International Journal of Pattern Recognition and Artificial Intelligence

NEAREST NEIGHBOR DISCRIMINANT ANALYSIS

XIPENG QIU
Media Computing & Web Intelligence Lab
Department of Computer Science and Engineering, Fudan University
No. 220, Handan Road, Shanghai, 200433, China
xpqiu@fudan.edu.cn

LIDE WU
Media Computing & Web Intelligence Lab
Department of Computer Science and Engineering, Fudan University
No. 220, Handan Road, Shanghai, 200433, China
ldwu@fudan.edu.cn

Linear Discriminant Analysis (LDA) is a popular feature extraction technique in statistical pattern recognition. However, it often suffers from the small sample size problem when dealing with high dimensional data. Moreover, while LDA is guaranteed to find the best directions when each class has a Gaussian density with a common covariance matrix, it can fail if the class densities are more general. In this paper, a novel nonparametric linear feature extraction method, nearest neighbor discriminant analysis(NNDA), is proposed from the view of the nearest neighbor classification. NNDA finds the important discriminant directions without assuming the class densities belong to any particular parametric family. It does not depend on the nonsingularity of the within-class scatter matrix either. Then we give an approximate approach to optimize NNDA and an extension to k-NN. We apply NNDA to the simulated data and real world data, the results demonstrate that NNDA outperforms the existing variant LDA methods.

Keywords: Nearest Neighbor Discriminant Analysis; Linear Discriminant Analysis; Nonparametric Discriminant Analysis

1. Introduction

The curse of high-dimensionality is a major cause of the practical limitations of many pattern recognition technologies, such as text classification and object recognition. In the past several decades, many dimensionality reduction techniques have been proposed. Linear discriminant analysis (LDA)[5] is one of the most popular supervised methods for linear feature extraction and dimensionality reduction. The purpose of LDA is to maximize the between-class scatter S_b while simultaneously minimizing the within-class scatter S_w. In many applications, LDA has been proven to be very powerful.

A major drawback of LDA is that it often suffers from the small sample size problem when dealing with the high dimensional data. When there are not enough

training samples, S_w may become singular, and it is difficult to compute the LDA vectors. For example, a 100×100 image data has 10000 dimensions, which requires more than 10000 training data to ensure that S_w is nonsingular. Several approaches[12,1,3,24,18] have been proposed to address this problem. A common problem with all these proposed variant LDA approaches is that they all lose some discriminative information more or less in the preprocessing stage.

Another disadvantage of LDA is that it assumes each class has a Gaussian density with a common covariance matrix. LDA guarantees to find the best projection directions when the distributions are unimodal and separated by the scatter of class means. However, if the class distributions are multimodal and share the same mean, it fails to find the discriminant direction[5]. Besides, the rank of S_b is $c-1$, where c is the number of class. So the number of extracted features is, at most, $c-1$. However, unless the posteriori probability functions are selected, $c-1$ features are suboptimal in Bayes sense, although they are optimal with regard to Fisher criterion[5].

In this paper, a new feature extraction method, stepwise nearest neighbor discriminant analysis (NNDA), is proposed. NNDA is a linear feature extraction method from the point of view of nearest neighbor classification (NN). Nearest neighbor classification [4] is an efficient method for performing nonparametric classification and often is used in the pattern classification field. Moreover, NN classifier is closely related to Bayes classifier. However, when nearest neighbor classification is carried out in a high-dimensional feature space, the nearest neighbors of a point can be very far away, causing bias and degrading the performance [10]. Hastie and Tibshirani[9] proposed a discriminant adaptive nearest neighbor (DANN) metric to stretch the neighborhood in the directions in which the class probabilities do not change much, but their method also suffers from the small sample size problem and is unable to deal with the high dimensional data.

NNDA can be regarded as an extension of nonparametric discriminant analysis[6], but it does not depend on the nonsingularity of the within-class scatter matrix. Moreover, NNDA finds the important discriminant directions without assuming the class densities belong to any particular parametric family.

There are some works related to our method. Zhang and Chen[25] proposed a classification algorithm based on the concept of symmetric maximized minimal distance in subspace (SMMS), but they just deal with the two-class problem intrinsically and first project the samples to null space of one-class samples, which is not fit for low dimensional data and multi-class problems. Liu[13] proposed a stochastic gradient algorithm for finding optimal linear representations of images and formulating an optimization problem on the Grassmann manifold. Goldberger[8] proposed a method for learning a Mahalanobis distance measure to be used in the KNN classification algorithm. The algorithm directly maximizes a stochastic variant of the leave-one-out KNN score on the training set. But they need consider all pairwise distances of samples in same class, in fact it is unnecessary for nearest neighbor classification. Gilad-Bachrach[7] introduced a margin based feature selection criterion and apply it to measure the quality of sets of features, whose objective function is similar with

our criterion. However, all the above methods used the iterative gradient descent based optimization. When the number or dimensionality of samples are high, those methods have low computational efficiency. Moreover, since the cost functions are not convex, those methods cannot guarantee global optima. Our method can be solved by eigen decomposition of a matrix and avoids these drawbacks.

