Nonlinear Discriminant Analysis on Embedded Manifold

Shuicheng Yan, Member, IEEE, Yuxiao Hu, Dong Xu, Hong-Jiang Zhang, Fellow, IEEE, Benyu Zhang, and Qiansheng Cheng

Abstract—Traditional manifold learning algorithms, such as ISOMAP, LLE, and Laplacian Eigenmap, mainly focus on uncovering the latent low-dimensional geometry structure of the training samples in an unsupervised manner where useful class information is ignored. Therefore, the derived low-dimensional representations are not necessarily optimal in discriminative capability. In this paper, we study the discriminant analysis problem by considering the nonlinear manifold structure of data space. To this end, firstly, a new clustering algorithm, called Intra-Cluster Balanced K-Means (ICBKM), is proposed to partition the samples into multiple clusters while ensuring that there are balanced samples for the classes within each cluster; approximately, each cluster can be considered as a local patch on the embedded manifold. Then, the local discriminative projections for different clusters are simultaneously calculated by optimizing the global Fisher Criterion based on the cluster weighted data representation. Compared with traditional linear/kernel discriminant analysis (KDA) algorithms, our proposed algorithm has the following characteristics: 1) it essentially is a KDA algorithm with specific geometry-adaptive-kernel tailored to the specific data structure, in contrast to traditional KDA in which the kernel is fixed and independent to the data set; 2) it is approximately a locally linear while globally nonlinear discriminant analyzer; 3) it does not need to store the original samples for computing the low-dimensional representation of a new data; and 4) it is computationally efficient compared with traditional KDA when the sample number is large.

The toy problem on artificial data demonstrates the effectiveness of our proposed algorithm in deriving discriminative representation for problems with nonlinear classification hyperplane. The face recognition experiments on YALE and CMU PIE databases show that our proposed algorithm significantly outperforms linear discriminant analysis (LDA) as well as Mixture LDA, and has higher accuracy than KDA with traditional kernels.

Index Terms—Kernel design, kernel machine, kernel selection, linear discriminant analysis (LDA), manifold learning, principal component analysis (PCA), subspace learning.

I. INTRODUCTION

Previous works on manifold learning [6], [10], [9], [12], [14] focus on uncovering the compact, low-dimensional representations of the observed high-dimensional unorganized data that lie on or nearly on a manifold in an unsupervised manner. According to the property of the mapping functions, these algorithms can be roughly classified into two types. One type is the algorithms without explicit mapping functions, such as ISOMAP [26], LLE [21], and Laplacian Eigenmap [3]. For this type, the low-dimensional representations are available often only for the sample data and by preserving certain local or global properties of the manifold structure. Another type is the algorithms with explicit mapping functions for the whole data space. Roweis et al. [25] proposed an algorithm that automatically aligns a mixture of local dimensionality reducers into a single global one. Brand [7] presented a similar work to merge local representations and construct a global nonlinear mapping function for the whole data space. He et al. [13] proposed the Locality Preserving Projections algorithm to linearly approximate the Laplacian Eigenmap algorithm. Bengio et al. [5] proposed the kernel explanations of ISOMAP, LLE, Laplacian Eigenmap and other spectral analysis algorithms to compute the low-dimensional representation for the out-of-sample data. All these algorithms are unsupervised and most of them are merely evaluated on simple toy problems. For real-world classification problems, such as face recognition, the unsupervised learning algorithms are unnecessarily optimal since they ignore the useful class label information of the sample data, which can be effectively applied to further improve classification performance.

In this paper, we study the problem to utilize the class label information of the training data for discriminant analysis by exploring the underlying nonlinear manifold structure of data space, and propose a novel algorithm for nonlinear discriminant analysis. This algorithm is motivated from the following observations. First, previous works on manifold learning focus on exploring the low-dimensional representations that best preserve some characteristics of a manifold, while the best representative features are not always the best discriminating ones for general classification task. Secondly, if we directly perform supervised learning based on the low-dimensional representations derived from previous manifold learning algorithms, there may be some important features for classification lost in the unsupervised dimensionality reduction step, which may be very strong in discriminative capability although weak in representative capability; and consequently it may degrade the performance of posterior discriminant analysis. Finally, linear discriminant analysis (LDA) can only well handle the linearly separable problem and is conducted directly on the Euclidean feature space without considering the possibly nonlinear manifold structure of the data space. Kernel discriminant analysis (KDA) [4], [16], [17], [29] utilizes the kernel trick to extend the LDA
for handling linearly inseparable classification problems; and owing to the capability to extract nonlinear discriminant features instead of linear ones, it has been widely studied and used in many applications. In the past years, many researches have been devoted to presenting more robust solutions to KDA and alleviating the ill-posed problem encountered by KDA. Mika et al. [19] proposed to add a small scalar matrix to the denominator matrix of objective function, consequently avoiding the singularity issue when using generalized eigenvalue decomposition method for computing the solution. To resolve the same issue, Baudat and Anouar [1] utilized the QR decomposition technique and Yang [30] proposed to use the principal component analysis (PCA) plus LDA strategy as in Fisher-faces [2] instead. When the sample number is large, KDA often suffers from the high computation cost and it requires to store all the training samples for computing the low-dimensional representation of a new data. Moreover, despite of the wide study of KDA, most works study KDA independent to the kernel itself, and the mapping function from the input feature space to the higher Hilbert space does not consider the manifold structure of the data, either. Following the above analysis, it is desired to propose an efficient algorithm for discriminant analysis by explicitly considering the possibly nonlinear manifold structure of data space.

