Robust Discriminant Analysis Based on Nonparametric Maximum Entropy

Ran He, Bao-Gang Hu, and Xiao-Tong Yuan

National Laboratory of Pattern Recognition,
Institute of Automation,
Chinese Academy of Sciences
95 Zhongguancun Donglu, Beijing 100190, China
{rhe,bghu,xtyuan}@nlpr.ia.ac.cn

Abstract. In this paper, we propose a Robust Discriminant Analysis based on maximum entropy (MaxEnt) criterion (MaxEnt-RDA), which is derived from a nonparametric estimate of Renyi's quadratic entropy. MaxEnt-RDA uses entropy as both objective and constraints; thus the structural information of classes is preserved while information loss is minimized. It is a natural extension of LDA from Gaussian assumption to any distribution assumption. Like LDA, the optimal solution of MaxEnt-RDA can also be solved by an eigen-decomposition method, where feature extraction is achieved by designing two Parzen probability matrices that characterize the within-class variation and the between-class variation respectively. Furthermore, MaxEnt-RDA makes use of high order statistics (entropy) to estimate the probability matrix so that it is robust to outliers. Experiments on toy problem , UCI datasets and face datasets demonstrate the effectiveness of the proposed method with comparison to other state-of-the-art methods.

1 Introduction

Feature extraction plays an important role in machine learning and computer vision. It is a common preprocessing step to learn a low-dimensional subspace from the raw input variables which might be strongly relevant and redundant [1]. From different viewpoints, there are two major categories of feature extraction: unsupervised and supervised.

In unsupervised feature extraction, the data class labels are unknown. The low-dimensional representation is learned by minimizing reconstruction error or preserving structural information of data. A best-known unsupervised method is principal component analysis (PCA) [2]. To deal with the Gaussian assumption problem in PCA, some important extensions of PCA are developed by locally optimizing the weighted scatter matrix, including locality preserving projections (LPP) [3], Laplacian PCA [4] and etc.

In supervised feature extraction [5][6] [7], the information of class labels is used to learn a low-dimensional subspace by maximizing the class differences. The linear discriminant analysis (LDA) [8] is the most representative one. It is a widely used in classification tasks due to its computational simplicity. Despite its wide use, LDA assumes that data distribution of each class is Gaussian. Thus, this will be certainly hard to make LDA adapt to data under non-Gaussian distribution. Among algorithms for solving this problem, nonparametric estimation is a most popularly used technique

Z.-H. Zhou and T. Washio (Eds.): ACML 2009, LNAI 5828, pp. 120–134, 2009.

and nonparametric LDAs have therefore been developed. The main difference between nonparametric LDAs and LDA is in that nonparametric LDAs introduce nonparametric within-class scatter and between-class scatter. Nonparametric LDA (NDA) [9], marginal Fisher analysis (MFA) [10] and linear Laplacian discrimination (LLD) [11] are most representatives of nonparametric methods. Although those nonparametric methods are useful for solving non-Gaussian data, they are often designed based on heuristic strategy and the learned subspace depends on the structure of training set. When there is noise, the learned subspace may be biased.

Another well-known supervised feature extraction category is information theoretic learning (ITL) [12], where the mutual information (or entropy) is selected as the objective. Feature extraction is achieved by directly maximizing the quadratic mutual information between the class label and the features [13]. Gaussian [14] and Parzen window [15] probability density functions are used to estimate mutual information. In [16], feature extraction is implemented by solving a minimum entropy problem (or maximizing the information potential) and an iterative algorithm based on half-quadratic is proposed. The experimental results demonstrate that the methods based on ITL have potential ability to discover principal curve of the data and are robust to noise [17][16]. However, the ITL based methods are often solved by iterative algorithms and hence have relatively high computation complexity.

In this paper, the maximum entropy (MaxEnt) criterion, which provides a natural means to process information in the form of constraints [18], is introduced in linear feature extraction. A MaxEnt Robust Discriminant Analysis (MaxEnt-RDA) is proposed where entropy is defined as the Renyi's quadratic entropy. The MaxEnt distribution is estimated by a nonparametric Parzen window density estimator. As a result, the calculation of differential entropy becomes a summation over pair wise interactions of the data. The proposed MaxEnt-RDA has several interesting perspectives: (1) it utilizes entropy maximum as objective and hence has a clearly theoretical foundation. The high order statistic information of data are preserved during feature extraction. And it is robust to noise. (2) it can be solved by an eigen-decomposition method which avoids iterative calculation of entropy. The optimal solution consists of the principal eigenvectors of two probability matrices corresponding to MaxEnt distribution. (3) Its assumption of distribution is free so that it can effectively capture the underline distribution of multimodal data statistics and deal with non-Gaussian distribution data.

The remainder of this paper is outlined as follows. The concepts of LDA and MFA are briefly reviewed in Section 2. The theoretical properties of MaxEnt objective function and MaxEnt-RDA are described in Section 3. We evaluate our method on UCI machine learning datasets and face datasets in Section 4. Finally, we conclude the paper in Section 5.

