wise Euclidean distances $\|x - x_i\|$, for $i = 1, \cdots, N$. Identify the k-nearest neighbors or ε-nearest-neighbors of x, which are assumed as $\{x_1, \cdots, x_k\}$.

Step 2. Compute the lengths of shortest paths for x by Eq. (3). Then, the feature vector of x is given as $f = [d_G(x, x_1), \cdots, d_G(x, x_N)]^T$.

Step 3. Low dimensional representation of x for classification is computed as $z = A^T f$.

Table II
BRIEF DESCRIPTIONS OF THE COMPARED ALGORITHMS

Name	Description
PCA, LDA, SDONPP	Linear algorithms
LDA, SDONPP	Supervised linear algorithms
SS-KDA, E-Isomap, Multi-MDA	Supervised nonlinear algorithms
SS-KDA, SDONPP, Multi-MDA	Semi-supervised algorithms

IV. EXPERIMENTS

In this section, we compare the proposed Multi-MDA algorithm with representative dimension reduction algorithms, PCA [2], LDA [1], SS-KDA [14], SDONPP [11] and E-Isomap [20]. The properties of the compared algorithms are summarized in Table II.

A. Data set description

USPS [21] is a benchmark handwritten digit database, which contains 1100 samples for each class from '0' to '1'. The data set used in our experiments consists of 1100 samples from classes '0' and 1100 samples from class '1', which form a binary classification data set.

MIT CBCL [22] database contains 2429 face images and 4548 non-face images. In the experiment, we use a subset of this database, which comprises 1000 face and 1000 non-face images.

B. Experimental settings

For any data set $\mathcal{X}$, we randomly select α_1 percents of the data points in each class as the training set $\mathcal{X}_{train}$ and leave the remaining $100 - \alpha_1$ percents of data points as the test set $\mathcal{X}_{test}$. Similarly, we randomly label α_2 percents of data points in $\mathcal{X}_{train}$, where $0 \leq \alpha_1, \alpha_2 \leq 100$. Therefore, $\mathcal{X}_{train}$ is a partially labeled training set. Let the number of data points in $\mathcal{X}_{train}$ is N and the number of labeled training points is l.

In the Multi-MDA algorithm, we set $\beta = 0.01$ throughout all the experiments. On every data sets, the neighborhood sizes for SDONPP, E-Isomap and Multi-MDA algorithms, the kernel width for SS-KDA are chosen by cross-validation.

Classification on the unlabeled samples in $\mathcal{X}_{train}$ is conducted as follows:

Step 1. Train an explicit dimension reduction mapping using $\mathcal{X}_{train} = \{(x_1, y_1), \cdots, (x_l, y_l), \cdots, x_{u+l}\}$. We apply the algorithms on $\mathcal{X}_{train}$, which provide low-dimensional representations of $\mathcal{X}_{train}$ and explicit dimension reduction mappings for test data points. Assume that $\mathcal{Z} = \{(z_i, y_i), z_{l+j}, i = 1, \cdots, l, j =$

$1, \cdots, u\}$ be the low-dimensional representation of $\mathcal{X}_{train}$.

Step 2. Considering $\{(z_i, y_i), i = 1, \cdots, l\}$ as the training set, we implement the nearest neighbor classifier on the unlabeled set $\{z_{l+j}, , j = 1, \cdots, u\}$.

Classification on the test data set $\mathcal{X}_{test}$ is conducted as follows:

Step 1. Apply the trained dimension reduction mappings on $\mathcal{X}_{test}$, where the computed low dimensional representations are assumed as $\{z_j^{test}, j = 1, \cdots, T\}$.

Step 2. Considering $\{(z_i, y_i), i = 1, \cdots, l\}$ as the training set, we implement the nearest neighbor classifier on the test set $\{z_j^{test}, j = 1, \cdots, T\}$.

In the following, for each setting, we split the data sets and conduct experiments five times. The averaged accuracy of the compared algorithms are reported accordingly.

C. Experiments on the assessments of Multi-MDA algorithm

In this section, let $\alpha_1 = 90$ and $\alpha_2 = 10$, we compare the performances of our Multi-MDA, Multi-MDA without the k-CG graph construction method (using the k-NN method), SSDA and MMC on the USPS '0' and '1' classification data set and an extended CBCL data set, where the extended version of CBCL data set contains 1500 face samples and 1500 non-face samples.

The neighborhood size k is set as 15 for both the multi-MDA and k-NN based Multi-MDA. Recognition curves of the compared algorithms are shown in Fig. 1. As can be seen from Fig. 1, the feature vectors built by graph distances on k-CG graph are more discriminative than the feature vectors built on the k-NN graph.

