sensors

MDPI

Article

Hierarchical Discriminant Analysis

Di Lu †, Chuntao Ding *,†, Jinliang Xu and Shangguang Wang

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China; ludi8418@gmail.com (D.L.); jlxu@bupt.edu.cn (J.X.); sgwang@bupt.edu.cn (S.W.)

* Correspondence: ctding@bupt.edu.cn

† These authors contributed equally to this work.

Received: 6 December 2017; Accepted: 13 January 2018; Published: 18 January 2018

Abstract: The Internet of Things (IoT) generates lots of high-dimensional sensor intelligent data. The processing of high-dimensional data (e.g., data visualization and data classification) is very difficult, so it requires excellent subspace learning algorithms to learn a latent subspace to preserve the intrinsic structure of the high-dimensional data, and abandon the least useful information in the subsequent processing. In this context, many subspace learning algorithms have been presented. However, in the process of transforming the high-dimensional data into the low-dimensional space, the huge difference between the sum of inter-class distance and the sum of intra-class distance for distinct data may cause a bias problem. That means that the impact of intra-class distance is overwhelmed. To address this problem, we propose a novel algorithm called Hierarchical Discriminant Analysis (HDA). It minimizes the sum of intra-class distance first, and then maximizes the sum of inter-class distance. This proposed method balances the bias from the inter-class and that from the intra-class to achieve better performance. Extensive experiments are conducted on several benchmark face datasets. The results reveal that HDA obtains better performance than other dimensionality reduction algorithms.

Keywords: Internet of Things; intelligent data; subspace learning; marginal fisher analysis; dimensionality reduction; discriminant neighborhood embedding

1. Introduction

In recent years, the high penetration rate of Internet of Things (IoT) in all activities of everyday life is fostering the belief that for any kind of high-dimensional IoT data there is always a solution able to successfully deal with it. The proposed solution is the dimensionality reduction algorithm. It is well known that the high dimensionality of feature vectors is a critical problem in practical pattern recognition. Naturally, dimensionality reduction technology has been shown to be of great importance to data preprocessing, such as face recognition [1] and image retrieval [2]. It reduces the computational complexity through reducing the dimension and improving the performance at the same time. A general framework [3] for dimensionality reduction defines a general process according to describing the different purposes of the dimensionality reduction algorithm that was proposed. Among this general framework, the dimensionality reduction is divided into unsupervised algorithms and supervised algorithms.

Unsupervised algorithms conduct datasets without labels, including principle component analysis (PCA) [4], locality preserving projection (LPP) [5], neighborhood preserving embedding (NPE) [6], etc. Supervised algorithms conduct datasets with labels that aim to present better performance and low complexity. Linear discriminant analysis (LDA), local discriminant embedding (LDE) [7], discriminant sparse neighborhood preserving embedding (DSNPE) [8], regularized coplanar discriminant analysis (RCDA) [9], marginal Fisher analysis (MFA) [3,5,10], discriminant neighborhood

embedding (DNE) [11], locality-based discriminant neighborhood embedding(LDNE) [12], and double adjacency graphs-based discriminant neighborhood embedding (DAG-DNE) [13] are typical supervised algorithms.

The unsupervised algorithm PCA [4] adopts linear transformation to achieve dimensional reduction commendably. However, PCA shows a noneffective performance in handling manifold data. Then, LLE [14] is proposed, which firstly uses linear coefficients to represent the local geometry of a given point, then explores a low-dimensional embedding for reconstruction in the subspace. However, LLE only defines mappings on the training data which means it is ill-defined in defining mappings on the testing data. To remedy this problem, NPE [6], orthogonal neighborhood preserving projection (ONPP) [15] and LPP are put on the table. They figure out the problem of LLE while preserving the original structure. NPE and ONPP solve the generalized eigenvalue problem with different constraint conditions. LPP [5] reserves the local structure by preferring an embedding algorithm. This algorithm extends to new points easily. However, as an unsupervised dimensionality reduction algorithm, LPP only shows good performance in datasets without labels.

To address the problem that unsupervised algorithms cannot work well in classification tasks, many supervised algorithms are proposed, such as linear discriminant analysis (LDA) [16], which maximizes the inter-class scatter, minimizes the intra-class scatter simultaneously and finds appropriate project directions for classification tasks. However, LDA still has some limitations. For instance, LDA only considers the global Euclidean structure. Marginal Fisher analysis [10] is proposed as an improved algorithm to surmount this problem. MFA constructs the penalty graph and the intrinsic graph to hold the local structure thereby solving the limitation problem of LDA. Both LDA and MFA can discover the projection directions that simultaneously maximize the inter-class scatter and minimize the intra-class scatter. DAG-DNE compensates for the small inter-class scatter in the subspace of DNE [11]. DAG-DNE constructs a homogeneous neighbor graph and heterogeneous neighbor graph to maintain the original structure perfectly. It maximizes the margin between the inter-class scatter and intra-class scatter so that points in the same class are compact and points in the different classes become separable in the subspace at the same time. However, DAG-DNE and all above dimensionality reduction algorithms, optimize intra-class and inter-class simultaneously. The inter-class scatter is larger than the intra-class scatter, thus it will aim at inter-class but ignore the intra-class. Thus, the huge difference between the sum of inter-class distance and the sum of intra-class distance for distinct datasets may cause a bias problem, which means the impact of intra-class distance is tiny in optimization tasks. Naturally, the separation optimization idea comes to mind.