In this work, the sample data of multiple classes are assumed lying on or nearly on a curved low-dimensional manifold; whereas it is often the case that the globally linearly inseparable manifold may be easily separable locally. The intuition of this work is to reside multiple local linear discriminant analyzers on the curved manifold, then merge these local analyzers into a global discriminant analyzer by optimizing the global Fisher Criterion. In summary, there are two subproblems to be solved: one is how to place the local discriminant analyzers on the curved manifold; the other is how to merge these local discriminant analyzers into a single global discriminant analyzer.

For the first subproblem, traditional methods like mixture factor analysis (MFA) [25] and K-means [11, 18] cannot be directly applied, since they cannot guarantee that there are balanced samples for all the classes within a cluster and the performance of the local discriminant analyzers will be degraded in the cases with unbalanced samples or even no sample for some classes. In this work, we formulate this task as a special supervised clustering problem and present a novel clustering approach that ensures there are balanced samples for the classes in a cluster, thus, called Intra-Cluster Balanced K-Means (ICBKM). Two extra terms penalizing the unbalance between clusters and within each cluster are imposed to the objective function of ICBKM; in addition to the reassignment optimization method utilized in K-means, a novel method called exchange optimization is proposed to minimize the sum of the intra-cluster variances, while keeps the penalizing terms constant.

Taking the advantage of the clustering results of ICBKM, the sample data are reset as multiple clusters, and local discriminant analysis can be conducted within each cluster. The traditional way to classify a new data using these local analyzers is to conduct the classification using the nearest local analyzer; therefore, these local analyzers are independent in both learning and inferring steps, which is not optimal since it is obvious that different analyzers can collaborate to improve the classification performance. To solve the second subproblem, we propose a novel approach in which the local analyzers are mutually dependent in both learning and inferring stages, and the optimal discriminative features for each cluster are computed simultaneously to optimize the global Fisher Criterion. First, PCA is conducted within each cluster; then the posterior probability of each cluster for a given data, i.e., \( p(c|x) \) can be obtained. The optimal discriminative features for each cluster are computed by maximizing the global Fisher criterion, i.e., maximizing the ratio of the weighted global inter-class and intra-class scatter, where the scatter are computed based on the \( p(c|x) \) weighted representations for the samples. In the inferring stage, the low-dimensional representation for new data is derived as the \( p(c|x) \) weighted sum of the projections from different clusters and the classification can be conducted using the Nearest Neighbor (NN) algorithm based on the derived low-dimensional representations. This algorithm can be understood from two different perspectives: 1) naively, it automatically merges local linear discriminant analyzers and conducts dimensionality reduction and classification in a single coordinate framework; and 2) in theory, it is essentially a special KDA algorithm with geometry-adaptive-kernel tailored to the special nonlinear manifold structure of sample data, in contrast to traditional KDA in which the kernel is predefined and dependent to sample data. Recently, Kim et al. [15] proposed a similar yet independently developed algorithm for locally LDA, which can be considered as the K-means-Daemon algorithm as discussed in our experiment part.

It is worthwhile to highlight some characteristics of our proposed algorithm here.

1) A supervised clustering algorithm is proposed to ensure that there are balanced samples for all classes in each cluster, which facilitates the following local discriminant analysis and successfully makes the local analyzers escape from the unbalanced problem existed in other clustering algorithm such as K-means.

2) The local discriminant analyzers collaborate to optimize the global Fisher Criterion in the learning stage, and the final classification is also based on the fused results from different local discriminant analyzers, which is superior to the Mixture LDAs that are learned and performed independently.

3) Our proposed algorithm is supervised and considers the nonlinear manifold structure of sample data; thus, it is natural to be superior to other unsupervised manifold learning algorithms, such as ISOMAP, LLE, Laplacian Eigenmap, and the recent proposed approach for out-of-sample extensions.

4) The kernel justification proves that our proposed algorithm is sound in theory. Moreover, the geometry-adaptive-kernel is tailored to the special nonlinear geometry structure of the sample data and superior to other fixed kernels in terms of classification performance.

The rest of the paper is structured as follow. The ICBKM clustering method and the locally linear while global nonlinear discriminant analysis algorithm based on the clustering result are introduced in Section II. In Section III, we present the kernel...
justification for the proposed algorithm. The toy problem on the artificial data and the real world face recognition experimental results compared with LDA, Mixture LDA and KDA on the YALE and CMU PIE database are illustrated in Section IV. Finally, we give the conclusion remarks in Section V.

II. NONLINEAR DISCRIMINANT ANALYSIS ON EMBEDDED MANIFOLD

Suppose $X = \{x_1, x_2, \ldots, x_N\}$ be a set of sample points that lie on or nearly on a low-dimensional manifold embedded in the high-dimensional observed space. For each sample $x_i \in \mathbb{R}^D$, a class label is given as $l_i \in \{1, 2, \ldots, L\}$. Most previous works on manifold learning are unsupervised; in this section, we show how to utilize the class information for nonlinear discriminant analysis by considering the nonlinear manifold structure of sample data. A continuous manifold may be considered as a combined set of a series of open sets, and when specific to the discrete sample data on it, they are the combination of a series of clusters. Furthermore, the globally linearly inseparable manifold may be easily separable within these local clusters. The above analysis motivates us to conduct local discriminant analysis within each local cluster, and then merge these local analyzers into a global discriminant analyzer. Following this idea, we first segment the sample data into multiple clusters.