D. Classification with different dimensions

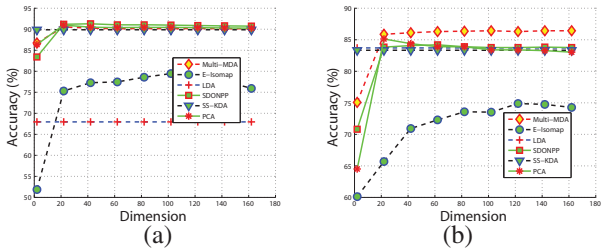

Figure 2. Recognition results of the compared algorithms with various dimension d on the unlabeled points of (a) USPS, (b) MIT CBCL.

Fixing $\alpha_1 = 90$ and $\alpha_2 = 10$, we evaluate the performances of compared algorithms with different dimensions, which varies from 2 to 200. In the experiments, the neighborhood sizes for E-Isomap and Multi-MDA are set as 12; the Gaussian kernel width for SS-KDA is tuned and chosen by cross validation, and the parameters for SDONPP algorithm we used are the same as used in their original paper. The results of the compared algorithms on unlabeled points are reported in Fig. 2, and the results of on test points are reported in Fig. 3.

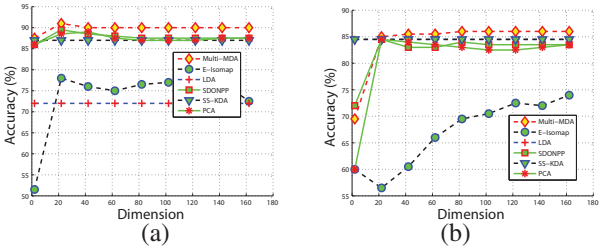

Figure 3. Recognition results of the compared algorithms with various dimension d on test points of (a) USPS, (b) MIT CBCL.

Table III
CLASSIFICATION WITH DIFFERENT PERCENTAGES OF LABELED TRAINING SAMPLES

	USPS			CBCL		
Unlabeled	α_2= 10	α_2=30	α_2=50	α_2=10	α_2=30	α_2=50
PCA	87.9	90.0	**91.7**	83.2	89.6	93.4
LDA	72.3	84.1	89.5	83.1	86.0	89.2
SS-KDA	78.3	89.1	89.6	85.1	90.7	94.2
SDONPP	87.9	89.8	91.7	83.5	89.6	93.6
E-Isomap	86.2	87.7	89	74.6	86.8	90.7
Multi-MDA	**88.7**	**90.3**	91.2	**85.6**	**91.2**	**95.1**
Test	α_2= 10	α_2=30	α_2=50	α_2=10	α_2=30	α_2=50
PCA	86.5	87.5	87.5	85.0	**92.5**	95.5
LDA	75.0	86.5	89.0	84.0	82.0	92.5
SS-KDA	**89.0**	**91.0**	88.5	85.3	87.0	95.0
SDONPP	88.5	88.0	87.5	85.0	92	95.5
E-Isomap	84.5	91.0	90.0	75.0	88.0	90.0
Multi-MDA	87.5	87.5	**90.0**	**86.0**	91.5	**96.0**

In Fig. 2(a), we can see that on the USPS data set, SDONPP has the best recognition rate. But there is not significant improvement when SDONPP is compared with Multi-MDA and SS-KDA methods. As can be seen from Fig. 2(b), Multi-MDA has significant advantages over compared algorithms on different dimensions on unlabeled data points from the CBCL data sets.

Fig. 3(a) indicates that Multi-MDA accomplishes the best recognition rate on the test set of USPS data set. In Fig. 3(b), we can see that the recognition rates of Multi-MDA algorithm are 2 or 3 percents higher than those of compared algorithms.

E. Classification with different portion of labeled training samples α_2

Fixing the training set portion $\alpha_1 = 90$ and target dimension $d = 100$, we test the compared algorithms with different percentages of labeled points in $\mathcal{X}_{train}$, i.e., α_2 changes from 10, 30, to 50. For each α_2, experiments are conducted five times and the averaged accuracy rates on both the unlabeled and test points are reported in Table III.

F. Discussion

According to the experiments being performed on the data sets, we make several observations:

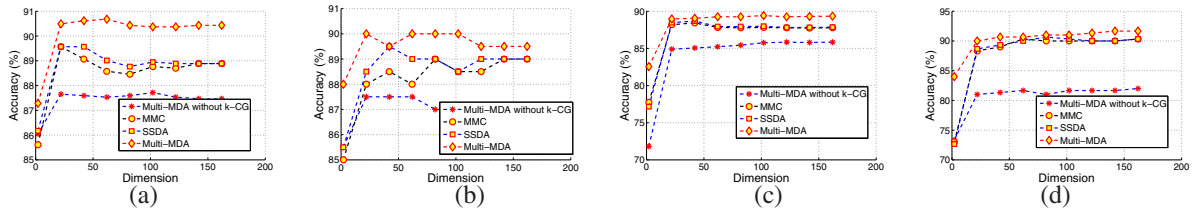

Figure 1. Recognition results of Multi-MDA, SSDA, MMC, and k-NN based Multi-MDA with various dimension d on (a) the unlabeled set of USPS, (b) the test set of USPS, (c) the unlabeled set of CBCL data set, (d) the test set of CBCL data set.