2 LDA and MFA

Suppose that we have a matrix $X = [x_1, \ldots, x_n]$ of n d-dimensional samples and a projection matrix $U = [u_1, \ldots, u_m]$ whose columns constitute the bases of the m-dimensional subspace. There are c classes in the data set X and each class C_j has n_j samples. According to definitions in [19], sample mean for all of the classes and for each class can be defined as,

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \mu_j = \frac{1}{n_j}\sum_{x_j \in C_j} x_j \tag{1}$$

and the covariance matrices as:

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T \tag{2}$$

$$\widehat{\Sigma}_j = \frac{1}{n_j}\sum_{x_j \in C_j}(x_j - \mu_j)(x_j - \mu_j)^T \tag{3}$$

We define the scatter matrices S_T, S_W and S_B by,

$$S_T = n\widehat{\Sigma} = X(I - W_t)X^T \tag{4}$$

$$S_W = \sum_{j=1}^{c} n_j \widehat{\Sigma}_j = \sum_{j=1}^{c} X_j(I - W_j)X_j^T = X(I - W_w)X^T \tag{5}$$

$$S_B = S_T - S_W \tag{6}$$

where W_t is a $n \times n$ matrix with all the elements equal to $1/n$, W_j is a $n_j \times n_j$ matrix with all the elements equal to $1/n_j$, and W_w reads the diagonal block form of $W_w = diag(W_1, \ldots, W_c)$. S_W and S_B are often called within-class scatter matrix and between-class scatter matrix respectively. The LDA tries to solve the following maximum problem:

$$J_{LDA}(u) = \frac{u^T S_T u}{u^T S_W u} \sim \frac{u^T S_B u}{u^T S_W u} \tag{7}$$

where u is a d-dimension vector. It is easy to show that the vector u that maximizes J_{LDA} must satisfy

$$S_T u = \lambda S_W u \tag{8}$$

If S_W is nonsingular, the solution of LDA can be obtained by a conventional eigenvalue problem:

$$(X(I - W_w)X^T)^{-1}X(I - W_t)X^T u = \lambda u \tag{9}$$

Since LDA assumes that data distribution of each class is Gaussian, LDA could not adapt to data under non-Gaussian distribution. MFA is an important extension of LDA from viewpoint of Graph Embedding. Two graphs in MFA are designed to characterize the within-class compactness matrix and the between-class separability, respectively. The within-class compactness matrix of MFA is defined as

$$\begin{aligned} S_w^{MFA} &= \sum_{i=1}^{n} \sum_{x_i \in N_{k_1}(x_j)\ or\ x_j \in N_{k_1}(x_i)} ||u^T x_i - u^T x_j||^2 \\ &= 2u^T X(D_w^{MFA} - W_w^{MFA})X^T u \\ (D_w^{MFA})_{ii} &= \sum_{\mathrm{j} \neq i} (W_w^{MFA})_{ij} \\ (W_w^{MFA})_{ij} &= 1 \quad if\ x_j \in N_{k_1}(x_i)\ or\ x_i \in N_{k_1}(x_j); 0\ else \end{aligned} \tag{10}$$

where $N_{k_1}(x_i)$ means the k_1 nearest neighbors in the same class of the sample x_i and D_w^{MFA} is a diagonal matrix. The between-class compactness matrix of MFA is defined as

$$\begin{aligned} S_b^{MFA} &= \sum_{i=1}^{n} \sum_{x_i \in P_{k_1}(x_j)\ or\ x_j \in P_{k_1}(x_i)} ||u^T x_i - u^T x_j||^2 \\ &= 2u^T X (D_b^{MFA} - W_b^{MFA}) X^T u \\ (D_b^{MFA})_{ii} &= \sum_{\mathrm{j} \neq i} (W_b^{MFA})_{ij} \\ (W_w^{MFA})_{ij} &= 1 \quad if\ x_j \in P_{k_2}(x_i)\ or\ x_i \in P_{k_2}(x_j); \quad 0\ else \end{aligned} \tag{11}$$

Where $P_{k_2}(x_i)$ is a set of data pairs that are the k_2 nearest pairs among $\{(x_i, x_j), x_i \in C_i, x_j \notin C_i\}$ and D_b^{MFA} is a diagonal matrix. MFA provides a new approach to deal with Gaussian assumption in LDA and thus can improve accuracy of classification rates in non Guassian distribution applications such as face recognition.

However, a main drawback of MFA is that MFA depends on the structure of the data points near the boundary of different classes. When there is noise, the performance of MFA will decrease (see experiment 4.3).