The traditional clustering algorithms like $K$-means [18] and Normalized Cut [23] cannot be directly applied to the problem discussed here since there may be unbalanced samples for the classes in a cluster and even only a single class in some clusters, which makes the local discriminant analysis difficult or even impossible. To address this problem, we propose a novel clustering algorithm called ICBKM to ensure that the sample numbers for the classes in a cluster are balanced. Second, we search for local optimal features in each cluster by following the global Fisher Criterion in which we maximize the ratio of the cluster weighted inter- and intra-class scatters. In the following subsections, we will introduce the ICBKM and the global Fisher Criterion in detail, respectively.

A. Intra-Cluster Balanced $K$-Means Clustering (ICBKM)

$K$-Means clustering algorithm aims at putting more similar samples in the same cluster. It is unsupervised, thus, cannot guarantee that there are balanced samples in a cluster as desired. Compared to the traditional clustering algorithms, the clustering problem we concern here may have the following characteristics: 1) the class label for each sample is available, thus, the clustering process can be conducted in a supervised manner; 2) its purpose is not only to put the similar samples in the same cluster, but also to ensure that the samples for the classes in each cluster are balanced such that the local discriminant analysis can be conducted within each cluster. Cheung et al. [8] proposed a variation $K$-Means approach called cluster balance $K$-Means (CBKM), in which the concept cluster balance was proposed. However, CBKM only ensures that the sample number in each cluster is balanced and does not take into account the class label information and does not require the class balance. To provide a solution to this special clustering problem, we propose a novel supervised clustering approach, namely ICBKM here. ICBKM satisfies the requirement that there are balanced samples for classes in each cluster by adding an extra regularization term to constrain the sample number variation for the classes in each cluster. The cluster variance and the class variation constraints for the samples in each cluster collaborate to derive the clustering result with both cluster compactness and class balance.

Formally, the objective function of ICBKM is represented as

$$
\text{arg} \min_{K_i \in \{1, 2, \ldots, K\}} \sum_{i=1}^{N} \frac{|x_i - \bar{x}_{K_i}|^2}{\delta^2} + \alpha \sum_{k=1}^{K} |N^k - \bar{N}|^2 + \beta \sum_{k=1}^{K} \sum_{c=1}^{C} |N^k_c - \bar{N}^k_c|^2
$$

subject to: $c_{K_i} \geq 2$ ($k = 1, 2, \ldots, K$) (1)

where $K_i$ is the cluster index for the sample $x_i$; $\bar{x}_{K_i}$ is the average of the samples in cluster $k$; $N^k$ is the sample number in cluster $k$; $\bar{N}$ is the average sample number for each cluster; $N^k_c$ is the sample number of the $c$-th class in cluster $k$; $\bar{N}^k_c$ is the average sample number for each class in cluster $k$; $c_{K_i}$ is the class number in cluster $k$; $\delta$ is the standard deviation of the sample data; and $\alpha$ and $\beta$ are the weighting coefficients for the last two terms.

In the objective function of ICBKM, the first term minimizes cluster variance so as to make clusters compact; the second term makes sample numbers of different clusters balanced; and the third term is to ensure that each class has similar number of samples in a given cluster. The objective function is not trivial and we cannot obtain the closed form solution directly. Here, we apply an iterative procedure as traditional $K$-Means does to optimize the objective function. The pseudo-code is listed in Fig. 1.

In this procedure, the first optimization step is called Assignment Optimization, which is similar to $K$-Means. For each sample, we first check whether there exist at least two classes of other samples within the cluster to which this sample belongs. If yes, this sample is reassigned to each of the other clusters, and is finally put to the one resulting in the minimal objective function value. We can see for each reassignment, only the terms related to the original cluster and the assigned one are required to be updated, hence the process to find the target cluster for each sample is fast.

Unlike $K$-Means, an extra step called exchange optimization is proposed in ICBKM. It is designed to solve the problem that traditional assignment optimization approach cannot well handle. That is, when a sample is reassigned to another cluster, the in-cluster variance is reduced but the other two terms increase; In the exchange optimization step, the first term is optimized while the last two terms remain constant, thus, the optimization conflict between the first term and last two terms is avoided. In the exchange optimization step, the exchange result in the issue of only one class of samples within a cluster will be canceled; and to avoid the heavy computational cost, the exchange pair is constrained to the ones near the cluster margins as denoted in the Fig. 1. These two steps are iteratively to minimize the objective function until a local optimum is obtained.