In this paper, we proposed a novel supervised subspace learning algorithm to further increase the performance, called hierarchical discriminant analysis. More concretely, we construct two adjacency graphs by distinguishing homogeneous and heterogeneous neighbors in HDA. Then HDA optimizes inter-class distance and intra-class distance independently. We minimize the intra-class distance first, then maximize the inter-class distance. Because of the hierarchical work, the optimization of intra-class distance and inter-class distance are detached. The process of optimization does not have to be biased to the inter-class scatter. Both of them get the best results. Thus, the influence between classes can be eliminated and we can find a good projection matrix.

The rest of this paper is structured as follows. In Section 2 we introduce the background on MFA, LDNE and DAG-DNE.I. Section 3, we dedicate to introducing HDA and revealing its connections with MFA, LDNE and DAG-DNE. In Section 4, several simulation experiments applying our algorithm are presented to verify its effectiveness. This is followed by the conclusions in Section 5.

2. Backgroud

In this section we emphatically introduce three classical algorithms, MFA [10], LDNE [12] and DAG-DNE [13], which are related to our research point.

2.1. Marginal Fisher Analysis

MFA proposes a new criteria. It can conquer the limitation of LDA by characterizing the intra-class compactness and inter-class separability. MFA algorithm follows the steps below:

(1) Construct the adjacency matrices. There are intrinsic graph characterizing the intra-class compactness and penalty graph characterizing the inter-class separability. For each sample $\mathbf{x}_i$, the intra-class matrix $\mathbf{S}^w$ is defined as:

$$S_{ij}^{w} = \begin{cases} 1, \ \mathbf{x}_i \in N_{K_1}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_{K_1}(\mathbf{x}_i) \\ 0, \ \text{otherwise} \end{cases} \tag{1}$$

where $N_{K_1}(\mathbf{x}_i)$ represents the index set of the K_1 nearest neighbors of node $\mathbf{x}_i$ in the same class.

For each sample $\mathbf{x}_i$, the inter-class matrix $\mathbf{S}^b$ is defined as:

$$S_{ij}^{b} = \begin{cases} 1, \ \mathbf{x}_i \in P_{K_2}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in P_{K_2}(\mathbf{x}_i) \\ 0, \ \text{otherwise} \end{cases} \tag{2}$$

where $P_{K_2}(\mathbf{x}_i)$ expresses the set of the K_2 nearest neighbors of point $\mathbf{x}_i$ that are in the different classes.

(2) Marginal Fisher Analysis Criterion. Find the optimal projection direction by minimizing the intra-class compactness and maximizing the inter-class separability at the same time.

$$\min_{\mathbf{P}} \frac{tr\left(\mathbf{P}^T\mathbf{X}\left(\mathbf{D^w}-\mathbf{S^w}\right)\mathbf{X}^T\mathbf{P}\right)}{tr\left(\mathbf{P}^T\mathbf{X}\left(\mathbf{D^b}-\mathbf{S^b}\right)\mathbf{X}^T\mathbf{P}\right)} \tag{3}$$

$tr(\cdot)$ is the trace of a matrix, $D_{ii}^{w} = \sum_j S_{ij}^{w}, D_{ii}^{b} = \sum_j S_{ij}^{b}$, and $\mathbf{P}$ is composed of the optimal r projection vectors, $\mathbf{X}$ is the set of samples.

2.2. Locality-Based Discriminant Neighborhood Embedding

LDNE uses a weight function instead of 1 or 0 to adopt the adjacency graph. It maximizes the difference between the inter-class scatter and the intra-class scatter to capture the top projection matrix. The LDNE algorithm is done by the following steps:

(1) Use K-nearest neighbors to construct the adjacency. The adjacency weight matrix $\mathbf{S}$ is declared as:

$$S_{ij} = \begin{cases} -\exp(\frac{-||\mathbf{x}_i-\mathbf{x}_j||^2}{\beta}), \ \mathbf{x}_i \in S_K^w(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in S_K^w(\mathbf{x}_i) \\ +\exp(\frac{-||\mathbf{x}_i-\mathbf{x}_j||^2}{\beta}), \ \mathbf{x}_i \in S_K^b(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in S_K^b(\mathbf{x}_i) \\ 0, \qquad\qquad\qquad\quad \text{otherwise} \end{cases} \tag{4}$$

where $S^{\omega}(\mathbf{x}_i)$ denotes the intra-class neighbors of each sample point $\mathbf{x}_i$, $S^b(\mathbf{x}_i)$ denotes the inter-class neighbors of $\mathbf{x}_i$, and the parameter β is a regulator.

(2) Feature mapping: Optimize the following objective function:

$$\begin{cases} \max\limits_{\mathbf{P}} \ \text{tr}\{\mathbf{P}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{P}\} \\ \text{s.t.} \ \ \mathbf{P}^T\mathbf{P} = \mathbf{I} \end{cases} \tag{5}$$

where $\mathbf{H} = \mathbf{D} - \mathbf{S}$, and $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \sum_j S_{ij}$. Function (5) can be considered as a generalized eigen-decomposition problem:

$$\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{P} = \lambda\mathbf{P} \tag{6}$$

The optimal projection $\mathbf{P}$ consists of r eigenvectors corresponding to the r largest eigenvalues.