3 MaxEnt Robust Discriminant Analysis

Renyi's quadratic entropy of a dataset X with Probability Density Function (PDF) $f_X(x)$ is defined by

$$H(X) = -\log \int f_X^2(x) dx \tag{12}$$

If Parzen window method is used to estimate the P.D.F., $f_X(x)$ can be obtained as

$$\widehat{f}_{X;\sigma}(x) = \frac{1}{n} \sum_{i=1}^{n} G(x - x_i, \sigma) \tag{13}$$

where $G(x - x_i, \sigma)$ is the Gaussian kernel with bandwidth σ

$$G(x - x_i, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x - x_i)^2}{2\sigma^2}) \tag{14}$$

Substitute $f_X(x)$ in (12) with (13), the estimate of entropy by Parzen method can be obtained as [13]:

$$H(X) = -\log(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} G(x_j - x_i, \sigma)) \tag{15}$$

In supervised linear feature extraction, one considers the following constraint MaxEnt problem:

$$\max_U H(U^T X) \quad s.t.\ H(U^T X|C) = c_1\ \&\ U^T U = I \tag{16}$$

where conditional entropy $H(X|C)$ is defined as [20]:

$$H(X|C) = \sum_{j=1}^{c} p(C_j) H(X|C = C_j) \tag{17}$$

When the formula of $f_X(x)$ is given, the MaxEnt distribution in (16) is only relative to the subspace U, i.e., the MaxEnt objective is a function of subspace U. The different subspace will give a different system state measured by entropy. We expect that in the subspace U entropy of all data is maximized so that the loss of information is minimized, meanwhile entropy of each class (measured by conditional entropy in (17)) is nearly invariant so that the structure of each class is not broken. After the dimension reduction, the information loss is minimized meanwhile the structural information of individual class is preserved. Since the distribution is estimated by Parzen window method, it can model the data's distribution more accurately. Note that the orthogonal constraint is necessary and important. It is in accord with the idea that the system of coordinates carries no information [18].

Theorem 1. *The optimal solution for (16) is given by the following eigenvector problem*

$$XL_t(u)X^T u = \lambda X L_w(u) X^T u \tag{18}$$

where

$$L_t(u) = D^t(u) - W^t(u) \tag{19}$$

$$L_w(u) = D^w(u) - W^w(u) \tag{20}$$

$$W_{ij}^t(u) = \frac{2G(u^T x_i - u^T x_j, \sigma)}{\sigma^2 \sum_{i=1}^{n}\sum_{j=1}^{n} G(u^T x_i - u^T x_j, \sigma)}$$

$$W_{ij}^w(u) = I_{(c_i = c_j)} p(c_i) \frac{2G(u^T x_i - u^T x_j, \sigma)}{\sigma^2 \sum_{i=1}^{n_i}\sum_{j=1}^{n_i} G(u^T x_i - u^T x_j, \sigma)}$$

$$D_{ii}^t(u) = \sum_{j=1}^{n} W_{ij}^t(u), \quad D_{ii}^w(u) = \sum_{j=1}^{n} W_{ij}^w(u)$$

(D^t and D^w are diagonal matrices)

Proof. We follow the standard theory of numerical optimization and applying the Lagrangian factor on (16) (Here we only consider the first constraint). The PDF $f_X(x)$ is defined in (13), we have:

$$J_H(u) \triangleq -\log \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} G(u^T x_i - u^T x_j, \sigma) + \lambda\Big(\sum_{j=1}^{c} p(C_j) \log \frac{1}{n_j^2} \sum_{k=1}^{n_j}\sum_{l=1}^{n_j} G(u^T x_k - u^T x_l, \sigma) - c_1\Big) \tag{21}$$

where the Lagrangian multiplier λ for enforcing the first constraints. The KKT condition for optimal solution specifies that the gradient of $J_H(u)$ must be zero:

$$\begin{aligned}\frac{\partial J_H(u)}{\partial u} &= \sum_{i=1}^{n}\sum_{j=1}^{n}(W_{ij}^t(u) - \lambda W_{ij}^w(u))((u^T x_i - u^T x_j)x_i^T)^T \\ &= XL_t(u)X^T u - \lambda XL_w(u)X^T u = 0\end{aligned}$$

The above equation gives the fixed point relation

$$XL_t(u)X^T u = \lambda XL_w(u)X^T u$$

Intuitively, the optimal u is the eigenvectors of (18).

In Theorem 1, we can find that the format of solution of MaxEnt is similar to that of Graph Embedding. They all can be solved by eigen-decomposition method. $W_{ij}^w(U)$ can be seen as a within-class scatter matrix, which represents interactions between pairs of samples inside each class. $W_{ij}^t(U)$ consists of interactions between all pairs of samples, regardless of class information. Given a U, $W_{ii}^w(U)$ (or $W_{ii}^t(U)$) is an approximate of probability contribution on x_i under the jth Parzen estimate, and $D_{ii}^w(U)$ (or $D_{ii}^t(U)$) is an approximate of probability value on x_i under the Parzen estimate. Hence, we denote and $W_{ij}^w(U)$ and $W_{ij}^t(U)$ as within-class and between-class Parzen probability matrices respectively.