The rest of the paper is organized as follows: Section 2 gives the review and analysis of the current existing variant LDA methods. Then we describe stepwise nearest neighbor discriminant analysis in Section 3, an extending to k-NN is also described. Experimental evaluations of our method and existing variant LDA methods are presented in Section 4. Finally, we give the conclusions in Section 5.

2. Review and Analysis of Variant LDA Methods

The purpose of LDA is to maximize the between-class scatter while simultaneously minimizing the within-class scatter.

The between-class scatter matrix S_b and the within-class scatter matrix S_w are defined as

$$S_b = \sum_{i=1}^{c} p_i (m_i - m)(m_i - m)^T \tag{1}$$

$$S_w = \sum_{i=1}^{c} p_i S_i, \tag{2}$$

where c is the number of classes; m_i and p_i are the mean vector and a priori probability of class i, respectively; $m = \sum_{i=1}^{c} p_i m_i$ is the global mean vector; S_i is the covariance matrix of class i.

The LDA method tries to find a set of projection vectors $W \in R^{D \times d}$ maximizing the ratio of determinant of S_b to S_w,

$$W = \arg\max_{W} \frac{|W^T S_b W|}{|W^T S_w W|}, \tag{3}$$

where D and d are the dimensionalities of the data before and after the transformation respectively.

From Eq.(3), the transformation matrix W must be constituted by the d eigenvectors of $S_w^{-1} S_b$ corresponding to its first d largest eigenvalues[5].

However, when the small sample size problem occurs, S_w becomes singular and S_w^{-1} does not exist. Moreover, if the class distributions are multimodal or share the same mean, it can fail to find the discriminant direction[5]. Many methods have been proposed for solving the above problems. In the following subsections, we give more detailed review and analysis of these methods.

2.1. *Methods Aimed at Singularity of S_w*

In recent years, many researchers have noticed the problem about singularity of S_w and tried to overcome the computational difficulty with LDA.

To avoid the singularity of S_w, a two-stage PCA+LDA approach is used[1]. Principal component analysis[5] (PCA) is first used to project the high dimensional data into a low dimensional feature space. PCA aims to maximize the covariance of all data regardless of the class information and is very useful in reducing noise in the data. Its goal is to find a set of mutually orthogonal basis functions that capture the directions of maximum variance in the data and for which the coefficients are pairwise decorrelated. Then LDA is performed in the reduced PCA subspace, in which S_w is non-singular. But this method is obviously suboptimal due to discarding much discriminative information.

Liu *et al.*[12] modified Fisher's criterion by using the total scatter matrix $S_t = S_b + S_w$ as the denominator instead of S_w. It has been proven that the modified criterion is exactly equivalent to Fisher criterion. However, when S_w is singular, the modified criterion reaches the maximum value, namely 1, for any transformation W in the null space of S_w. Thus the transformation W cannot guarantee the maximum class separability $|W^T S_b W|$ is maximized. Besides, this method still needs calculate an inverse matrix, which is time consuming. Chen *et al.*[3] suggested that the null space spanned by the eigenvectors of S_w with zero eigenvalues contains the most discriminative information. A LDA method (called NLDA)in the null space of S_w was proposed. It chooses the projection vectors maximizing S_b with the constraint that S_w is zero. But this approach discards the discriminative information outside the null space of S_w. However, the null space of S_w probably contains no discriminant information and could not avoid overfit. Thus, it is obviously suboptimal because it maximizes the between-class scatter in the null space of S_w instead of the original input space. Besides, the performance of the NLDA drops significantly when $N - c$ is close to the dimension D, where N is the number of samples and c is the number of classes. The reason is that the dimensionality of the null space is too small in this situation and too much information is lost[11]. Yu *et al.*[24]proposed a direct LDA (DLDA) algorithm, which first removes the null space of S_b. They assume that no discriminative information exists in this space. Unfortunately, it can be shown that this assumption is incorrect because the optimal discriminant vectors do not necessarily lie in the subspace spanned by the class centers[22].

2.2. *Methods Aimed at Limitations of S_b*

When the class conditional densities are multimodal, the class separability represented by S_b is poor. Especially in the case that each class shares the same mean, it fails to find the discriminant direction because there is no scatter of the class means[5].

Notice the rank of S_b is $c - 1$, so the number of extracted features is, at most, $c-1$. However, unless the posteriori probability functions are selected, $c-1$ features are suboptimal in Bayes sense, although they are optimal with regard to Fisher criterion[5].

In fact, if classification is the ultimate goal, we only need to estimate the class

density well near the decision boundary[10].

Fukunaga and Mantock[6] presented a nonparametric discriminant analysis in an attempt to overcome these limitations presented in LDA. In nonparametric discriminant analysis the between-class scatter S_b is of nonparametric nature. This scatter matrix generally is of full rank, thus loosening the bound on extracted feature dimensionality. Also, the nonparametric structure of this matrix inherently leads to the extracted features that preserve relevant structures for classification. Bressan *et al.*[2] explored the nexus between nonparametric discriminant analysis (NDA) and the nearest neighbors (NN) classifier and gave a slight modification of NDA which extends the two-class NDA to a multi-class version.

