Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 725204, 9 pages
http://dx.doi.org/10.1155/2014/725204

Research Article

Trace Ratio Criterion for Feature Extraction in Classification

Guoqi Li,[1] Changyun Wen,[2] Wei Wei,[3] Yi Xu,[2] Jie Ding,[2] Guangshe Zhao,[4] and Luping Shi[1]

[1] *Department of Precision Instrument, Tsinghua University, Beijing 100084, China*
[2] *School of EEE, Nanyang Technological University, Singapore*
[3] *School of Computing, Xi'an Technological University, Shaanxi, China*
[4] *School of Aerospace, Xi'an Jiaotong University, Shaanxi, China*

Correspondence should be addressed to Guoqi Li; liguoqi@mail.tsinghua.edu.cn

Received 12 November 2013; Revised 28 March 2014; Accepted 28 March 2014; Published 29 May 2014

Academic Editor: Yang Tang

A generalized linear discriminant analysis based on trace ratio criterion algorithm (GLDA-TRA) is derived to extract features for classification. With the proposed GLDA-TRA, a set of orthogonal features can be extracted in succession. Each newly extracted feature is the optimal feature that maximizes the trace ratio criterion function in the subspace orthogonal to the space spanned by the previous extracted features.

1. Introduction

Linear discriminant analysis (LDA) [1–3] has been proposed as a class separatory measure, which has been intensively used to reduce dimensionality of a classification problem as well as improve the generalization capability of a pattern classifier. Generally speaking, LDA method is to optimize the ratio criterion of the between-class distance and within-class distance constructed based on the available learning data. Such optimization can be realized by solving a generalized eigenvalue problem of the between-class and within-class scatter matrices [4].

In our opinion, there are three main problems for LDA methods. The first problem is its difficulty of dealing with high-dimensional data, where the number of observed samples is much lower than the samples' feature dimension [5]. Many methods have been studied and proposed to address this problem; see, for example, the linear discriminant feature selection (LDFS) [6], Sparse Discriminant Analysis [5, 7], and Sparse Tensor Discriminant Analysis [8].

The second problem is the well-known undersampled problem [9] in LDA method, in which scatter matrices may become singular due to insufficient samples. The solutions to this problem have also been well investigated in the following works such as the Regularized LDA [10, 11] using regularization techniques [12, 13] and the Penalized LDA [14], the Pseudo Fisher Linear Discriminant [15], the Generalized Singular Value Decomposition [16], and the Uncorrelated LDA [17] and the Orthogonal LDA [17].

Basically the above two problems are quite similar and they can be unified as the same problem, which has also been extensively investigated in the above schemes. However, the third problem due to the LDA method can only extract quite limited features for classification problems [4]. For example, in two-class classification, one can only find one nonzero eigenvalue (extracted feature), as the between-class scatter matrix is a rank-one matrix. To the best of our knowledge, currently there is no good way to deal with this problem yet.

We focus on the third problem in this paper. A generalized LDA based on trace ratio criterion [18–23] is proposed to overcome such a problem and an algorithm called generalized linear discriminant analysis based on trace ratio criterion algorithm (GLDA-TRA) is derived to extract features from the input feature space. The algorithm first extracts a feature which maximizes the trace ratio criterion by solving a generalized eigenvalue problem. It is shown that such a generalized eigenvalue problem is the same as the generalized eigenvalue problem of LDA. Then, the learning data are projected to a subspace orthogonal to the space spanned by the extracted features. In that orthogonal subspace, the algorithm continues to extract a feature which maximizes the proposed trace ratio criterion. This process continues

and, in this way, a set of orthogonal features is obtained iteratively. It is proven that each newly extracted feature is the optimal feature that maximizes the trace ratio criterion in the subspace orthogonal to the space spanned by the previous extracted features. Finally the extracted features are shown to give a sequence of trace ratios with magnitudes monotonically decreasing.