Recent theoretical results [21][16] illustrate that there is a close relationship between entropy objective and robust estimators [22]. Algorithms based on information theoretic objectives often can significantly improve the robustness [21][16]. For MaxEnt-RDA, if there are outliers that are significantly faraway from the rest of the data points, those outliers will obtain small values in within-class and between-class Parzen probability matrices and hence they will less affect the objective.

Look at (18) in Theorem 1, we recognize that it is still hard to find a closed form for this problem, because this eigenvalue problem is nonlinear. Hence the objective of $J_H(u)$ is still difficult to be directly optimized. In next section, we introduce an approximate strategy for optimization of the problem.

3.1 Algorithm of MaxEnt-RDA

From now on, we derive an algorithm to solve the MaxEnt problem. From theorem 1, we can learn that solution of MaxEnt-RDA can be given by an eigen-decomposition method. However, the eigen-decomposition problem in (18) is nonlinear and the solution is dependent on U in a non-trivial way. The estimate of PDF is performed on the reduced dimension instead of original input feature space. To address this problem, we approximate each Gaussian kernel term by its first-order Taylor expansion. Since expanding $exp(-z)$ at z_0 leads to $exp(-z) \approx exp(-z_0) - exp(-z_0)(z - z_0)$, let $z = \frac{||U^T x_i - U^T x_j||^2}{\sigma^2}$ and $z_0 = \frac{||x_i - x_j||^2}{\sigma^2}$, we have

$$G(U^T x_i - U^T x_j, \sigma) \approx -G(x_i - x_j, \sigma)||U^T x_i - U^T x_j||^2 + const \qquad (22)$$

Substitute (22) into (16), we finally reduce the MaxEnt objective to a constraint graph embedding problem:

$$\max tr(U^T X L_t(I) X^T U) \quad s.t. \;\; tr(U^T X L_w(I) X^T U) = c_1 \;\&\; U^T U = I \tag{23}$$

where we remove the $log(x)$ function because $log(x)$ is a convex function. Since the $XL_t(I)X^T$ and $XL_w(I)X^T$ are symmetric and semi-positive definite, by using the Lagrangian technique, it is easy to prove that the solution is given by

$$X L_t(I) X^T U = \Lambda X L_w(I) X^T U \tag{24}$$

where Λ is a diagonal matrix whose diagonal elements are m largest eigenvalues. Compared with (18) and (24), we can learn that, in the Taylor approximation solution, we assume that the MaxEnt distribution on the subspace U and original input feature are quite similar so that $L_t(U) \approx L_t(I)$ and $L_w(U) \approx L_w(I)$. The merit of this solution based on Taylor expansion is that it is a unique and global solution.

The bandwidth σ is an important parameter in MaxEnt-RDA, which is used in Parzen estimate of $f_X(x)$. The bandwidth controls all properties of the estimator [21]. Considering the theoretical analysis of nonparametric entropy estimators [15], we set the bandwidth as a factor of average distance between projected samples:

$$\sigma^2 = \frac{1}{sn^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2 \tag{25}$$

where s is a scale factor.

The detailed description of MaxEnt-RDA is summarized as Algorithm 1.

```
Input: data matrix $X$, desired dimensionality $m$, bandwidth parameter $\sigma$
Output: orthonormal matrix $U$
  1. Project the data set into the PCA subspace by retaining $n - c$ dimensions.
     Let $W_{PCA}$ denote the transformation matrix of PCA.
  2. Construct the approximated within-class and between-class Parzen probability
     matrices on the PCA subspace according to (20) and (19).
  3. Finding the optimal projection direction $U^*$ by solving (24).
     $U^*$ is given by the eigenvectors corresponding to $m$ largest eigenvalues.
  4. Output the final linear projection direction as $U = W_{PCA} * U^*$
```

Algorithm 1. MaxEnt-RDA

3.2 Relation to Mutual Information and LDA

By Lagrangian factor method, we can rewrite (16) as

$$\max_U J_I(U) = H(U^T X) - \lambda_W (H(U^T X|C) - c_1) \quad s.t. \quad U^T U = I \tag{26}$$

When λ_W is set to 1, $J_I(U)$ becomes the mutual information which has been used as the objective in [15][14].

Theorem 2. *If the data is Gaussian distributed and* $\lambda_W = 1$*, the solution of LDA gives of a lower bound of MaxEnt in (26):*

$$J_I(u) \geq \tfrac{1}{2}\log(J_{LDA}(u)) \tag{27}$$

Proof. Since $\frac{d}{2}\log 2\pi + \frac{d}{2} = \sum\limits_{j=1}^{c} p(C_j)(\frac{d}{2}\log 2\pi + \frac{d}{2})$, we have

$$\begin{aligned}
J_I(u) &= H(u^T X) - H(u^T X|C) \\
&= \frac{1}{2}\log u^T \varSigma u - \frac{1}{2}\sum_{j=1}^{c} p(C_j)\log(u^T \varSigma_j u) \\
&\geq \frac{1}{2}\log u^T \varSigma u - \frac{1}{2}\log(\sum_{j=1}^{c} \frac{n_j}{n} u^T \varSigma_j u) \\
&= \frac{1}{2}\log(\frac{1}{n}u^T S_T u) - \frac{1}{2}\log(\sum_{j=1}^{c} \frac{n_j}{n}(\frac{1}{n_j}u^T S_j u)) \\
&= \frac{1}{2}\log(\frac{\frac{1}{n}u^T S_T u}{\frac{1}{n}u^T S_W u})
\end{aligned}$$