2.3. Double Adjacency Graphs-Based Discriminant Neighborhood Embedding

DAG-DNE is a linear manifold learning algorithm based on DNE, which constructs homogeneous and heterogeneous neighbor adjacency graphs. This algorithm follows the steps below:

(1) Construct two adjacency graphs. Let $\mathbf{A}^w$ and $\mathbf{A}^b$ be the intra-class and inter-class adjacency matrices. $\mathbf{x}_i$ is the sample point. The intra-class adjacency matrix $\mathbf{A}^w$ is defined as

$$A_{ij}^w = \begin{cases} 1, \ \mathbf{x}_i \in S_K^w(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in S_K^w(\mathbf{x}_i) \\ 0, \ \text{otherwise} \end{cases} \tag{7}$$

The inter-class adjacency matrix $\mathbf{A}^b$ is defined as

$$A_{ij}^b = \begin{cases} 1, \ \mathbf{x}_i \in S_K^b(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in S_K^b(\mathbf{x}_i) \\ 0, \ \text{otherwise} \end{cases} \tag{8}$$

(2) Optimize the following objective function by finding a projection $\mathbf{P}$.

$$\begin{cases} \max\limits_{\mathbf{P}} \ \text{tr}\{\mathbf{P}^T\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{P}\} \\ \text{s.t.} \ \ \mathbf{P}^T\mathbf{P} = \mathbf{I} \end{cases} \tag{9}$$

where $\mathbf{G} = \mathbf{D}^b - \mathbf{A}^b - \mathbf{D}^w + \mathbf{A}^w$, and $\mathbf{D}^b$ and $\mathbf{D}^w$ are diagonal matrices with $D_{ii}^b = \sum_j A_{ij}^b$ and $D_{ii}^w = \sum_j A_{ij}^w$. The projection matrix $\mathbf{P}$ can be solved through the generalized eigenvalue problem as follows:

$$\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{P} = \lambda\mathbf{P} \tag{10}$$

The optimal projection $\mathbf{P}$ consists of r eigenvectors corresponding to the r largest eigenvalues.

From optimizing the equation of the related work we can see that the optimization work is achieved through maximizing the distance of inter-class minus the distance of the intra-class or maximizing the distance inter-class divided by the intra-class. Thus, the inter-class and intra-class are simultaneously optimized. Our hierarchical discriminant analysis separates the inter-class and the intra-class. This method avoids the interference between the inter-class and the intra-class. The detail of our algorithm will be presented in the next part.

3. Hierarchical Discriminant Analysis

This section discusses a novel dimensional reduction algorithm called hierarchical discriminant analysis. In MFA and DAG-DNE, two adjacency matrices for a node's homogenous neighbors and heterogeneous neighbors are optimized simultaneously. The huge difference between the sum of inter-class distance and the sum of intra-class distance for distinct data may cause a bias problem, which is too weak for intra-class scatter to play a fundamental role. HDA considers the adjacency graphs respectively, which optimizes the sum of distances between each node and its neighbors of the same class firstly, and then optimizes the sum of distances between the nodes of different classes.

3.1. HDA

Suppose $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ is the set of training points, where N is the number of training points, $\mathbf{x}_i \in \mathbb{R}^D$, $y_i \in \{1, 2, \ldots, c\}$, y_i is the class label of $\mathbf{x}_i$, d is the dimensionality of points, and c is classes' number. For these unclassified training sets, we come up with a potential manifold subspace. We use matrix transformation to project the original high-dimensional data into a low-dimensional space. By doing so, homogeneous samples are compacted and the heterogeneous samples are scattered in low-dimensional data. Consequently, the computational complexity and classification performance are improved.

Similar to DAG-DNE [13], HDA constructs two adjacency matrices respectively, the intra-class and inter-class adjacency matrices. Given a sample $\mathbf{x}_i$, we suppose the set of its k homogenous and heterogeneous neighbors are $\pi_k^+(\mathbf{x}_i)$ and $\pi_k^-(\mathbf{x}_i)$. We construct the intra-class adjacency matrix $\mathbf{F}^w$ and the inter-class adjacency matrix $\mathbf{F}^b$.

$$F_{ij}^w = \begin{cases} +1, & \mathbf{x}_i \in \pi_k^+(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \pi_k^+(\mathbf{x}_i) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

$$F_{ij}^b = \begin{cases} +1, & \mathbf{x}_i \in \pi_k^-(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \pi_k^-(\mathbf{x}_i) \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

And the local intra-class scatter is defined as:

$$\begin{aligned} \Phi(\mathbf{P_1}) &= \sum_{\mathbf{x}_i \in \pi_k^+(\mathbf{x}_j)} \sum_{\mathbf{x}_j \in \pi_k^+(\mathbf{x}_i)} \left\| \mathbf{P_1}^T \mathbf{x}_i - \mathbf{P_1}^T \mathbf{x}_j \right\|^2 F_{ij}^w \\ &= \sum_{\mathbf{x}_i \in \pi_k^+(\mathbf{x}_j)} \sum_{\mathbf{x}_j \in \pi_k^+(\mathbf{x}_i)} \{\mathbf{P_1}^T (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{P_1}\} F_{ij}^w \\ &= \sum_{\mathbf{x}_i \in \pi_k^+(\mathbf{x}_j)} \sum_{\mathbf{x}_j \in \pi_k^+(\mathbf{x}_i)} \{\mathbf{P_1}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{P_1} - \mathbf{P_1}^T \mathbf{x}_i \mathbf{x}_j^T \mathbf{P_1} - \mathbf{P_1}^T \mathbf{x}_j \mathbf{x}_i^T \mathbf{P_1} + \mathbf{P_1}^T \mathbf{x}_j \mathbf{x}_j^T \mathbf{P_1}\} \\ &= 2 \sum_{\mathbf{x}_i \in \pi_k^+(\mathbf{x}_j)} \{\mathbf{P_1}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{P_1}\} D_{ii}^w - 2 \sum_{\mathbf{x}_i \in \pi_k^+(\mathbf{x}_j)} \sum_{\mathbf{x}_j \in \pi_k^+(\mathbf{x}_i)} \{\mathbf{P_1}^T \mathbf{x}_i \mathbf{x}_j^T \mathbf{P_1}\} F_{ij}^w \\ &= 2\{\{\mathbf{P_1}^T (\sum_{\mathbf{x}_i \in \pi_k^+(\mathbf{x}_j)} \mathbf{x}_i D_{ii}^w \mathbf{x}_i^T) \mathbf{P_1}\} - \{\mathbf{P_1}^T (\sum_{\mathbf{x}_i \in \pi_k^+(\mathbf{x}_j)} \sum_{\mathbf{x}_j \in \pi_k^+(\mathbf{x}_i)} \mathbf{x}_i F_{ij}^w \mathbf{x}_j^T) \mathbf{P_1}\}\} \\ &= 2\mathbf{P_1}^T \mathbf{X} (\mathbf{D}^w - \mathbf{F}^w) \mathbf{X}^T \mathbf{P_1} \end{aligned} \tag{13}$$