2. Problem Formulation

Let $(x, y) \in R^d \times \mathcal{Y}$ be a sample, where R^d denotes a d-dimensional feature space and $\mathcal{Y} = \{1, 2, \ldots, C\}$ is a label set. Let x_{ij} denote the ith sample in the jth class. The within-class scatter matrix S_W and the between-class scatter matrix S_B [24] are, respectively, defined as

$$S_W = \frac{1}{n}\sum_{j=1}^{C}\sum_{i=1}^{n_i}\left(x_{ij} - \mathbf{m}_j\right)\left(x_{ij} - \mathbf{m}_j\right)', \qquad S_B = \frac{1}{n}\sum_{j=1}^{C}\sum_{i=1}^{n_i}\left(\mathbf{m}_j - \mathbf{m}\right)\left(\mathbf{m}_j - \mathbf{m}\right)', \tag{1}$$

where n_i is the number of samples in the ith class and n is the total number of samples and $\mathbf{m}_j$ is the sample mean of the jth class and $\mathbf{m}$ is the sample mean of all classes and the notation $'$ means matrix transpose. Without loss of generality, we assume that $n \gg d \gg C$. Then S_W can be constructed as a full rank matrix of rank d while S_B is at most with rank $C - 1$. So, in this paper, S_W is considered as a symmetrical and positive definite matrix and S_B is a symmetrical and nonnegative definite matrix. The LDA method extracts features, that is, the column vectors in matrix W_{opt}, in such a way that the ratio of the between-class scatter and the within-class scatter is maximized [25]. Consider

$$W_{\text{opt}} = \arg\max_{W}\frac{\left|W'S_BW\right|}{\left|W'S_WW\right|} = \left[w_1 \;\; w_2 \;\; \cdots \;\; w_m\right], \tag{2}$$

where $|\cdot|$ is the determinant of a matrix and $\{w_i \mid i = 1, \ldots, m\}$ is the set of generalized eigenvectors of S_B and S_W corresponding to the m largest generalized eigenvalues $\{\lambda_i \mid i = 1, \ldots, m\}$ such that

$$S_Bw_i = \lambda_iS_Ww_i, \quad i = 1, \ldots, m. \tag{3}$$

Unfortunately, there are at most $C - 1$ nonzero generalized eigenvalues as the rank of S_B is at most $C - 1$. As such, for a two-class classification problem, LDA can only extract one feature.

To overcome this problem, we hope that one can continue to extract w_2 after w_1 is extracted. Generally speaking, we hope to extract w_{i+1} after the features $w_1, \ldots, w_i$ are extracted with i starting from 0. We now use the trace ratio criterion function in [24, 26] to formulate the above-mentioned feature extraction problem. The optimization model we proposed is shown as follows:

$$w_{i+1} = \arg\max_{w_{i+1}} F(W) = \frac{1}{2}\frac{\operatorname{tr}\left(W'S_BW\right)}{\operatorname{tr}\left(W'S_WW\right)} \quad \text{subject to} \quad W = W_{i+1} = \left[W_i \;\; w_{i+1}\right], \quad w_{i+1} \text{ is orthonormal to span } \{W_i\}, \tag{4}$$

where $W_i = [w_1 \;\; w_2 \;\; \cdots \;\; w_i]$ is a matrix denoting all the features extracted with W_0 being empty, w_{i+1} is the feature to be determined in the space orthogonal to span$\{W_i\}$, and

$$F(W) = \frac{1}{2}\frac{w_1S_Bw_1 + \cdots + w_iS_Bw_1 + w_{i+1}S_Bw_{i+1}}{w_1S_Ww_1 + \cdots + w_1S_Ww_1 + w_{i+1}S_Ww_{i+1}}, \tag{5}$$

with tr$(\cdot)$ denoting the trace of a matrix.

Remark 1. Later it can be shown that the first extracted vector w_1 which maximizes $F(W)$ in the space R^d is the same w_1 found by LDA in (2). Here it is also necessary to point out that the orthogonal constraint condition in W is needed in our formulated problem (4). As in (2), the numerator is the determinant of matrix $W'S_BW$. Implicitly there is a constraint that $w_1, \ldots, w_i$ cannot be the same. Because when $w_1 = \cdots = w_i$, the numerator $|W'S_BW|$ would be zero. But in (4), the numerator is the trace of $W'S_BW$, which does not exclude the possibility that $w_1, \ldots, w_i$ are the same. With the constraint in (4), such a possibility can be avoided.