1. Kernel methods are sensitive to the number of training points. When the number of training points changes or the number of labeled points changes, the kernel width needs to be updated accordingly. Otherwise, kernel method gives poor results.
2. The results demonstrate that if the training set is large enough to characterizes the data distribution (manifold distribution), Multi-MDA outperforms the compared algorithms. Otherwise, the linear algorithms such as PCA and SDONPP are more effective than nonlinear algorithms.

V. Conclusion

In this paper, we have proposed a new multi-manifold learning algorithm. It combines semi-supervised multi-manifold modeling, nonlinear feature extraction and a new semi-supervised discriminant analysis method to achieve better performance in geodesic feature extraction and classification tasks. Experiments show that the proposed algorithm yields good results on projecting the data to a comparably low-dimensional space. Future works will be concentrated on exploring other nonlinear features from data.

Acknowledgement

This work is supported by NSFC (Grant Nos. 61100147, 61203241, 61272341 and 61231002), NSF of Zhejiang Province (Grant Nos. LQ12F03004 and LY12F03016), and the Open Project Program of the State Key Lab of CAD&CG (Grant No. A1202), Zhejiang University. ZZ is supported in part by US NSF (IIS-0812114, CCF-1017828), National Basic Research Program of China (2012CB316400), ZJU–Alibaba Joint Lab, and Zhejiang Engineering Center on Media Data Cloud Processing and Analysis.

References

[1] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 6, pp. 711-720, Jul. 1997.

[2] I. Jolliffe, "Principal Component Analysis", *Springer-Varlag, New York,* 1989.

[3] H. Li, T. Jianh, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. on Neural Networks*, vol. 17, no. 1, pp. 157 - 165, Jan. 2006.

[4] J. Tenenbaum, V. Sliva, and J. Landford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, December, 2000.

[5] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by local linear embedding," *Science*, vol. 290, no. 5500, pp. 2323-2326, December, 2000.

[6] M. Belkin, and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373-1396, Jun. 2003.

[7] D. Donoho, and C. Grimes, "Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data," *Proc. of National Academy of Sciences*, vol. 100, no. 10, pp. 5591-5596, May, 2003.

[8] Z. Y. Zhang and H. Y. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM Journal on Scientific Computing*, vol. 26, no. 1, pp. 313-338, 2005.

[9] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using laplacianfaces" *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328-340, March, 2005.

[10] E. Kokiopoulou, and Y. Saad, "Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2143-2156, December, 2007.

[11] T. Zhang, K. Huang, X. Li, J. Yang, and D. Tao, "Discriminative Orthogonal Neighborhood-Preserving Projections for Classification," *IEEE Trans on Systems, Man, and Cybernetics Part B*, vol. 40, no. 1, pp. 253-263, Feb. 2010.

[12] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40-51, Jan. 2007.

[13] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, H.-J. Zhang, "Multilinear Discriminant Analysis for Face Recognition," *IEEE Trans. Image Processing* vol. 16, no.1, pp. 212 - 220, Jan. 2007.

[14] D. Cai, X. He, and J. Han, " Semi-supervised Discriminant Analysis" *Proc. IEEE Intl Conf. Computer Vision,* pp. 1 - 7, 2007.

[15] Y. Song, F. Nie, C. Zhang, and S. Xiang, "A unified framework for semi-supervised dimensionality reduction," *Pattern Recognition*, vol. 41, no. 9, pp. 2789–2799, September, 2008.

[16] W. Zhang, Z. Lin, and X. Tang, "Learning Semi-Riemannian Metrics for Semisupervised Feature Extraction," *IEEE Transactions on Knowledge and Data Engineering* vol. 23, no.4, pp. 600-611, April, 2011.

[17] J. Lu, Y. Tan, and G. Wang, "Discriminative multi-manifold analysis for face recognition from a single training sample per person," *Proc. of IEEE International Conference on Computer Vision,* pp. 1943-1950, 2011.

[18] W. Yang, C. Sun, and L. Zhang, " A multi-manifold discriminant analysis method for image feature extraction," *Pattern Recognition,* vol. 44, no. 8, pp. 1649 - 1657, August, 2011.

[19] R. Wang and X. Chen, " Manifold Discriminant Analysis," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition,* pp. 429 - 436, June, 2009.

[20] M. Yang, "Extended isomap for pattern classification," *Proc. of 16th International Conference on Pattern Recognition,* no. 3, pp. 615-618, 2002.

[21] J. Hull, "A database for handwritten text recognition research" *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550-554, May, 1998.

[22] CBCL Face Database 1, MIT Center For Biological and Computation Learning, http://www.ai.mit.edu/projects/cbcl.