where $\mathbf{D}^w$ is a diagonal matrix and its entries are column sum of $\mathbf{F}^w$, i.e., $D_{ii}^w = \sum_j F_{ij}^w$. First, we minimize the intra-class scatter, which means:

$$\begin{cases} \min \Phi(\mathbf{P_1}) \\ \text{s.t.} \mathbf{P_1}^T \mathbf{P}_1 = \mathbf{I} \end{cases} \tag{14}$$

The objective function can be rewritten by some algebraic steps:

$$\begin{aligned} \Phi(\mathbf{P_1}) &= 2tr\left\{\mathbf{P_1}^T \mathbf{X} (\mathbf{D^w} - \mathbf{F^w}) \mathbf{X}^T \mathbf{P_1}\right\} \\ &= 2tr\left\{\mathbf{P_1}^T \mathbf{X} \mathbf{S} \mathbf{X}^T \mathbf{P_1}\right\} \end{aligned} \tag{15}$$

where $\mathbf{S} = \mathbf{D^w} - \mathbf{F^w}$. Therefore, the form of trace optimization function can be rewritten as

$$\begin{cases} \min\limits_{\mathbf{P_1}} \text{tr}\{\mathbf{P_1}^T \mathbf{X} \mathbf{S} \mathbf{X}^T \mathbf{P_1}\} \\ \text{s.t.} \mathbf{P_1}^T \mathbf{P_1} = \mathbf{I} \end{cases} \tag{16}$$

While the local inter-class scatter is defined as:

$$\begin{aligned} \Psi(\mathbf{P}) &= \sum_{\mathbf{x}_i \in \pi_k^-(\mathbf{x}_j)} \sum_{\mathbf{x}_j \in \pi_k^-(\mathbf{x}_i)} \left\| \mathbf{P}^T \mathbf{x}_i - \mathbf{P}^T \mathbf{x}_j \right\|^2 \\ &= 2\mathbf{P}^T \mathbf{X} \left(\mathbf{D}^b - \mathbf{F}^b\right) \mathbf{X}^T \mathbf{P} \end{aligned} \tag{17}$$

where $\mathbf{D}^b$ is a diagonal matrix and its entries are the column sum of $\mathbf{F}^b$, i.e., $D_{ii}^b = \sum_j F_{ij}^b$.

Now, we maximize the inter-class scatter:

$$\begin{cases} \max \Psi(\mathbf{P}) \\ \text{s.t.} \mathbf{P}^T\mathbf{P} = \mathbf{I} \end{cases} \tag{18}$$

It is worth noting that the training set has changed since first minimization, and we rewrite this function as:

$$\begin{aligned} \Psi(\mathbf{P}) &= 2tr\left\{\mathbf{P}^T\mathbf{X}_{new}\left(\mathbf{D}^b - \mathbf{F}^b\right)\mathbf{X}_{new}^T\mathbf{P}\right\} \\ &= 2tr\left\{\mathbf{P}^T\mathbf{X}_{new}\mathrm{M}\mathbf{X}_{new}^T\mathbf{P}\right\} \end{aligned} \tag{19}$$

where $\mathbf{X}_{new} = \mathbf{X}\mathbf{P}_1$.

So that the optimization problem as shown below:

$$\begin{cases} \max\limits_{\mathbf{P}} \mathrm{tr}\left\{\mathbf{P}^T\mathbf{X}_{new}\mathbf{M}\mathbf{X}_{new}^T\mathbf{P}\right\} \\ \text{s.t.} \mathbf{P}^T\mathbf{P} = \mathrm{I} \end{cases} \tag{20}$$

Since **S** and **M** are real symmetric matrices, the optimization scatters (17) and (20) are same as the eigen-decomposition problem of the matrices $\mathbf{X}\mathbf{S}\mathbf{X}^T$ and $\mathbf{X}\mathbf{M}\mathbf{X}^T$. The two projection matrices **P** are composed of the egienvetors that corresponding to the egienvalues of $\mathbf{X}\mathbf{S}\mathbf{X}^T$ and $\mathbf{X}\mathbf{M}\mathbf{X}^T$. The optimal solutions $\mathbf{P_1}$ and **P** have the form $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_r]$.

Therefore, the image of any point $\mathbf{x}_i$ can be represented as $\mathbf{v}_i = \mathbf{P}^T\mathbf{x}_i$. The details about HDA are given below (Algorithm 1).