3. Proposed Algorithm and Analysis

In this section, we present and analyze the proposed feature extraction algorithm. Our idea is summarized as follows. We first extract a feature w_1 by maximizing the trace ratio criterion function involving S_W and S_B in (4). When the current extracted features become $W_i = [w_1, \ldots, w_i]$, let span$\{W_i\}$ denote a space spanned by the linear combination of all the columns of W_i and let span$\{W_i\}^\perp$ denote the space orthogonal to span$\{W_i\}$. Then, S_W and S_B are projected onto the subspace span$\{W_i\}^\perp$ by using projection operators $(I - W_iW_i^+)$ and $W_iW_i^+$, respectively, where $W_i^+ = (W_i'W_i)^{-1}W_i'$ is the generalized matrix inverse of a column full rank matrix W_i. We continue the process to find w_{i+1} by optimizing (4) until all m $(m \le d)$ features are extracted.

We first present the algorithm in Section 3.1. In Section 3.2., we will show how this algorithm is derived and then analyze its properties.

3.1. Proposed Algorithm. For convenience, we present the definition of a generalized eigenvalue as follows.

Definition 2. A number λ is called a generalized eigenvalue of matrix B with respect to A if λ satisfies that $Bx = \lambda Ax$ for a nonzero vector x, where A is a positive definite symmetrical matrix. When $A = I$, λ is a normal eigenvalue of B.

Now the algorithm is given as follows.

Initialization Step: construct symmetrical matrices $S_{W_1} = S_W \in R^{d\times d}$ and $S_{B_1} = S_B \in R^{d\times d}$ as shown in (1) based on the available learning data with W_0 as an empty matrix.

Step i $(i = 1, \ldots, m)$ is as follows.

(a) Calculation stage: find a unit vector in the direction of a generalized eigenvector which corresponds to the maximum eigenvalue of the generalized eigenvalue problem of S_{B_i} with respect to S_{W_i}. This can be achieved by the following process:

do the Cholesky decomposition $S_{W_i} = G_i G_i'$. Let $S_i = G_i^{-1} B_{i-1} (G_i^{-1})'$ and obtain its maximum eigenvalue λ_i together with its corresponding eigenvector x_i. Choose $w_i = (G_i^{-1})' x_i / \|(G_i^{-1})' x_i\|_2$ and $W_i = [W_{i-1} \ w_i]$.

(b) Update stage is as follows:

$$\begin{aligned} S_{W_{i+1}} &= \left(I - W_i W_i^+\right) S_W \left(I - W_i W_i^+\right) \\ &\quad + \mu W_i W_i^+ S_W W_i W_i^+, \\ S_{B_{i+1}} &= \left(I - W_i W_i^+\right) S_B \left(I - W_i W_i^+\right), \end{aligned} \tag{6}$$

where u is a sufficiently small positive number.

Replace i by $i + 1$ until m features are extracted in W.

Remark 3. (1) As W is orthonormal, (6) can be rewritten as the following recursive from:

$$\begin{aligned} S_{W_{i+1}} &= \left(I - w_i w_i^+\right) S_{W_i} \left(I - w_i w_i^+\right) \\ &\quad + \mu \sum_{j=1}^{i} w_j w_j^+ S_W w_j w_j^+, \\ S_{B_{i+1}} &= \left(I - w_i w_i^+\right) S_{B_i} \left(I - w_i w_i^+\right). \end{aligned} \tag{7}$$

Later in Lemma 5, it will be shown that S_{W_i} for $i = 1, \ldots, m$ can still be positive definite for a sufficiently small positive u. At each step i, S_{W_i} is positive definite and the generalized eigenvalue of S_{B_i} with respect to S_{W_i} exists.

(2) The reason why we need to find the maximum eigenvalue of S_i in the Calculation stage is shown in Lemma 8.

(3) Suppose that i features $w_1, \ldots, w_i$, $(i < d)$ have been extracted. Then Theorem 12 shows that w_{i+1} is found to ensure the trace ratio criterion function in (4) to attain its maximum value.

(4) When $m = 1$, the LDA and GLDA-TRA extract the same feature.