Although these nonparametric methods overcome the limitations of S_b, they still depend on the nonsingularity of S_w. The rank of S_w must be no more than $N - c$.

3. Nearest Neighbor Discriminant Analysis

In this section, we describe a new feature extraction method, nearest neighbor discriminant analysis (NNDA). NNDA uses nonparametric between-class and within-class scatter matrix and it does not depend on the nonsingularity of the within-class scatter matrix. We first propose a nearest neighbor discriminant analysis criterion, then a stepwise dimensionality reduction process is presented.

3.1. *Nearest Neighbor Discriminant Analysis Criterion*

Assuming a multi-class problem with classes $\omega_i (i = 1, \ldots, c)$, we define the extra-class nearest neighbor of a sample $x_n \in \omega_i$ as

$$x_n^E = \arg\min_z ||z - x_n||, \forall z \notin \omega_i. \tag{4}$$

The intra-class nearest neighbor of the sample $x \in \omega_i$ is defined as

$$x_n^I = \arg\min_z ||z - x_n||, \forall z \in \omega_i, z \neq x_n. \tag{5}$$

Then the nonparametric extra-class and intra-class differences are defined as

$$\Delta_n^E = x_n - x_n^E, \tag{6}$$

$$\Delta_n^I = x_n - x_n^I. \tag{7}$$

.

The nonparametric between-class and within-class scatter matrix are defined as

$$\hat{S}_b = \sum_{n=1}^{N} w_n (\Delta_n^E)(\Delta_n^E)^T, \tag{8}$$

$$\hat{S}_w = \sum_{n=1}^{N} w_n (\Delta_n^I)(\Delta_n^I)^T, \tag{9}$$

where w_n is the sample weight defined as

$$w_n = \frac{||\Delta_n^I||^\alpha}{||\Delta_n^I||^\alpha + ||\Delta_n^E||^\alpha}, \tag{10}$$

where α is a control parameter between zero and infinity.

This sample weight is introduced to deemphasize the samples in the class center and give emphasis to the samples near to the other class. The sample that has a larger ratio between the nonparametric extra-class and intra-class differences is given a light influence on the scatter matrix. The sample weights in Eq.(10) take values close to 0.5 near the classification boundaries and drop to zero as we move toward class center. The control parameter α adjusts how fast this happens, and can be chosen by cross-validation.

From Eq.(6) and (7), we can see that $||\Delta_n^E||$ represents the distance between the sample x_n and its nearest neighbor in the different classes, and $||\Delta_n^I||$ represents the distance between the sample x_n and its nearest neighbor in the same class. Given a training sample x_n, the accuracy of the nearest neighbor classification can be directly computed as the difference

$$\Theta_n = ||\Delta_n^E||^2 - ||\Delta_n^I||^2, \tag{11}$$

where Δ^E and Δ^I are nonparametric extra-class and intra-class differences and defined in Eq.(6) and (7).

If the difference Θ_n is positive, x_n will be correctly classified. Otherwise, x_n will be classified wrongly. The larger the difference Θ_n is, the more accurately the sample x_n is classified.

Assuming that we extract features by the $D \times d$ linear projection matrix W, the projected sample $x^{new} = W^T x$. The projected nonparametric extra-class and intra-class differences can be written as $\delta^E = W^T \Delta^E$ and $\delta^I = W^T \Delta^I$. So we expect to find the optimal W to make the difference $||\delta_n^E||^2 - ||\delta_n^I||^2$ in the projected subspace as large as possible.

$$\widehat{W} = \arg\max_W \sum_{n=1}^{N} w_n(||\delta_n^E||^2 - ||\delta_n^I||^2). \tag{12}$$

This optimization problem can be interpreted as: find the linear transform that maximizes the distance between classes, while minimizing the expected distance among the samples of a single class.

Considering that,

$$\begin{aligned}
&\sum_{n=1}^{N} w_n(||\delta_n^E||^2 - ||\delta_n^I||^2) \\
&= \sum_{n=1}^{N} w_n(W^T\Delta_n^E)^T(W^T\Delta_n^E) - \sum_{n=1}^{N} w_n(W^T\Delta_n^I)^T(W^T\Delta_n^I) \\
&= tr(\sum_{n=1}^{N} w_n(W^T\Delta_n^E)(W^T\Delta_n^E)^T) - tr(\sum_{n=1}^{N} w_n(W^T\Delta_n^I)(W^T\Delta_n^I)^T)
\end{aligned}$$

$$
\begin{aligned}
&= tr(W^T(\sum_{n=1}^{N} w_n \Delta_n^E (\Delta_n^E)^T)W) - tr(W^T(\sum_{n=1}^{N} w_n \Delta_n^I (\Delta_n^I)^T)W) \\
&= tr(W^T \hat{S}_b W) - tr(W^T \hat{S}_w W) \\
&= tr(W^T(\hat{S}_b - \hat{S}_w)W), \qquad (13)
\end{aligned}
$$

where $tr(\cdot)$ is the trace of matrix, $\hat{S}_b$ and $\hat{S}_w$ are the nonparametric between-class and within-class scatter matrix, as defined in Eq.(8) and (9).