Algorithm 1 Hierarchical Discriminant Analysis

Input: A training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, and the dimensionality of discriminant subspace r.

Output: Projection matrix **P**.

1: Compute the intra-class adjacency graph $\mathbf{F}^w$

$$F_{ij}^w = \begin{cases} +1, & \mathbf{x}_i \in \pi_k^+(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \pi_k^+(\mathbf{x}_i) \\ 0 & \text{otherwise} \end{cases}$$

and the inter-class adjacency matrix $\mathbf{F}^b$.

$$F_{ij}^b = \begin{cases} +1, & \mathbf{x}_i \in \pi_k^-(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \pi_k^-(\mathbf{x}_i) \\ 0 & \text{otherwise} \end{cases}$$

2: Minimize the intra-class distance by decomposing the matrix $\mathbf{X}\mathbf{S}\mathbf{X}^T$, where $\mathbf{S} = \mathbf{D^w} - \mathbf{F^w}$. Let egienvalues be $\lambda_i, i = 1, \ldots, d$ and their corresponding egienvectors be $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$.

3: Choose the first r smallest egienvalues so that return $\mathbf{P_1} = [\mathbf{p}_1, \ldots, \mathbf{p}_r]$.

4: Compute the new input $\mathbf{X}_{new} = \mathbf{X}\mathbf{P}_1$.

5: Maximize the inter-class distance by decomposing the matrix $\mathbf{X}\mathbf{M}\mathbf{X}^T$, where $\mathbf{S} = \mathbf{D^b} - \mathbf{F^b}$. Let egienvalues be $\lambda_i, i = 1, \ldots, d$ and their corresponding egienvectors be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.

6: Choose the first r largest egienvalues so that return $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_r]$.

3.2. Comparisons with MFA, LDNE and DAG-DNE

HDA, LDNE and DAG-DNE are all dimensionality reduction algorithms. In this section, we will probe into the relationships between HDA and the other three algorithms.

3.2.1. HDA vs. MFA

These two algorithms both build an intrinsic(intra-class) graph and penalty(inter-class) graph to keep the original structure of the given data. MFA designs two graphs to find the projection matrix which characterizes the intra-class closely and inter-class separately. However, MFA chooses the

dimensionality of the discriminant subspace by experience, which causes some important information to be lost. HDA maximizes the distance of the inter-class and minimizes the distance of the intra-class hierarchically. It uses this method to find the projection matrix to estimate the dimension of the discriminant subspace.

3.2.2. HDA vs. LDNE

Both HDA and LDNE are supervised subspace learning algorithms and build two adjacency graphs to keep the original structure of the given data. LDNE assigns different weights to intra-class neighbors and inter-class neighbors of a given point. It follows the 'locality' idea of LPP, and produces balanced links through the set up connection between each point and its heterogeneous neighbors. By analyzing the experimental result, we get an effectiveness result on the subspace.

3.2.3. HDA vs. DAG-DNE

DAG-DNE algorithm maintains balanced links by constructing two adjacency graphs. Through this method, samples in the same class are compact and samples in different classes are separable in the discriminant subspace. However, DAG-DNE optimizes the heterogeneous neighbors and homogeneous neighbors simultaneously. By doing this, the distance of inter-class scatter is not wide enough in the subspace. The HDA algorithm also constructs two graphs, and separately minimizes the within-class distance first, then maximizes the inter-class distance. Because of the hierarchical work, the optimization of intra-class distance and inter-class distance are detached. The process of optimization is not biased to the inter-class scatter. As a consequence, the intra-class is compact while the inter-class is separated in the subspace.

4. Experiments

We illustrate a set of experiments to confirm the performance of HDA on image classification in this part. Several benchmark datasets are used, i.e., Yale, Olivetti Research Lab (ORL) and UMIST. The samples from three datasets are shown in Figures 1–3. We compare HDA with other representative dimensional reduction algorithms, including LPP, MFA and DAG-DNE.W. Then we choose the nearest neighbor parameter K for those algorithms when constructing the adjacency graphs. We exhibited the results in the form of mean recognition rate.

Figure 1. Samples from the Yale face dataset.

Figure 2. Samples from the ORL face dataset.

Figure 3. Samples from the UMIST face dataset.

4.1. Yale Dataset

The Yale datset is constructed by the Yale Center for Computational Vision and Control. It contains 165 gray scale images of 15 individuals. These images display variations in lighting condition and facial expression (normal, happy, sad, sleepy, surprised and wink). Each image is 32×32 dimension.

We choose 60% from the dataset as training samples, and the rest of the samples for testing. The nearest neighbor parameter K can be taken as 1, 3 and 5. Figure 4 shows the average accuracy rate of $K = 1$ after running the four algorithms 10 times. Figure 5 with $K = 3$ and Figure 6 with $K = 5$ also obtain similar results like Figure 4 after being run 10 times. From these three figures, we can see that HDA has a higher accuracy rate than other three algorithms and we can gain the best performance.

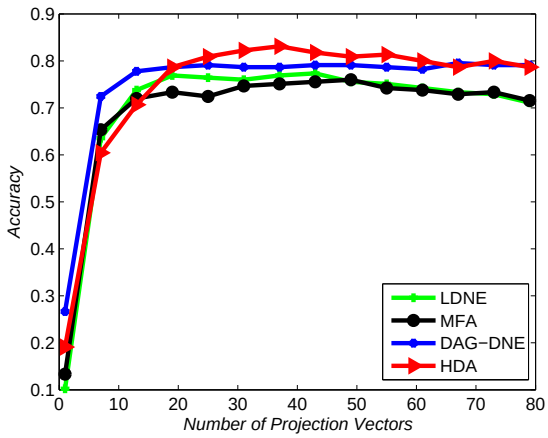

Figure 4. Accuracy vs. Number of Projection Vectors on the Yale dataset under $K = 1$.