According to the definition of S_B in (6), we have

$$J_I(u) \geq \frac{1}{2}\log(\frac{u^T S_B u}{u^T S_W u}) = \frac{1}{2}\log(J_{LDA})$$

From theorem 2, we can learn that if the data is Gaussian distributed and $\lambda_W = 1$, LDA provides a lower bound of the MaxEnt problem. This theorem also gives a theoretical explanation that LDA subspace is often selected as an initial guess of optimal subspace in gradient ascend method based ITL algorithms.

Theorem 3. *The maximum entropy in (26) is bounded and nonincreasing when* $0 \leq \lambda_W \leq 1$*, i.e.,*

$$0 \leq J_I(U_F^T) \leq J_I(U) \tag{28}$$

where $U \in R^{d\times d}$ *is an orthonormal matrix,* $U_F^T : R^d \rightarrow R^m$ *and* $m < d$*.*

Proof. Since $H(X) \geq H(X|C) \geq \lambda_W H(X|C)$ [20], we have

$$H(U_F^T X) - \lambda_W H(U_F^T X|C) \geq 0 \tag{29}$$

Let $U_B \in R^{d\times(d-m)}$ be a matrix which is the complement subspace of U_F, define the matrix U as

$$U^T X = [U_F^T X \quad U_B^T X], \quad U = [U_F \quad U_B] \tag{30}$$

Since U is orthonormal matrix and H(X) is differential entropy, it follows that $H(X) = H(U^T X)$ [20], we have

$$\begin{aligned}
J_I(U) &= H(X) - \lambda_W H(X|C) \\
&= H(U^T X) - \lambda_W H(U^T X|C) \\
&= H(U_F^T X \, U_{\bar{F}}^T X) - \lambda_W H(U_F^T X \, U_{\bar{F}}^T X|C) \\
&= H(U_F X) + H(U_{\bar{F}} X|U_F X) - \lambda_W H(U_F X|C) - \lambda_W H(U_{\bar{F}} X|U_F X, C) \\
&= H(U_F X) - \lambda_W H(U_F X|C) + (H(U_{\bar{F}} X|U_F X) - \lambda_W H(U_{\bar{F}} X|U_F X, C)) \\
&\geq H(U_F X) - \lambda_W H(U_F X|C)
\end{aligned}$$

Theorem 3 states that the maximum entropy of any orthonormal subspace of the original feature space is bounded and nonincreasing. Furthermore, because the entropy $H(X)$ is a concave function of $f_X(x)$, there is at least a maxima solution of (26).

It is clear that MaxEnt-RDA is a nature extension of LDA from Gaussian distribution assumption to any distribution assumption. Without the prior information on data distributions, the between-class variance of MaxEnt-RDA can better characterize the separability of different classes than the between-class variance as in LDA. Compared with $L_t(U)$ and S_T, we can learn that although the distribution assumptions of MaxEnt-RDA and LDA are different, their solutions have similar matrix format.

According to (26), when λ_W is set to 1, the objective of MaxEnt-RDA becomes the mutual information. Different from previous ITL algorithms, we propose an approximate algorithm to solve the MaxEnt problem. This eigen-decomposition method not only can save computation cost but also has a unique and global solution.

Recently, linear regression based discriminant methods [23][24][25] are proposed in subspace learning for computational convenience. They use multinomial class indicator matrix [26][23] as the regression targets. Theoretical results show that the linear regression based method is equal to LDA in some conditions [24][25]. Since the mean square criterion in linear regression is sensitive to outliers and non-Gaussian noise [21], the linear regression based discriminant methods are often sensitive to outliers [16]. Compared regression based methods, MaxEnt-RDA is robust to outliers due to its MaxEnt objective.

4 Experiments

In this section, we applied the proposed MaxEnt-RDA algorithm to several real-world pattern recognition problems and compared its performance with PCA, LDA and MFA.

4.1 A Toy Problem

In the toy problem, a two-class problem is discussed. The data for each class is non-Gaussian distribution. As shown in Fig. 1, class one has 100 2D points from a bimodal Gaussian distribution (blue triangles), with centers at (-3.5, 3.5) and (3.5, 1.0); class two has also 100 2D points (red circles) which are drawn from a bimodal Gaussian distribution with centers at (3.5, -3.5) and (0.5, 1.0). The scale factor s in (25) of MaxEnt-RDA was set to 8.