So Eq.(12) is equivalent to

$$
\widehat{W} = \arg\max_{W} tr(W^T(\hat{S}_b - \hat{S}_w)W) \qquad (14)
$$
$$
\texttt{subject to } W^T W = 1.
$$

We call Eq.(14) the *nearest neighbor discriminant analysis criterion(NNDA).*

The projection matrix $\widehat{W}$ must be constituted by the d eigenvectors of $(\hat{S}_b - \hat{S}_w)$ corresponding to its d largest eigenvalues.

Fig. 1 gives comparisons among NNDA, LDA and PCA on four simulated dataset. The data of each class are generate by sampling from single or multiple Gaussian distributions. When the class density is unimodal ((a)), NNDA is approximately equivalent to LDA. But in the cases where the class density is multimodal or that all the classes share the same mean ((b),(c) and (d)), NNDA outperforms LDA greatly.

We also compare the performance of NNDA with such as PCA, LDA and NCA[8] for more complicated data sets. The visualization results is shown in Fig.2. The leave-one-out classification accuracies of nearest neighbor classifier are given for the 2-D data. The three data sets are generated from gaussian distributions, spirals and circles in 3-D space, then random noises are added to each dimensionality of data. In NCA, the maximum number of iterations is set to 500. In NNDA, the weight parameter α in Eq.10 is set to 0.

3.2. *Discussions*

NNDA has an advantage that there is no need to calculate the inverse matrix, so it is a more efficient and stable method.

However, a drawback of NNDA is the computational inefficiency in finding the neighbors when the original data are high dimensional. An alleviative method is that PCA is first used to reduce the dimension of data to $N-1$ (the rank of the total scatter matrix) through removing the null space of the total scatter matrix. Then, NNDA is performed in the transformed space. Yang *et al.*[23] shows that no discriminant information is lost in this transformed space. Another drawback of NNDA is that it is time-consuming in the training procedure due to the stepwise dimensionality reduction process.

However, once the final transform matrix $\widehat{W}$ is found, it is unnecessary to perform the stepwise dimensionality reduction process to unknown test samples. Thus, NNDA is as efficient as the traditional LDA methods in the test phase.

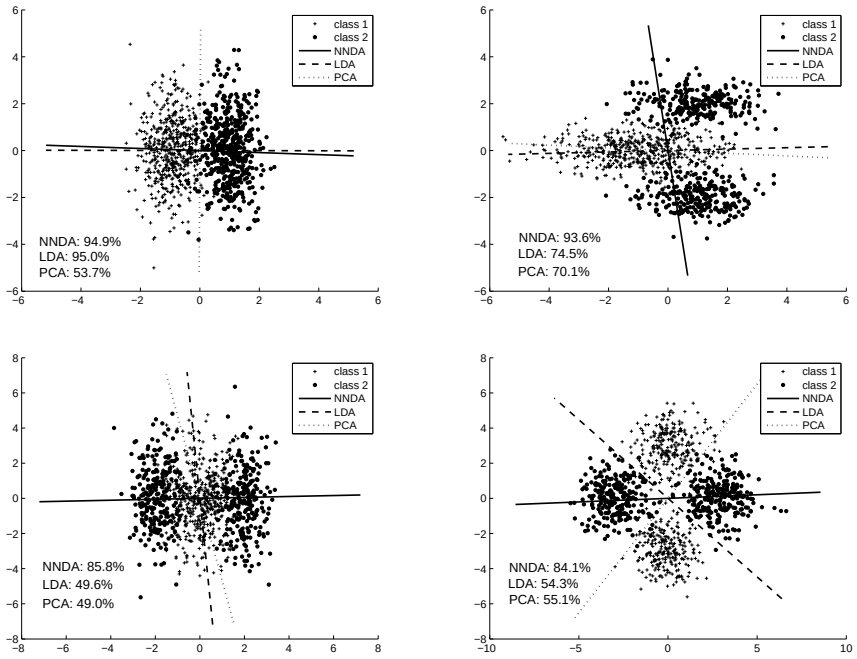


Fig. 1. First projected directions of NNDA, LDA and PCA on four artificial datasets. The leave-one-out accuracy rates are also given in 1-D projected subspace by NN classification. Top Left:One gaussian distribution for each of two classes; Top Right: One class is sampled from two gaussian mixtures; Bottom Left: One class is between two gaussian components of another class; Bottom Right: Two gaussian mixtures for each of two classes.

3.3. *Stepwise Dimensionality Reduction*

Although Fig. 1 and 2 give nice demonstrations for NNDA, there is still a potential risk. In the analysis of the nearest neighbor discriminant analysis criterion, notice that we calculate nonparametric extra-class and intra-class differences (Δ^E and Δ^I) in the original high dimensional space, then project them to the low dimensional space ($\delta^E = W^T\Delta^E$ and $\delta^I = W^T\Delta^I$). However, except for the orthonormal transformation case, we cannot guarantee that δ^E and δ^I agree exactly with the nonparametric extra-class and intra-class differences in projection subspace. A solution for this problem is to find the projection matrix $\widehat{W}$ by stepwise dimensionality reduction method. In each step, we re-calculate the nonparametric extra-class and intra-class differences in its current dimensionality. Thus, we keep the consistency of the nonparametric extra-class and intra-class differences in the process of dimensionality reduction.