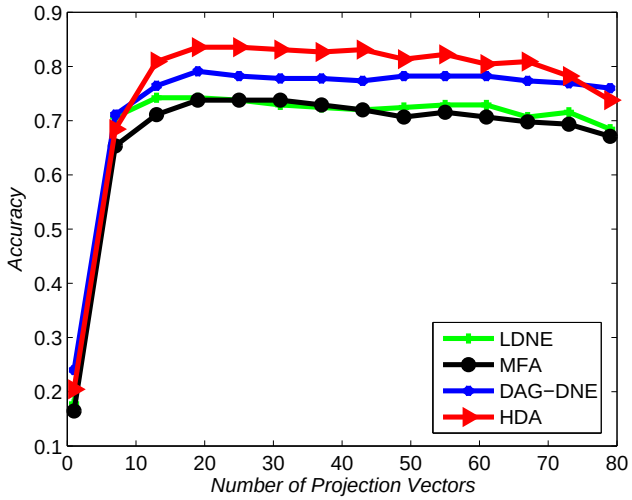

Figure 5. Accuracy vs. Number of Projection Vectors on the Yale dataset under $K = 3$.

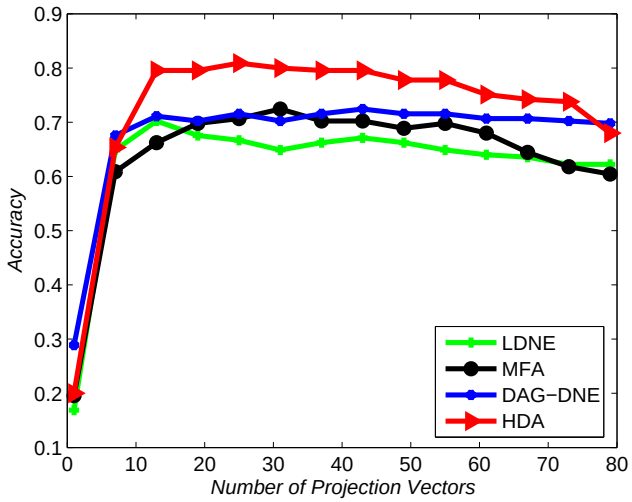

Figure 6. Accuracy vs. Number of Projection Vectors on the Yale dataset under $K = 5$.

4.2. UMIST Dataset

The UMIST dataset consists of 564 images of 20 individuals. The dataset takes race, sex and appearance into account. Each individual takes several poses from profile to frontal views. The original size of each image is 112 × 92 pixels. In our experiments, the whole dataset is resized to 32 × 32 pixels.

Similar to applying the Yale dataset, we use UMIST dataset to test the recognition rate of the proposed algorithm and the other three algorithms. We select 20% from the UMIST dataset as training samples and use others as testing samples. Figures 7–9 show the average accuracy rate of different K after running the four algorithms 10 times. As shown in these figures, our algorithm reaches the top and presents the best recognition accuracy compare to the other three algorithms.

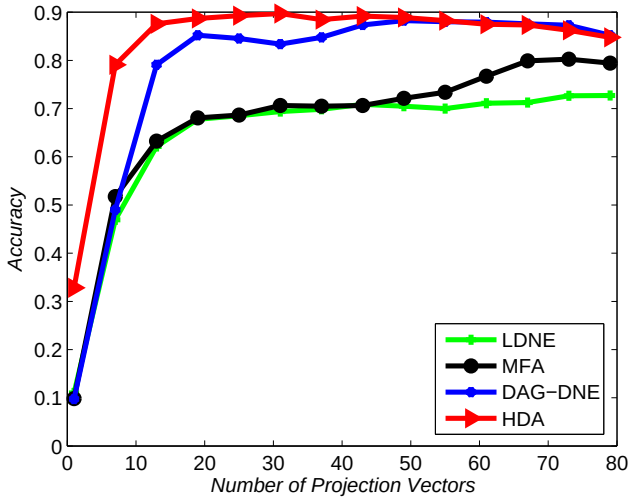

Figure 7. Accuracy vs. Number of Projection Vectors on the UMIST dataset under $K = 1$.

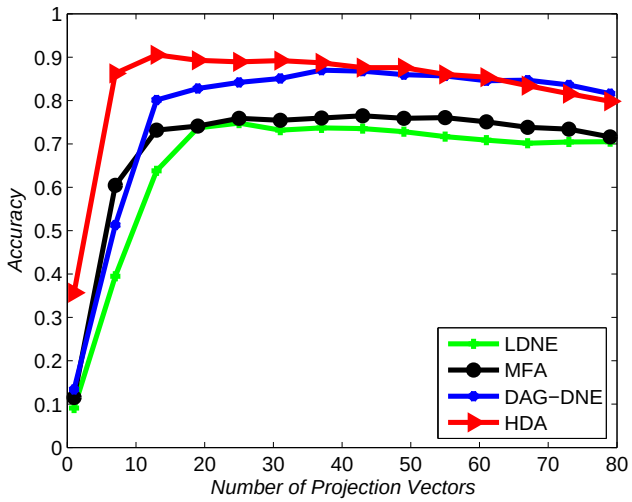

Figure 8. Accuracy vs. Number of Projection Vectors on the UMIST dataset under $K = 3$.