1) Discussion: The proposed ICBKM is based on the assumption that it is possible to have relatively balanced numbers
ICBKM: Given the class label set $S = \{1, 2, \ldots, l\}$, the data set $X$, the class label $l_i$ for each sample $x_i$ in $X$ and the final cluster number $K$. 
1. Initialization: Compute the standard deviation $\delta$ of the data set $X$; randomly select $\bar{x_1}, \bar{x_2}, \ldots, \bar{x_K}$ as the initial cluster centers, then assign each $x_i$ to the cluster whose center is the nearest to $x_i$. For each cluster with only one class of samples, randomly select another one of different class to this cluster.
2. Reset Cluster Centers: For each cluster $C^i$, reset the center as the average of all the samples assigned to cluster $C^i$.
3. Assignment Optimization: For each $x_i \in X$, assign it to the cluster that makes the objective function minimal and the result satisfies the constraint in (1).
4. Exchange Optimization: For each cluster, exchange the cluster labels for the sample in $C^i$ that is the farthest to cluster center and the sample of the same class $e \in C^i$ that is the nearest to cluster center. If no improvement or the requirement of multiple classes for each cluster unsatisfied, keep the previous labels.
5. Evaluation: If current step has no improvement, return the final clustering results $\{C^1, C^2, \ldots, C^K\}$; else, go step 2.

Fig. 1. Procedure for ICBKM.

of samples for the classes within a cluster, so it may be improper for certain extreme cases, such as the two-class problem with extremely unbalanced training samples, and the multi-class problem in which the sample number of one class is much larger than that of all other classes. For most applications such as face recognition, the training sample numbers of different classes are relatively balanced; hence ICBKM can be well applied in these applications.

B. Global Discriminant Analysis by Merging Local Analyzers

Taking the advantage of the proposed ICBKM approach, the sample data are separated into multiple clusters with balanced samples for different classes. The traditional way to utilize these clustering results is to conduct discriminant analysis within each cluster, then determine the class label of a new data according to the classification result from its nearest discriminant analyzer. In this way, the local analyzers are independent and the final classification uses only part of the available information. We propose to utilize the global Fisher criterion to combine the local discriminant analyzers into a globally nonlinear discriminant analyzer. The global Fisher criterion maximizes the ratio of the class weighed inter- and intra-cluster scatters. The algorithm has three steps and they are introduced in detail as follows.

1) PCA Projections: In each cluster, PCA is conducted for dimensionality reduction; moreover, like in Fisher-face [2], PCA step can prevent the algorithm from suffering from the singular problem when the sample number is less than the feature number. In our experiments, we retain 98% of the energy in the sense of reconstruction error. Thus, in each cluster, each data $x_i \in \chi$ projected into a low-dimensionality feature space as

$$z^k(x_i) = (W^k_{\text{pca}})^T(x_i - \bar{x}^k) \quad k = 1, \ldots, K$$

where $W^k_{\text{pca}} \in \mathbb{R}^{D \times n_k}$ is the $n_k$ leading eigenvectors of the covariance matrix from the sample data belonging to cluster $k$, and $\bar{x}^k(x_i) \in \mathbb{R}^{n_k}$. The conditional probability of cluster $k$ for a given data $x$, $p(C^k|x)$ also simplified as $p^k(x)$, can be obtained using a simple formulation as in [22]

$$p^k(x) = p(C^k|x) = \frac{p(x,C^k)}{\sum_{j=1}^{K} p(x,C^j)}$$

where $p(x,C^k) = \exp[-\alpha^k(x)]$ and $\alpha^k(x)$ is the activity signal of the data for cluster $k$. In our experiments, $\alpha^k(x)$ is set as the Mahalanobis Distance [20] of the data in the PCA space of cluster $k$.

2) Nonlinear Dimensionality Reduction by Optimizing Global Fisher Criterion: LDA algorithm cannot well handle the nonlinear classification problem; and the KDA may suffer from high computational cost in the classification stage when the sample number is too large, it is desirable to propose a novel efficient discriminant analysis algorithm to conduct nonlinear discriminant analysis in consideration of nonlinear manifold structure of the sample data. As previously described, each sample $x_i \in \chi$ can be represented as a low-dimensional vector $z^k(x_i)$ in cluster $k$. Denote the optimal feature directions for dimensionality reduction within cluster $k$ as $W_k^f \in \mathbb{R}^{k \times n_k}$ and the translations as $W_k^b \in \mathbb{R}^{1 \times n_k}$ in cluster $k$, and $W_k = (W_k^f)^T, (W_k^b)^T \in \mathbb{R}^{(n_k+1) \times n_k}$, then the optimal low-dimensional representation of $x_i$ can be represented as the weighted sum of the projections from different clusters as

$$\Gamma(x_i) = \sum_{k=1}^{K} p^k(x_i)(W_k^f)^T(W_k^b)^T + W_0^b)$$

where

$$z^T(x_i) = (p^k(x_i)z^k(x_i))^T, p^k(x_i) = \sum_{k=1}^{K} (n_k + 1)$$

\ldots, p^k(x_i)z^k(x_i)^T, p^k(x_i)) \in \mathbb{R}^{(n_k+1)}$$
and matrix  

\[ W = ((W^1)^T,(W^2)^T,\ldots,(W^K)^T)^T \in \mathbb{R}^{\sum_{k=1}^{K}(n_k+1) \times n}. \]