Figure 3 gives the algorithm of stepwise nearest neighbor discriminant analysis.

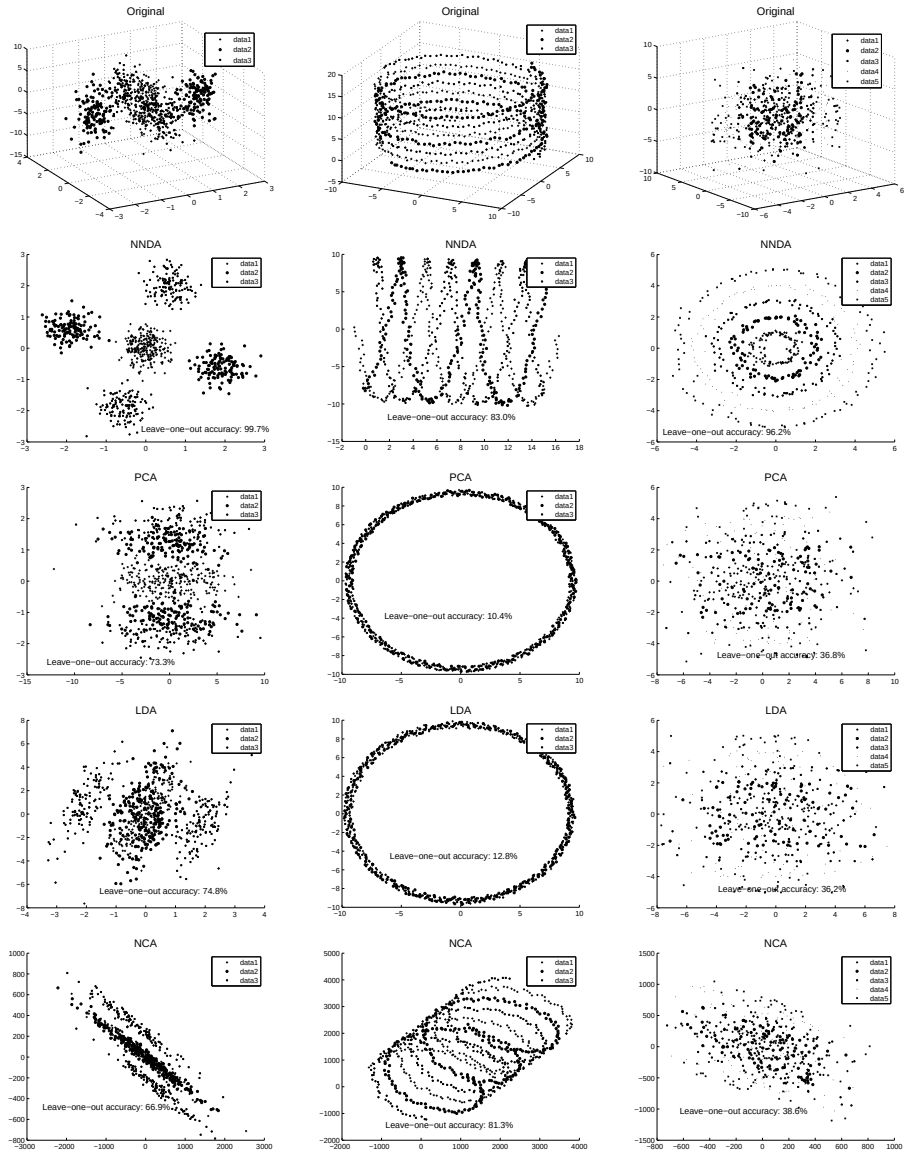


Fig. 2. 2-D visualization of NNDA, PCA, LDA and NCA applied to three synthetic datasets. The top row shows the three original 3-D datasets. The leave-one-out accuracy rates are also given in 2-D projected subspace by NN classification.

3.4. *Extending NNDA from* 1*-NN to* k*-NN*

Though NNDA optimizes the 1-NN classification, it is easy to extend it to the case of k-NN. Given the training sample $x \in \omega_i$ and its k nearest neighbors, if the

- Given D dimensional samples $\{x_1, \cdots, x_N\}$, we expect to find d dimensional discriminant subspace.
- Suppose that we find the projection matrix $\widehat{W}$ via T steps, we reduce the dimensionality of samples to d_t in step t, and d_t meet the conditions: $d_{t-1} > d_t > d_{t+1}$, $d_0 = D$ and $d_T = d$.
- For $t = 1, \cdots, T$
  (1) Calculate the nonparametric between-class $\hat{S}_b^t$ and within-class scatter matrix $\hat{S}_w^t$ in the current d_{t-1} dimensional space;
  (2) Calculate the projection matrix $\widehat{W}_t$; $\widehat{W}_t$ is $d_{t-1} \times d_t$ matrix.
  (3) Project the samples by the projection matrix $\widehat{W}_t$, $x = \widehat{W}_t^T \times x$.
- The final transformation matrix $\widehat{W} = \prod_{t=1}^{T} \widehat{W}_t$.