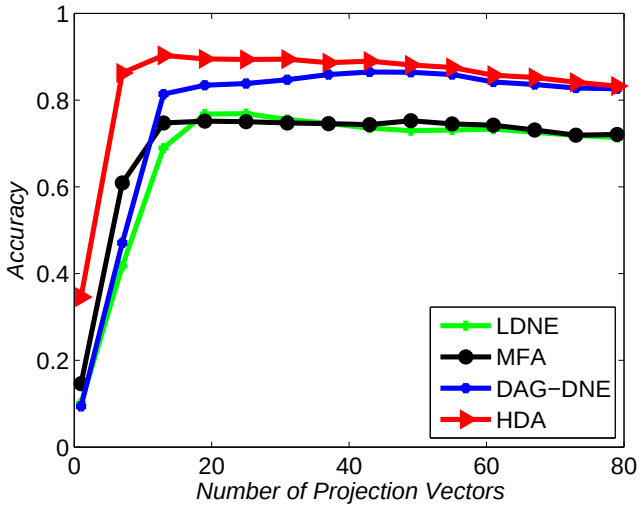

Figure 9. Accuracy vs. Number of Projection Vectors on the UMIST dataset under $K = 5$.

4.3. ORL Dataset

The Olivetti Research Lab dataset is composed of 40 distinct objects. Each of them contains 10 different images. The ORL dataset contains images in different conditions, such us light conditions, facial expressions, etc. All of the images were taken under a dark homogenous background in an upright and frontal position. The size of each image is 112 $\times$ 92 pixels, with 256 gray levels per pixel. Similarly, we resize those image to 32 $\times$ 32.

Similarly, we choose the nearest neighbor parameter K be 1, 3 and 5. Take 60% training samples from the original dataset, and the other 40% as testing samples. As shown in Figures 10–12, HDA presents the best accuracy rate compared to the other algorithms after 10 runs and HDA is almost the best one under a different number of projection vectors.

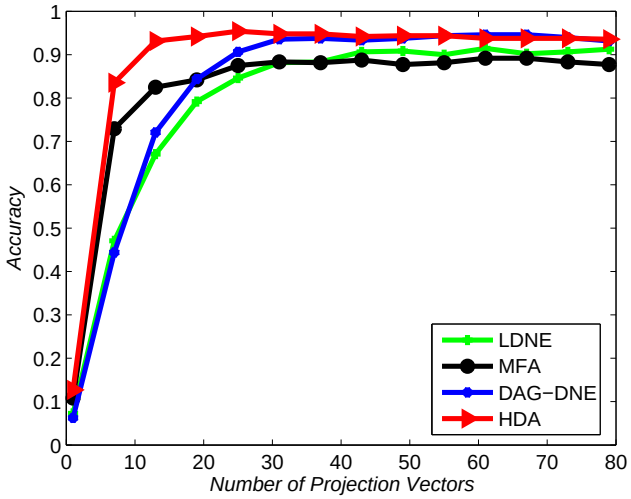

Figure 10. Accuracy vs. Number of Projection Vectors on the Olivetti Research Lab (ORL) dataset under $K = 1$.

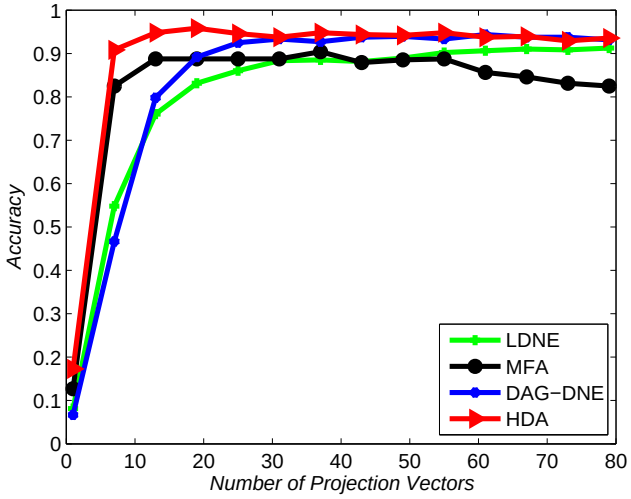

Figure 11. Accuracy vs. Number of Projection Vectors on the ORL dataset under $K = 3$.

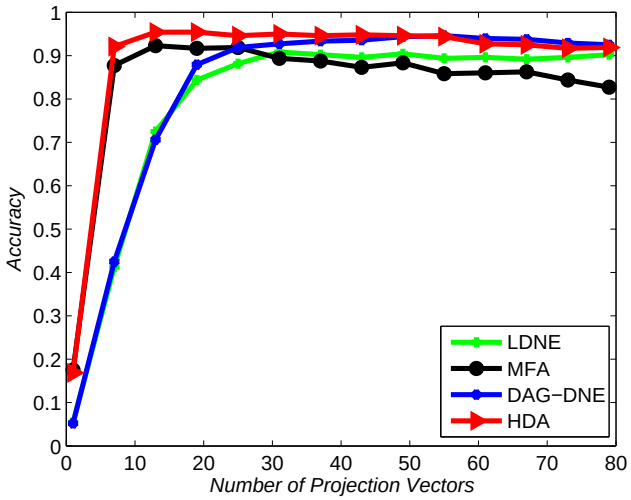

Figure 12. Accuracy vs. Number of Projection Vectors on the ORL dataset under $K = 5$.