The global intra-class and inter-class scatters can be represented as

\[ S_w = \frac{1}{N} \sum_{i=1}^{N} (\Gamma(x_i) - \bar{\Gamma}) (\Gamma(x_i) - \bar{\Gamma})^T \]

\[ = W^T \sum_{i=1}^{N} (z(x_i) - \bar{z}) (z(x_i) - \bar{z})^T W \]

\[ S_b = \sum_{l=1}^{L} N_l (\Gamma_l - \bar{\Gamma}) (\Gamma_l - \bar{\Gamma})^T \]

\[ = W^T \sum_{l=1}^{L} N_l (\bar{z} - \bar{z}) (\bar{z} - \bar{z})^T W \]

\[ = W^T M_b W \]  

where

\[ M_w = \sum_{i=1}^{N} (z(x_i) - \bar{z}) (z(x_i) - \bar{z})^T \]

\[ M_b = \sum_{l=1}^{L} N_l (\bar{z} - \bar{z}) (\bar{z} - \bar{z})^T \]  

and \( \bar{\Gamma} \) is the mean of \( \Gamma(x_i) \) where \( x_i \) belongs to class \( l \) and \( \bar{z} \) is the mean of \( z(x_i) \) where \( x_i \) belongs to class \( l \) and \( \bar{z} = (1/N) \sum_{i=1}^{N} z(x_i) \). The global Fisher Criterion is to maximize the cluster weighted inter-class scatter while minimize the cluster weighted intra-class scatter with respect to \( W \), i.e.,

\[ W^* = \arg \max \frac{W^T M_b W}{W^T M_w W}. \]

It has closed form solution and can be directly computed out by solving the generalized eigenvalue decomposition problem [9] as

\[ M_b w_i = \lambda_i M_w w_i, \quad W = [w_1, w_2, \ldots, w_n]. \]

3) Nonlinear Dimensionality Reduction for Classification:

For a new data \( x \), the posterior probabilities for each cluster can be computed according to (3) and its low-dimensional representation is obtained via the following nonlinear mapping function in term of the derived local optimal discriminant features for each cluster

\[ M(x) = \sum_{k=1}^{K} p(C^k|x)(W^k x^k T W^k p(x - \bar{x}^k)^T + W^k_0). \]

It is an explicit nonlinear mapping function from the original data space to the low-dimensional space. The consequent classification can be conducted based on these low-dimensional representations by using the traditional approaches like NN or Nearest Feature Line (NFL). In all our experiments, we used the NN method for final classification for simplicity.

III. KERNEL JUSTIFICATION AND DISCUSSIONS

The two approaches described above are integrated into an algorithm for conducting discriminant analysis on embedded manifold (Daemon). That is, Daemon consists of two steps: 1) separate the sample set into a set of class balanced clusters; and 2) merge the local discriminant analyzers into a single global nonlinear discriminant analyzer by following the global Fisher Criterion. It supervises the local analyzers and automatically decides the responsibility of each analyzer, which is somewhat like the background procedure named Daemon in UNIX system, thus, called Daemon. As described above, the intuition of Daemon is to merge the local discriminant analyzers into a unified coordinate framework. In this section, we analyze Daemon from a different point of view and justify that Daemon is a specific KDA algorithm, in which the kernel is data dependent and geometry adaptive, unlike traditional kernel machines that are independent to the data set to be analyzed. Then, we will discuss the relationship between Daemon and Mixture LDA as well as LLE variant [25].

A. Kernel Justification

Daemon follows the global Fisher Criterion in the learning stage and essentially is a discriminant analysis algorithm [1]; yet, as shown in Section II, it is nonlinear and is adaptive to the geometry structure of the data set. As shown in (4), Daemon can be considered a process in which the training sample, denoted as \( x \), is mapped into another data space, denoted as \( z(x) \), then, LDA are conducted on this new feature space by considering \( z(x) \) as the new object to be analyzed. Therefore, Daemon can be considered a special KDA algorithm with the kernel defined as

\[ k(x,y) = \phi(x) \cdot \phi(y) \]

where

\[ \phi(x) = z(x) = (p^1(x)z^1(x)^T, p^2(x)z^2(x)^T, \ldots, p^K(x)z^K(x)^T)^T \]

as in (4). This kernel has the following characteristics: 1) it has explicit mapping function from the input space to another feature space as the polynomial kernel does; and 2) it is dependent on the training samples and adaptive to the geometry structure of the sample data, which directly leads to its superiority over traditional kernels that are fixed as independent to the data set. Moreover, it is a special Marginalized kernel as introduced below.

Marginalized Kernel [27] defines a kernel between two visible variables \( x, x' \) with an extra hidden variable \( h \in \mathcal{H} \), where \( \mathcal{H} \) is a finite set. Let \( \xi = (x, h) \), the marginalized kernel is derived by taking the expectation with respect to the hidden variable as

\[ K(x,x') = \sum_{h \in \mathcal{H}} \sum_{h' \in \mathcal{H}} p(h|x)p(h'|x') K_{\xi}(\xi,\xi') \]

in which the joint kernel \( K_{\xi}(\xi,\xi') \) is designed between two combined variables \( \xi \) and \( \xi' \).
In Daemon, the kernel $k(x, y) = \phi(x) \cdot \phi(y)$ can be rewritten as

$$K(x, x') = \phi(x) \cdot \phi(x') = \sum_k p^k(x)p^k(x')(x^k \cdot x'^k + 1) = \sum_k p(C^k|x)p(C^k|x')(x - x^k)^T W_{pca}^k W_{pca}^k(x' - x^k) + 1.$$  

Thus, it can be considered as a special marginalized kernel with $h = k, \mathcal{H} = \{1, 2, \ldots, K\}$ and the joint kernels are defined as

$$K_\xi\xi' = (x - \overline{x}^h)^T W_{\text{pca}}^h W_{\text{pca}}^h (x' - \overline{x}^h) + 1. \quad (12)$$

B. Discussions

In the following, we discuss the connections between Daemon and other algorithms, including Mixture LDA, LLE, and its variant [25]. Then, we analyze the computational complexity of Daemon.