Fig. 3. Stepwise Nearest Neighbor Discriminant Analysis

amount of neighbors which belong to ω_i is larger than the amount of neighbors which belong to any other class, k-NN classifier will give x the correct class label.

We give a stricter criterion for k-NN. If the majority (no less than $[k/2]+1$) of its k nearest neighbors belong to ω_i, x will classified correctly by k-NN, where $[k/2]$ represents the largest one among the integers that are less than $k/2$. In this criterion, we can extend NNDA of 1-NN to k-NN easily.

Equivalently, we define $x^E_{[k/2]}$ as the extra-class $[k/2]$-th nearest neighbor for the sample x. The intra-class $([k/2]+1)$-th nearest neighbor of the sample x is defined as $x^I_{([k/2]+1)}$. If the distance from x to $x^I_{([k/2]+1)}$ is less than the distance from x to $x^E_{[k/2]}$, the majority $([k/2]+1)$ of its k nearest neighbors must belong to ω_i, and x will be classified correctly by k-NN classifier accordingly.

We rewrite the nonparametric extra-class and intra-class differences in Eq. (6) and (7) as follow:

$$\Delta^E = x - x^E_{[k/2]}, \tag{15}$$

$$\Delta^I = x - x^I_{([k/2]+1)}. \tag{16}$$

.

Thus, the accuracy of the k-NN classification can be estimated by examining the difference

$$\Theta_n = ||\Delta^E_n||^2 - ||\Delta^I_n||^2, \tag{17}$$

where Δ^E and Δ^I are nonparametric extra-class and intra-class differences and defined in Eq.(15) and (16). The larger the difference Θ_n is, the more accurately the sample x_n is classified by k-NN classifier.

The remainder analysis is the same as NNDA of 1-NN. Thus, we extend NNDA from 1-NN to k-NN just through replacing Eq. (6) and (7) by Eq.(15) and (16)

respectively.

A more stricter extension can be found in [17,19].

4. Experiments

In this section, we apply NNDA to recognize handwritten digital character and faces, and compare it with the existing variant LDA methods, such as PCA[21],MMC[11], PCA+LDA[1], NLDA[3], NDA[2] and Bayesian[15] approaches.

All the following experiments are repeated 5 times independently and the average results are calculated. The classifier is nearest neighbor classifier. In NNDA, the stepwise dimensionality reduction process is performed by 5 steps in our experiments. The intervals of dimensionality reduction are same in every step.

4.1. *Handwritten Digital Character Recognition*

The handwritten Digital character dataset[14] contains the binary 20×16 digits of "0" through "9" and capital "A" through "Z". There are 39 examples of each class. Since the dimensionality of data is relative low and null space of the within-scatter matrix does not exist, NLDA fails to deal with this dataset. In NNDA, we set the weight parameter α in Eq. 10 is 0.

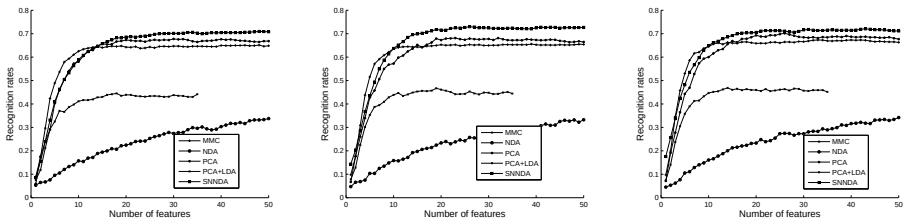


Fig. 4. Recognition rates with the different number of features on the handwritten digits character dataset by kNN classification. (Left: k=1; Middle: k =3; Right: k=5). SNNDA means NNDA with stepwise dimensionality reduction process.

Fig.4 gives the results of k-NN on handwritten digits character dataset by different methods. NNDA is better than other methods on all the three classifiers.

4.2. *Face Recognition*

To evaluate the robustness of NNDA for the high-dimensional data, we perform the experiments on three datasets from the popular ATT face image database[20] and FERET face database[16]. The images are formulated in the vector representation. The descriptions of the three datasets are below:

ATT Dataset This dataset is the ATT face database (formerly 'The ORL Database of Faces'), which contains 400 images (112×92) of 40 persons, 10

images per person. The images are taken at different times, varying lighting slightly, facial expressions (open/closed eyes, smiling/non-smiling) and facial details (glasses/no-glasses). All the images are taken against a dark homogeneous background. The faces are in up-right position of frontal view, with slight left-right out-of-plane rotation. Each image is linearly stretched to the full range of pixel values of $[0, 255]$. The set of the 10 images for each person is randomly partitioned into a training subset of 5 images and a test set of the other 5.