Table 1 shows the comparison of average recognition accuracy for different algorithms and different K values. HDA has higher recognition accuracy than the other three algorithms. It improves the classification performance through a dimensional reduction method.

Table 1. Performance comparison of the algorithms on three datasets with different numbers of K.

Algorithm/Result	Yale			ORL			UMIST		
	$K = 1$	$K = 3$	$K = 5$	$K = 1$	$K = 3$	$K = 5$	$K = 1$	$K = 3$	$K = 5$
8-10 LDNE	0.7733	0.7600	0.7067	0.9208	0.9146	0.9208	0.7369	0.7480	0.7775
MFA	0.7600	0.7378	0.7244	0.8937	0.9042	0.9229	0.8034	0.7701	0.7627
DAG-DNE	0.8000	0.7911	0.7378	0.9500	0.9437	0.9521	0.8869	0.8699	0.8655
HDA	0.8444	0.8400	0.8178	0.9542	0.9708	0.9583	0.8973	0.9106	0.9061

5. Conclusions

The collection of a huge amount of high-dimensional sensor intelligent data from the Internet of Things (IoT) causes an information overload problem, so the importance of this research is even more prominent. In this paper, a novel algorithm named hierarchical discriminant analysis is proposed, which aims to find a good latent subspace and preserves the intrinsic structure of the high-dimensional intelligent data. Our proposed algorithm constructs two adjacency graphs to preserve the local structure and deal with optimization problems separately. The experimental results show that HDA is more effective than MFA, LDNE and DAG-DNE on three real image datasets. In other words, hierarchical discriminant analysis can generate a good discriminant subspace. However, HDA is still a linear algorithm, so future work will focus on extending it to be nonlinear to improve the classification performance in cloud computing environments [17–19].

Acknowledgments: This work was supported by the National Science Foundation of China (61472047).

Author Contributions: Chuntao Ding and Shangguang Wang conceived and designed the experiments; Di Lu performed the experiments, analyzed the data. Di Lu and Jinliang Xu wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Turk, M.A.; Pentland, A.P. Face recognition using eigenfaces. In Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Maui, HI, USA, 3–6 June 1991; pp. 586–591.
2. He, X.; Cai, D.; Han, J. Learning a Maximum Margin Subspace for Image Retrieval. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 189–201.
3. Yan, S.C.; Xu, D.; Zhang, B.Y.; Zhang, H.J. Graph embedding: A general framework for dimensionality reduction. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 830–837.
4. Martinez, A.M.; Kak, A.C. PCA versus LDA. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 228–233.
5. Xu, Y.; Zhong, A.N.; Yang, J.; Zhang, D. LPP solution schemes for use with face recognition. *Pattern Recognit.* **2010**, *43*, 4165–4176.
6. He, X.F.; Cai, D.; Yan, S.C.; Zhang, H.J. Neighborhood preserving embedding. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Beijing, China, 17–21 October 2005; Volume 1, pp. 1208–1213.
7. Wong, W.K.; Zhao, H.T. Supervised optimal locality preserving projection. *Pattern Recognit.* **2012**, *45*, 186–197.
8. Gui, J.; Sun, Z.N.; Jia, W.; Hu, R.X.; Lei, Y.K.; Ji, S.W. Discriminant sparse neighborhood preserving embedding for face recognition. *Pattern Recognit.* **2012**, *45*, 2884–2893.
9. Huang, K.K.; Dai, D.Q.; Ren, C.X. Regularized coplanar discriminant analysis for dimensionality reduction. *Pattern Recognit.* **2017**, *62*, 87–98.
10. Xu, D.; Yan, S.; Tao, D.; Lin, S.; Zhang, H.J. Marginal Fisher Analysis and Its Variants for Human Gait Recognition and Content- Based Image Retrieval. *IEEE Trans. Image Process.* **2007**, *19*, 2811–2821.
11. Zhang, W.; Xue, X.Y.; Lu, H.; Guo, Y.F. Discriminant neighborhood embedding for classification. *Pattern Recognit.* **2006**, *39*, 2240–2243.
12. Gou, J.P.; Zhang, Y. Locality-Based Discriminant Neighborhood Embedding. *Comput. J.* **2013**, *56*, 1063–1082.
13. Ding, C.T.; Zhang, L. Double adjacency graphs-based discriminant neighborhood embedding. *Pattern Recognit.* **2015**, *48*, 1734–1742.
14. Roweis, S.T.; Saul, L.K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290*, 2323.
15. Kokiopoulou, E.; Saad, Y. Orthogonal Neighborhood Preserving Projections: A Projection-Based Dimensionality Reduction Technique. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2143–2156.
16. Yan, S.; Xu, D.; Zhang, B.; Zhang, H.J.; Yang, Q.; Lin, S. Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 40–51.
17. Wang, S.; Zhou, A.; Hsu, C.; Xiao, X.; Yang, F. Provision of Data-intensive Services through Energy- and QoS-aware Virtual Machine Placement in National Cloud Data centers. *IEEE Trans. Emerg. Top. Comput.* **2016**, *4*, 290–300.
18. Zhou, A.; Wang, S.G.; Li, J.L.; Sun, Q.B.; Yang, F.C. Optimal mobile device selection for mobile cloud service providing. *J. Supercomput.* **2016**, *72*, 3222–3235.
19. Lei, T.; Wang, S.G.; Li, J.L.; Yang, F.C. AOM: Adaptive Mobile Data Traffic Offloading for M2M Networks. *Pers. Ubiquitous Comput.* **2016**, *20*, 863–873.