1) Relationship With Mixture LDA: A nature extension of LDA to handle nonlinear classification problem is to train specific LDA within each cluster, and we call this strategy as Mixture LDA here. Similar to Daemon, Mixture LDA algorithm also separates the sample data into multiple clusters and conducts local discriminant analysis. However, Daemon is different from Mixture LDA in both learning and inferring steps.

1) In the training stage, for Mixture LDA, the local discriminant features are learned directly from the samples within the corresponding cluster by optimizing the local Fisher Criterion; hence, these local discriminant analyzers are independent. For Daemon, the local discriminant features for different clusters are calculated simultaneously by optimizing the global Fisher Criterion. Daemon aims at aligning the local discriminant features with a single global coordinate framework.

2) In the inferring stage, Mixture LDA classifies a new data only by the nearest local discriminant analyzer thus, can only use part of the available information for classification. In Daemon, the low-dimensional representation of a new data is the weighted sum of the projections from all the local discriminant analyzers and the classification is based on all the sample data. Fig. 2 demonstrates the difference between Daemon and Mixture LDA, and notice that, for Daemon, the final most similar sample of a new data may not exist in the nearest cluster.

2) Relationship With LLE and its Variant: LLE [21] is a manifold learning algorithm that maps the input data to a lower dimensional feature space by preserving the relationship among the neighboring points. First, the sparse local reconstruction coefficient matrix $M$ is computed, such that $\sum_{j \in N(i)} M_{ij} = 1$ where the set $N(i)$ is the index set of the $k$ nearest neighbors of the sample $x_i$ and $\sum_{j \in N(i)} \|x_i - M_{ij}x_j\|^2$ is minimized; then, the low-dimensional representation $y_i$ for sample $x_i$ is obtained by minimizing $\sum_i \sum_{j \in N(i)} \|y_i - M_{ij}y_j\|^2$.

Roweis [25] proposed a procedure to align disparate local linear presentations into a global coherent coordinate system by preserving the relationship between neighboring points as LLE does. It has some similar ideas as Daemon, but they are intrinsically different in two-folds: 1) Daemon is supervised, but LLE variant [25] is unsupervised, and Daemon presents a novel clustering algorithm to derive desirable clustering results for supervised learning; 2) the merging strategy in LLE variant is similar to that in Daemon, whereas Daemon derives and understands it from the kernel design perspective which presents an insight between the kernel machine and manifold learning. With this in-depth understanding, the strategy to globally merge local analyzers can be used as a general kernel design procedure for other kernel-based algorithms and we are planning to explore this extension in our future work.

3) Computational Complexity: In the training stage of Daemon, the ICBKM is required for clustering and, hence, it is often computationally more expensive than traditional KDA. In real application, namely testing stage, Daemon has the superiority over traditional KDA in terms of computational efficiency. Denote the time complexity for a multiplication operator of two real values as $T_a$, and the time complexity for an exponential operator is $T_e$. The complexity to compute a feature for Daemon is $O(2KDT_a + KT_e)$, while for Gaussian kernel-based KDA, it is $O(NDT_a + NT_e)$, and for polynomial kernel, it is $O(NDT_a + NpT_e)$ where $p$ is the exponent.

We can see that the complexity for the testing of Daemon is independent to the sample number and it is often the case that $K \ll N$. Therefore, it is much more computationally efficient compared with KDA when the sample number is large.

IV. EXPERIMENTS

In this section, we present three sets of experiments with both artificial data and real-world data to evaluate the effectiveness of our proposed Daemon algorithm. The first toy problem on the artificial data demonstrates the effectiveness of Daemon in deriving discriminative feature for nonlinear classification problem; the face recognition results on YALE and CMU PIE databases show that Daemon significantly outperforms Face

Fig. 2. Classification with M-LDA and Daemon. Note that 1) for Mixture-LDA, the small semi-transparent area means the sample area that will be used for classification when the new data in figure comes, and we can see only part of the training samples will be used for final classification; and 2) for Daemon, the large semi-transparent area means the sample area that will be used for classification when the new data in figure comes, and we can see that all the samples are used and the samples most similar to the new data may not be in the nearest cluster.
well as Mixture LDA. Another toy problem explores the property of the geometry-adaptive-kernel when applied for Kernel PCA (KPCA) [30] algorithm.

A. Toy Problem

The objective of this experiment is two folds. One is to explore the performance of the supervised algorithm for clustering, namely ICBKM, in deriving the clustering results with balanced samples for each class within a cluster. Another is to explicitly introduce the characteristics of the local discriminant analyzers in deriving satisfying projection directions.

As shown in the top-left image of Fig. 3, the original data is composed of two classes of samples and they cannot be separated linearly. Formally, these data are synthesized according to the following distribution:

\[
\begin{align*}
    x_i^k &= 0.03 \cdot i + \delta, \\
    y_i^k &= \sin(\pi x_i^k) + k + \delta, \quad k = 0, 1, \quad \delta \sim N(0, 0.1).
\end{align*}
\]

We have systematically compared the clustering results of three K-Means-like algorithms. The original K-Means algorithm produced clustering result that aims at obtaining least sums of intra-cluster variances. As shown in the up-right image of Fig. 3, the sample numbers for the classes in a cluster is not balanced and some clusters have only one class of samples. It makes the consequent local discriminant analysis impossible in these clusters. The cluster-balanced K-means algorithm produced similar result as that of the original K-Means algorithm; yet, the sample numbers for different clusters are balanced.