The solid line in Fig. 1 represents the derived classification hyperline for MaxEnt-RDA; and the dashed line is the optimal classification hyperline for LDA. It is obvious

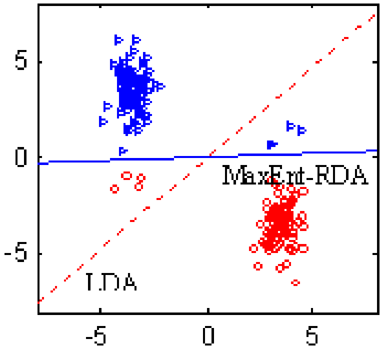


Fig. 1. A toy problem: comparative optimal classification hyperlines for a two-class bimodal Gaussian distribution

Table 1. UCI datasets used in the experiments

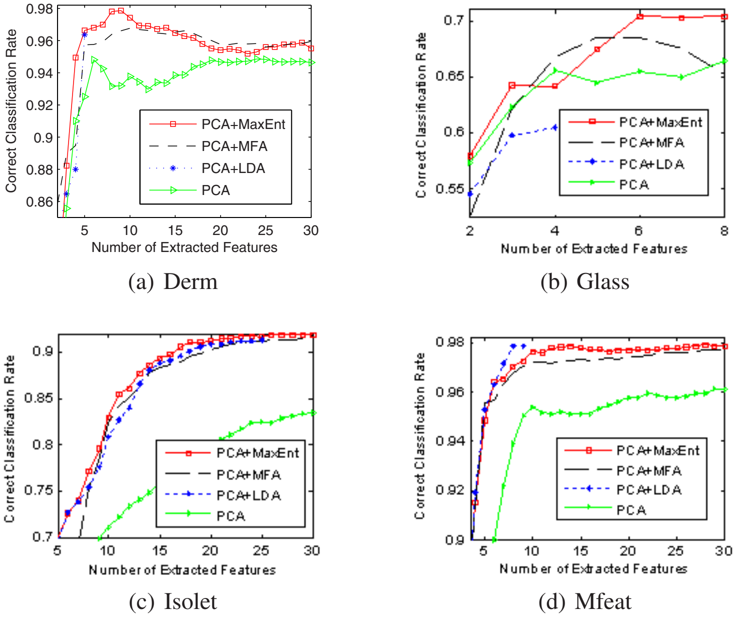
data set	Dimension	Classes	Instances
Derm	34	6	366
Glass	9	6	214
Isolet	617	26	1559
Mfeat	216	10	2000

that optimal hyperline of MaxEnt-RDA can perfectly separate the two-class data. But LDA fails to find an optimal direction in the non-Gaussian case. Given a suitable scale factor s in the bandwidth σ, MaxEnt-RDA will learn a correct hyperline. It provides an efficient means for discovering non-Gaussian structures in the data.

4.2 Classification on UCI Dataset

We applied MaxEnt-RDA to four UCI data sets in UCI machine learning repositories [27]. Tab. 1 shows a brief summary of the data sets used in this experiment, which have been used in many feature extraction studies [13] [14][15]. For each data set, we performed 10-fold cross validation (CV) 10 times and computed the average correct classification rate. Each dimension of raw data was normalized using the means and the standard variances. The Nearest-Neighbor [19] algorithm based on Euclidean distance, which is commonly applied as the classifier in feature extraction, is utilized as the classifier. The facial images in the second row of Fig. 3 illustrate examples of the noisy image.

Fig. 2 shows the average correct classification rates of each data set with various numbers of extracted features. For ”Glass” dataset, the number of extracted features m is varied from 2 to $d-1$. For data sets with a high input dimension such as the ”Derm”, ”Ionosphere” and ”Isolet” data sets, the number of extracted features in the Figure 2 was truncated at 30 for a clear view. The final reduced dimension after LDA is $c-1$.



(a) Derm

(b) Glass

(c) Isolet

(d) Mfeat

Fig. 2. Classification accuracy for UCI data sets

The results in Fig. 2 show that MaxEnt-RDA achieves the highest correct classification rate on four datasets. In Fig.2 (a) and (b), the highest correct classification rate of MaxEnt-RDA is clearly higher than that of other methods. In Fig. 2(c), MaxEnt-RDA, LDA and MFA are all perform significantly better than PCA. Although the results of MaxEnt-RDA, LDA and MFA are very close, MaxEnt-RDA still slightly outperforms LDA and MFA. In Fig. 2(d), MaxEnt-RDA outperforms MFA; and the highest correct classification rate of MaxEnt-RDA is slightly larger that of LDA. The results on UCI datasets further demonstrate that MaxEnt-RDA provides an efficient method for discriminant feature extraction.

4.3 Robustness to Noise

In real face recognition system, it is necessary to label large amount of training facial images to learn a robust classifier. But during collection of facial images, two types of noise may be occur. The one is mislabeling noise and the other is image noise. The mislabeling noise occurs when we mislabel one person's facial image to another; and the image noise occurs when the person's face is close to border of the camera so that we can only crop part of the face.