FERET Dataset 1 This dataset is a subset of the FERET database with 194 subjects only. Each subject has 3 images: (a) one taken under controlled lighting condition with a neutral expression; (b) one taken under the same lighting condition as above but with different facial expressions (mostly smiling); and (c) one taken under different lighting condition and mostly with a neutral expression. All images are pre-processed using zero-mean-unit-variance operation and manually registered using the eye positions. All the images are normalized by the eye locations and are cropped to the size of 75×65. A mask template is used to remove the background and the hair. Histogram equalization is applied to the face images for photometric normalization. Two images for each person are randomly selected for training and the rest one is used for test.

FERET Dataset 2 This dataset is a different subset of the FERET database. All the 1195 people from the FERET Fa/Fb data set are used in the experiment. There are two face images for each person. This dataset has no overlap between the training set and the galley/probe set according to the FERET protocol[16]. 500 people are randomly selected for training, and the remaining 695 people are used for testing. For each testing people, one face image is in the gallery and the other is for probe. All images are pre-processed by using the same method in FERET Dataset 1.

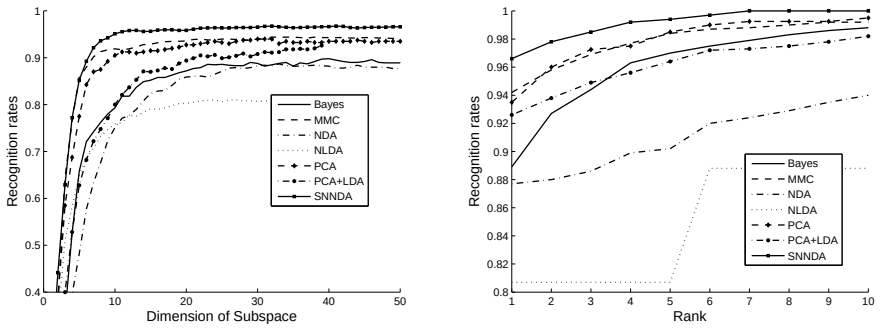


Fig. 5. Rank-1 (left) and cumulative (right) recognition rates with the different number of features on the ATT dataset. In cumulative recognition rates, the number of features is 39.

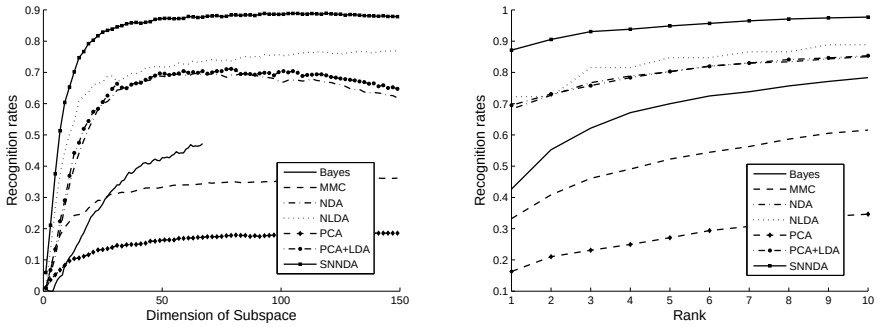


Fig. 6. Rank-1 (left) and cumulative (right) recognition rates with the different number of features on the FERET dataset 1. In cumulative recognition rates, the number of features is 60.

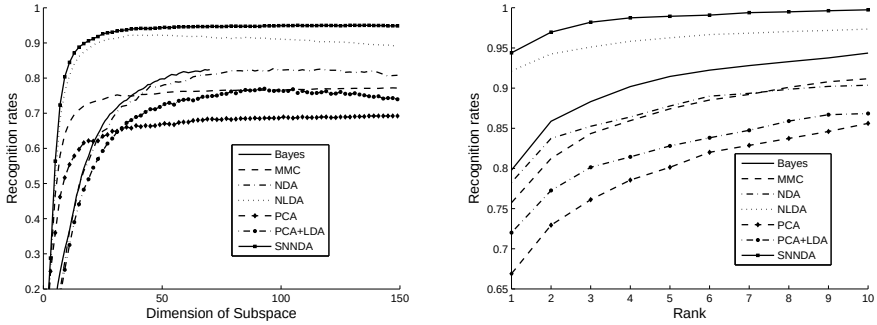


Fig. 7. Rank-1 (left) and cumulative (right) recognition rates with the different number of features on the FERET dataset 2. In cumulative recognition rates, the number of features is 60.

Fig. 5, 6 and 7 show the rank-1 and cumulative recognition rates[16] with the different number of features on the three different datasets. Note that SNNDA means NNDA with stepwise dimensionality reduction process. In NNDA, we set the weight parameter α in Eq. 10 is 6.

It is shown that NNDA outperforms the other methods. Moreover, NNDA does not suffer from the overfitting. Except NNDA and PCA, the rank-1 recognition rates of the other methods have a descent when the dimensionality increases.

When dataset contains noise (the changes of lighting condition in FERET Dataset 1), NNDA also has obviously better performance than the others.