As shown in the bottom-right plot of Fig. 3, the clustering result from our proposed ICBKM algorithm has the following properties: 1) the sample numbers for the classes in a cluster are balanced, which is helpful for the local discriminant analysis in each cluster; and 2) the two classes of samples in each cluster are linearly separable. It is obvious that the proposed ICBKM algorithm has produced more useful clustering result than the other two methods and intra-cluster balanced clustering result presents proper structure representation for the consequent analysis. We have also calculated the optimal local discriminant feature in each cluster to optimize the global Fisher Criterion, and the computed local feature direction in each cluster is illustrated in Fig. 4. It shows that the local feature direction is approximately optimal for the samples in each cluster in the senses of classification capability and they are merged into a global coordinate such that the samples of the same class are close to each other and samples of different classes are far to each other.

B. Face Recognition Experiments

The YALE [28] and CMU PIE [24] face databases were used to evaluate the effectiveness of our proposed Daemon algorithm for face recognition problem. In both experiments, each face image is normalized by fixing the positions of two eyes and scaling to size of $64 \times 64$ pixels. Yale face database was constructed by the Yale Center for Computational Vision and Control and it contains 165 grayscale images of 15 individuals. For each individual, six faces are used for training, and the other five are used for testing. Fig. 5 plots the 11 images of one person in the YALE database. In all our experiments, the reduced dimension number for face recognition using Fisher-face is $L = 1$ as in [2]. The kernel used for traditional Kernel-DA algorithm is Gaussian Kernel $K(x, y) = \exp \left(-\frac{\|x - y\|^2}{2\sigma^2} \right)$, and for a fair comparison, we tested 21 different parameters as $\eta_i = 2^{(i-11)/2\delta}$ where $\delta$ is the standard variance of the sample data. Moreover, in this experiment, we compare Daemon with Mixture LDA that learns different LDA models for different clusters derived from ICBKM, and we also compare Daemon with K-Means-Daemon algorithm which is similar to Daemon by replacing the ICBKM algorithm with the K-Means for clustering. For Daemon, we set $\alpha = 0.15, \beta = 0.1$ as in the toy problem; the cluster number $K$ is explored between 2–7 and the best result is reported. Table I illustrates the face recognition results of Fisher-face, KDA, Mixture LDA, K-Means-Daemon, and Daemon. It shows that Daemon significant outperforms LDA, Mixture LDA and K-Means-Daemon, and also has better results than traditional KDA with Gaussian Kernel.

Fig. 3. Toy problem on clustering on the synthesized data ($\alpha = 0.15, \beta = 0.1$). Note that top-left is the plot of the original samples of two classes, top-right is the plot of the clustering result of $K$-means, bottom-left is the plot of the clustering result of CBKM and bottom-right is the plot of the clustering result of our proposed ICBKM.

Fig. 4. Derived local feature directions using Daemon. We can see that the plotted directions are approximately optimal for local classification.
TABLE I

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Fisher-face</th>
<th>Kernel-DA</th>
<th>Mixture-LDA</th>
<th>K-Means-Daemon</th>
<th>Daemon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>80%</td>
<td>85.3%</td>
<td>82.7%</td>
<td>82.7% (K=2)</td>
<td>88% (K=2)</td>
</tr>
</tbody>
</table>

TABLE II

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Fisher-face</th>
<th>Kernel-DA</th>
<th>Daemon(K=5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>63.6%</td>
<td>68.4%</td>
<td>71.1%</td>
</tr>
</tbody>
</table>

TABLE III

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Fisher-face</th>
<th>Kernel-DA</th>
<th>Daemon(K=5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>88.5%</td>
<td>90.7%</td>
<td>93.2%</td>
</tr>
</tbody>
</table>

Fig. 5. Eleven images of one person in the YALE database and the images are aligned by fixing the positions of the two eyes.

Fig. 6. Some sample images of one person in the CMU PIE database and the images are aligned by fixing the positions of the two eyes.

We also conducted the face recognition experiments on the PIE database. Some example images of one person in PIE database are plotted in Fig. 6. The face images of pose 02, 37, 05, 27, 29, and 11 in the illumination directory are used in our experiments and referred to as PIE-1 subdatabase. The results over ten random splits with equal numbers of training and testing samples are averaged and the results listed in Table II show that Daemon outperforms the other two algorithms in the PIE-1 subdatabase as in YALE database. It also demonstrates that Daemon has strong capability to handle nonlinear classification problems and can improve the accuracy in the general classification problems compared with Fisher-face and Kernel-DA. Moreover, we also conducted a relative simple experiment on the frontal images with only 15 types of different illumination conditions in the illumination directory of the PIE database, referred to as PIE-2 subdatabase. The results in Table III show that our proposed Daemon algorithm performs best and Fig. 7 plots the comparative face recognition accuracies on different feature dimensions. Note that the directions from 68–80th are selected as the eigenvalues with larger nonzero eigenvalues produced by computational error.