In this experiment, we use CMU PIE face database [28] to validate different methods under above two noise. The facial images are collected from a subset of CMU PIE face database that contains more than 40,000 facial images of 68 subjects. These still images are acquired across different poses, illuminations and facial expressions. We choose the

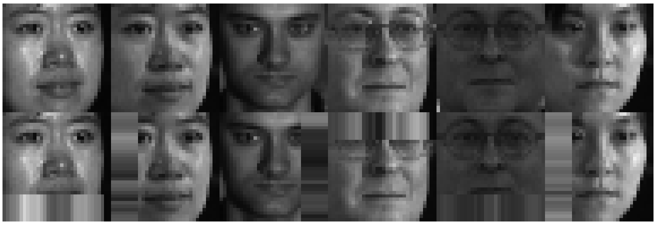


Fig. 3. Cropped facial images and their corresponding noisy images on CMU PIE database

five near frontal poses (C05, C07, C09, C27, C29) and use all the images under different illuminations and expressions, thus we get 170 images for each individual. All facial images are manually aligned and cropped. The size of each cropped image is 32×32. For each individual (20, 30 ,40) images are randomly selected for training and the rest are used for testing. And 5%, 10%,20% facial images per person is randomly selected as noisy samples. To eliminate statistical deviations, all experiments are averaged over 50 random splits and the mean as well as the standard deviation are reported. The Nearest-Center [19] algorithm based on Euclidean distance is used as the classifier.

Table 2. Comparison of different algorithms on mislabeling noise (average correct classification rate ± standard deviation). Best of all results are highlighted in bold.

	20 train			
	0%	5%	10%	20%
PCA + LDA	87.9 ± 0.71	87.1± 0.75	86.2 ± 0.82	84.4 ± 0.87
PCA + MFA	90.9 ± 0.58	88.8± 0.75	86.9 ± 0.90	83.4 ± 0.94
PCA + MaxEnt-RDA	**91.9** ± 0.56	**90.6**± 0.64	**89.3** ± 0.79	**86.6** ± 0.78
	30 train			
PCA + LDA	90.7 ± 0.37	90.2 ± 0.47	89.6 ± 0.50	88.4 ± 0.50
PCA + MFA	92.2 ± 0.53	91.0 ± 0.63	89.4 ± 0.57	86.8 ± 0.73
PCA + MaxEnt-RDA	**94.0** ± 0.31	**93.2** ± 0.40	**92.4** ± 0.40	**90.7** ± 0.43
	40 train			
PCA + LDA	92.1 ± 0.44	91.8 ± 0.44	91.5 ± 0.43	90.5 ± 0.55
PCA + MFA	93.0 ± 0.46	92.0 ± 0.53	90.7 ± 0.61	88.4 ± 0.71
PCA + MaxEnt-RDA	**94.9** ± 0.33	**94.5** ± 0.36	**94.0** ± 0.37	**92.8** ± 0.45

Tab. 2 shows the results of different methods under mislabeling noise. MaxEnt-RDA achieves the highest average correct classification rate on all testing datasets. Furthermore, its standard deviation is also smaller than two other methods. This means that MaxEnt-RDA can learn a stable result under mislabeling noise. Since the mislabeling noise changes the relationship of neighborhood near the boundary of different classes, MFA's performance decreases rapidly when more mislabeling samples are added into the training set. When 20% mislabeling samples are added, average correct classification rate of MFA is even lower than that of LDA. Since MaxEnt-RDA tries to estimate a MaxEnt distribution measured by Parzen window, it can efficiently deal with the mislabeling noise so that it performs better than other methods.

Tab. 3 shows the results of different methods under noisy images. Similar to mislabeling noise, when noise occurs, all methods' correct classification rates decrease. But it seems that all methods are more robust to image noise than mislabeling noise. MaxEnt-RDA still achieves the highest average correct classification rate, and its standard deviation is smaller than two other methods. Experimental results illustrate that MaxEnt-RDA is robust against training noise for both features (image) and labels.

Table 3. Comparison of different algorithms on image noise(average correct classification rate ± standard deviation). Best of all results are highlighted in bold.

	20 train			
	0%	5%	10%	20%
PCA + LDA	87.9 ± 0.71	87.5 ± 0.69	87.4 ± 0.65	87.0 ± 0.73
PCA + MFA	90.9 ± 0.58	89.5 ± 0.67	89.2 ± 0.68	88.7 ± 0.69
PCA + MaxEnt-RDA	**91.9** ± 0.56	**91.2** ± 0.55	**91.0** ± 0.57	**90.5** ± 0.58
	30 train			
PCA + LDA	90.7 ± 0.37	90.4 ± 0.39	90.3 ± 0.39	90.0 ± 0.40
PCA + MFA	92.2 ± 0.53	91.5 ± 0.52	91.2 ± 0.54	90.9 ± 0.51
PCA + MaxEnt-RDA	**94.0** ± 0.31	**93.5** ± 0.33	**93.3** ± 0.33	**93.0** ± 0.31
	40 train			
PCA + LDA	92.1 ± 0.44	91.9 ± 0.47	91.8 ± 0.44	91.6 ± 0.47
PCA + MFA	93.0 ± 0.46	92.4 ± 0.45	92.2 ± 0.50	91.9 ± 0.46
PCA + MaxEnt-RDA	**94.9** ± 0.33	**94.6** ± 0.39	**94.5** ± 0.37	**94.2** ± 0.38