Different from ATT dataset and FERET dataset 1, where the class labels involved in training and testing are the same, the FERET dataset 2 has no overlap between the training set and the galley/probe set according to the FERET protocol[16]. The ability of generalization from known subjects in the training set to

unknown subjects in the gallery/probe set is needed for each method. Thus, the result on FERET dataset 2 is more important for the robustness of methods. We can see that NNDA also gives the best performance than the other methods on FERET dataset 2.

The major characteristic, shown in the experimental results, is that NNDA always has a stable and high recognition rates on the three different datasets, while the other methods have unstable performances.

5. Conclusion

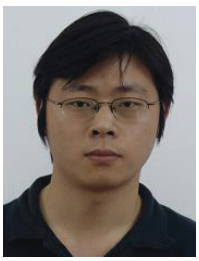
In this paper, we proposed a new feature extraction method, stepwise nearest neighbor discriminant analysis(NNDA), which finds the important discriminant directions without assuming the class densities belong to any particular parametric family. It does not depend on the nonsingularity of the within-class scatter matrix either. We also give an approximate algorithm and an extension to the k-NN case to optimize the performance of NNDA . Our experimental results demonstrate that NNDA outperforms the existing variant LDA methods greatly. Moreover, NNDA is very efficient, accurate and robust. In the further works, we will extend NNDA to non-linear discriminant analysis with the kernel method.

References

1. P. Belhumeur, J. Hespanda, and D. Kiregeman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(7):711–720, 1997.
2. M. Bressan and J. Vitrià. Nonparametric discriminant analysis and nearest neighbor classification. *Pattern Recogn. Lett.*, 24:2743–2749, 2003.
3. L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu. A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recogn.*, 33(10):1713–1726, 2000.
4. R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, New York, 2nd edition, 2001.
5. K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, Boston, 2nd edition, 1990.
6. K. Fukunaga and J. Mantock. Nonparametric discriminant analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 5:671–678, 1983.
7. R. Gilad-Bachrach, A. Navot, and N. Tishby. Margin based feature selection - theory and algorithms. In *Proc. of Int. Conf. on Mach. Learn. (ICML)*, Banff, Canada, 2004.
8. J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Proc. of Neural Inform. Processing Syst. (NIPS)*, 2004.
9. T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Machine Intell.*, 18:607–616, 1996.
10. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.
11. H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. In *Proc. of Neural Inform. Processing Syst. (NIPS)*, 2003.
12. K. Liu, Y. Cheng, and J. Yang. A generalized optimal set of discriminant vectors. *Pattern Recogn.*, 25(7):731–739, 1992.

13. X. Liu, A. Srivastava, and K. Gallivan. Optimal linear representations of images for object recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, 26(5):662–666, 2004.
14. S. Lucas. Handwritten digits. http://www.cs.toronto.edu/ roweis/data.html.
15. B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recogn.*, 33:1771–1782, 2000.
16. P.J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Comput.*, 16(5):295–306, 1998.
17. Xipeng Qiu and Lide Wu. Face recognition by stepwise nonparametric margin maximum criterion. In *Proc. of IEEE Conf. on Comput. Vision (ICCV)*, pages 1567–1572, Beijing, 10 2005.
18. Xipeng Qiu and Lide Wu. Null space-based LDA with weighted dual personal subspaces for face recognition. In *Proc. of IEEE Conf. on Image Processing (ICIP)*, pages 1665–1668, Genoa, 2005.
19. Xipeng Qiu and Lide Wu. Stepwise nearest neighbor discriminant analysis. In *Proc. of Int. Joint Conf. on Artif. Intell. (IJCAI)*, pages 829–834, Edinburgh, 2005.
20. Ferdinando Samaria and Andy Harter. Parameterisation of a stochastic model for human face identification. In *Proc. of IEEE Workshop on Appl. of Comput. Vision*, 1994.
21. M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.
22. X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. In *Proc. of IEEE Conf. on Comput. Vision and Pattern Recogn. (CVPR)*, 2004.
23. J. Yang and J. Yang. Why can LDA be performed in PCA transformed space? *Pattern Recogn.*, 36:563–566, 2003.
24. H. Yu and J. Yang. A direct LDA algorithm for high-dimensional data with application to face recognition. *Pattern Recogn.*, 34:2067–2070, 2001.
25. W. Zhang and T. Chen. Classification based on symmetric maximized minimal distance in subspace (SMMS). In *Proc. of IEEE Conf. on Comput. Vision and Pattern Recogn. (CVPR)*, 2003.

Biographical Sketch and Photo



Xipeng Qiu received the PhD degree in computer science from Fudan University in 2006. Currently, he is an assistant professor at Fudan University, Shanghai, China. His research interests include machine learning, Information Retrieval and Text Mining.

Lide Wu graduated from Fudan University, Shanghai, China, in 1958. He was a Visiting Scholar with Princeton University, Princeton, NJ, in 1980, and with Brown University, Providence, RI, in 1981. He was the Dean of the Department of Computer Science, Fudan University, from 1982 to 1983. He is now a professor with Fudan University. His main research interests are computer vision, natural language processing, and video database systems