C. Property of Geometry-Adaptive-Kernel

We have proved that Daemon is essentially a special KDA algorithm with geometry-adaptive-kernel, which is dynamic and adaptive with the data distribution and label information. In this section, we explore the property of this kernel when applied to KPCA for nonlinear PCA.

KPCA has the potential to capture nonlinear structure of the data set. Hence, we compare the linear PCA with the Kernel-PCA using four different kernels: Gaussian, polynomial, inverse multiquadric kernel and our proposed geometry-adaptive-kernel. The data set are sampled from a structure integrating the symbol “V” and a horizontal line, i.e., with both nonlinear and linear structures. Fig. 8 demonstrates the distributions of the data projections to the first principal component of KPCA and the lines are the contours of the projection values. It is evident that KPCA algorithm with geometry-adaptive-kernel best captures the intrinsic nonlinear structure of the data set. PCA cannot capture the nonlinear structure of the data as the linear property.

V. DISCUSSIONS AND FUTURE WORK

We have presented a novel algorithm called Daemon for general nonlinear classification problem. Daemon is a nonlinear discriminant analysis algorithm that effectively utilizes the underlying nonlinear manifold structure. In this work, the discrete sample data on a manifold are clustered using the ICBKM algorithm such that the sample numbers for the
classes within each cluster are balanced; and then the local optimal discriminant features are simultaneously derived by optimizing the global Fisher Criterion. Daemon can be justified from kernel-machine perspective and is a KDA algorithm with geometry-adaptive-kernel, a special marginalized kernel tailed to the nonlinear manifold structure of the sample data.

To the best of our knowledge, this is the first work to conduct nonlinear discriminant analysis while explicitly considering the embedded geometry structure of the data set. In this work, we have only utilized the basic property of manifold that a manifold can be covered by a series of open sets; how to combine the other topology properties of a manifold with discriminant analysis for general classification problem is the future direction of our work. Moreover, in the clustering task, the initialization issue and how to select the optimal cluster number are still open problems, and we also plan to explore these problems in our future work.

REFERENCES


Fig. 8. Toy problem to illustrate the capability of KPCA with geometry-adaptive-kernel to capture the principal curve of the nonlinearly distributed data. From left to right and top to bottom are the contour line image of the projection to the first principal component direction of PCA and KPCA with different types of kernels. Note that contour value at one point is computed by projecting its coordinates to the first principal direction. The curve with arrows arrow means the desired principal curve direction; and for a good result, the contour line should be orthogonal to the curve with arrow and the projection value should change monotonously along the direction of the curve with arrow. Therefore, KPCA with geometry-adaptive-kernel is the best.

**Shuicheng Yan** (M’06) received the B.S. and Ph.D. degrees from the Applied Mathematics Department, School of Mathematical Sciences, Peking University, Peking, China, in 1999 and 2004, respectively. Currently, he is a Postdoctoral Fellow at the University of Illinois at Urbana-Champaign. His research interests include computer vision and machine learning.

**Yuxiao Hu** received the B.S. and M.S. degrees in computer science from the Tsinghua University, Beijing, China, in 1999 and 2001, respectively. He is currently working toward the Ph.D. degree at the Beckman Institute, University of Illinois at Urbana-Champaign. His research interests include human computer interface, multimedia signal processing, machine learning, and face-related projects such as detection, tracking, pose estimation, and recognition.

**Dong Xu** received the B.S. and Ph.D. degrees from the Electronic Engineering and Information Science Department, University of Science and Technology of China, Hefei, China, in 2001 and 2005, respectively. He is currently a Postdoctoral Research Scientist at the Department of Electrical Engineering, Columbia University, New York City. His research interests include pattern recognition, computer vision and machine learning.

**Hong-Jiang Zhang** (F’04) received the B.S. degree from Zhengzhou University, China, in 1982 and the Ph.D degree from the Technical University of Denmark, Lyngby, Denmark, in 1991, and both in electrical engineering.

From 1992 to 1995, he was with the Institute of Systems Science, National University of Singapore, where he led several projects in video and image content analysis and retrieval and computer vision. From 1995 to 1999, he was a Research Manager at Hewlett-Packard Laboratories, responsible for research and technology transfers in the areas of multimedia management, computer vision and intelligent image processing. In 1999, he joined Microsoft Research Asia, Beijing, China, and is currently the Managing Director of the Advanced Technology Center. He has authored three books, more than 300 referred papers, eight special issues of international journals on image and video processing, content-based media retrieval, and computer vision, as well as more than 50 patents or pending applications.

Dr. Zhang currently serves on the Editorial Boards of five IEEE/ACM journals and a dozen committees of international conferences.

**Benyu Zhang** received the B.S. and M.S. degrees in computer science from Peking University, Peking, China, in 1999 and 2002, respectively. He joined Microsoft Research Asia, Beijing, China, in July 2002 to pursue his wide-ranging research interests in machine learning, information retrieval, data mining, and artificial intelligence.

**Qiansheng Cheng** received the B.S. degree in mathematics from Beijing University, Beijing, China, in 1963.

He is now a Professor in the Department of Mathematics and the Vice Director of the Institute of Mathematics, Beijing University, Beijing, China. His research interests include signal processing, time series analysis, system identification, and pattern recognition.

Prof. Cheng is the Vice Chairman of Chinese Signal Processing Society and has won the Third Chinese National Science Award.