5 Conclusion

A MaxEnt-RDA method is proposed for discriminant feature extraction based on the MaxEnt criterion. It utilizes Renyi's quadratic entropy as objective and hence has a clearly theoretical foundation. An eigen-decomposition algorithm is proposed to approximately solve the MaxEnt problem based on Taylor expansion. The proposed method bases on the Parzen window estimate of MaxEnt distribution and hence can effectively overcome the limitations of the traditional LDA algorithm in data distribution assumption. And it is robust against image noise and mislabeling noise. Experiments on UCI datasets and face databases have demonstrated the superiority of MaxEnt-RDA compared with traditional discriminant methods.

Acknowledgement

This work was supported in part by NSF of China (No. 60275025) and MOST of China (No. 2007DFC10740).

References

1. Guyon, I., Elissee, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
2. Jolliffe, I.T.: Principal component analysis. Springer, New York (1986)
3. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.: Face recognition using laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(3), 328–340 (2005)
4. Zhao, D., Lin, Z., Tang, X.: Laplacian pca and its applications. In: International Conference on Computer Vision (2007)
5. Zhu, M., Martinez, A.M.: Subclass discriminant analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(8), 1274–1286 (2006)
6. Hamsici, O.C., Martinez, A.M.: Bayes optimality in linear discriminant analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(4), 647–657 (2008)
7. Tao, D.C., Li, X.L., Wu, X.D., Maybank, S.J.: Geometric mean for subspace selection. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2), 260–274 (2009)
8. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(1), 711–720 (1997)
9. Fukunaga, K.: Statistical pattern recognition. Academic Press, London (1990)
10. Yan, S., Xu, D., Zhang, B., Zhang, H., Yang, Q., Lin, S.: Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(1), 40–51 (2007)
11. Zhao, D., Lin, Z., Xiao, R., Tang, X.: Linear laplacian discrimination for feature extraction. In: IEEE conference on Computer Vision and Pattern Recognition (2007)
12. Principe, J., Xu, D., Iii, J.W.F.: Information-theoretic learning, `http://www.cnel.ufl.edu/bib/./pdf_papers/chapter7.pdf`
13. Torkkola, K.: Feature extraction by nonparametric mutual information maximization. Journal of Machine Learning Research 3, 1415–1438 (2003)
14. Nenadic, Z.: Information discriminant analysis: feature extraction with an information-theoretic objective. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(8), 1394–1407 (2007)
15. Hild II, K.E., Erdogmus, D., Torkkola, K., Principe, C.: Feature extraction using information-theoretic learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(9), 1385–1392 (2006)
16. Yuan, X., Hu, B.: Robust feature extraction via information theoretic learning. In: International Conference on Machine Learning (ICML 2009), Montreal, Canada (2009)
17. Rao, S., Liu, W., Principe, J., de Medeiros Martins, A.: Information theoretic mean shift algorithm. In: Machine Learning for Signal Processing (2006)
18. Caticha, A., Giffin, A.: Updating probabilities. In: The 26th International Workshop on Bayesian Inference and Maximum Entropy Methods, Paris, France (2006)
19. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. Wiley-Interscience, Hoboken (2000)
20. Cover, T., Thomas, J.: Elements of Information Theory, 2nd edn. John Wiley, New Jersey (2005)
21. Liu, W., Pokharel, P.P., Principe, J.C.: Correntropy: Properties and applications in non-gaussian signal processing. IEEE Transactions on Signal Processing 55(11), 5286–5298 (2007)
22. Huber, P.: Robust statistics. Wiley, Chichester (1981)
23. Baek, J., Son, Y.S.: Local linear logistic discriminant analysis with partial least square components. In: Li, X., Zaïane, O.R., Li, Z.-h. (eds.) ADMA 2006. LNCS (LNAI), vol. 4093, pp. 574–581. Springer, Heidelberg (2006)

24. Ye, J.: Least squares linear discriminant analysis. In: International Conference on Machine Learning, ICML (2007)
25. Cai, D., He, X., Han, J.: Spectral regression for efficient regularized subspace learning. In: International Conference on Computer Vision, pp. 1–7 (2007)
26. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: Data mining, inference, and prediction. Springer, Heidelberg (2001)
27. Newman, D., Hettich, S., Blake, C., Merz, C.: Uci repository of machine learning databases (1998), `http://www.ics.uci.edu/mlearn/MLRepository.html`
28. Sim, T., Baker, S., Bsat, M.: The cmu pose, illumination, and expression database. IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 1615–1618 (2005)