(5) GLDA-TRA extracts $m \le d$ features one by one. When $i = m = d$, $S_{B_{i+1}}$ becomes a zero matrix, and the algorithm will not extract any more features.

3.2. Derivation and Analysis of GLDA-TRA

Lemma 4. *For an arbitrary full column rank matrix $W_i \in R^{d\times i}$, $W_i W_i^+ = W_i (W_i' W_i)^{-1} W_i'$ and $I - W_i W_i^+$ are projection operators which project a vector onto* span$\{W_i\}$ *and* span$\{W_i\}^\perp$, *respectively.*

Proof. Suppose that $v \in R^{m\times 1}$ belongs to span$\{W_i\}$; there exists a vector $x \in R^{i\times 1}$ such that $v = W_i x$. It can be obtained that $W_i W_i^+ v = W_i (W_i' W_i)^{-1} W_i' W_i x = W_i x = v$ and $(I - W_i W_i^+) v = (I - W_i (W_i' W_i)^{-1} W_i') v = (I - W_i (W_i' W_i)^{-1} W_i') W_i x = 0$. This lemma holds. □

Lemma 5. *$S_{W_{i+1}}$ in (6) for $i = 1, \ldots, m$ are positive definite for a sufficiently small positive μ.*

Proof. Note that $x S_{W_1} x > 0$ for any nonzero $x \in R^{m\times 1}$ as $S_{W_1} = S_W$. Let $x_1 = (I - W_i W_i^+) x$ and $x_2 = W_i W_i^+ x$. Then $x = (I - W_i W_i^+) x + W_i W_i^+ x = x_1 + x_2$. As $x \neq 0$, x_1 and x_2 cannot be zero at the same time. Thus

$$\begin{aligned} x' S_{W_{i+1}} x &= x' \left(I - W_i W_i^+\right) S_W \left(I - W_i W_i^+\right) x \\ &\quad + \mu x' \left(W_i W_i^+\right) S_W \left(W_i W_i^+\right) x \\ &= \left(\left(I - W_i W_i^+\right) x\right)' S_W \left(I - W_i W_i^+\right) x \\ &\quad + \mu \left(W_i W_i^+ x\right)' S_W W_i W_i^+ x \\ &= x_1' S_W x_1 + \mu x_2' S_W x_2 > 0, \end{aligned} \tag{8}$$

for a sufficiently small positive number μ. □

Lemma 6. *For matrices $W \in R^{d\times m}$, $X \in R^{d\times d}$, define $g(W) =$ tr$(W' X W)$. Then $dg/dW = (X' + X)W$.*

Proof. Let $W = [w_1 \ \cdots \ w_i \ \cdots \ w_m]$ and $X = [x_1 \ \cdots \ x_d]$, where $w_i \in R^{d\times 1}$ for $i = 1, \ldots, m$, and $x_i \in R^{d\times 1}$ for $i = 1, \ldots, d$. We have

$$\begin{aligned} g(W) &= \operatorname{tr}\left(W' X W\right) \\ &= \operatorname{tr}\left(\sum_{i=1}^{d} w_1' x_i w_{i1} + \cdots + \sum_{i=1}^{d} w_k' x_i w_{ik} \right. \\ &\qquad \left. + \cdots + \sum_{i=1}^{d} w_m' x_i w_{i1}\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} w_{j1} x_{ji} w_{i1} + \cdots \\ &\quad + \sum_{i=1}^{d}\sum_{j=1}^{d} w_{jk} x_{ji} w_{ik} + \sum_{i=1}^{d}\sum_{j=1}^{d} w_{jm} x_{ji} w_{im}. \end{aligned} \tag{9}$$

Then, for $k_0 = 1, \ldots, m$, $j_0 = 1, \ldots, d$, we get

$$\begin{aligned} \frac{dg(W)}{dw_{j_0 k_0}} &= \sum_{i=1}^{d} x_{j_0 i} w_{i k_0} + \sum_{j=1}^{d} w_{j k_0} x_{j j_0} \\ &= \sum_{i=1}^{d} x_{j_0 i} w_{i k_0} + \sum_{i=1}^{d} x_{i j_0} w_{i k_0}. \end{aligned} \tag{10}$$

So (10) gives $dg/dW = (X' + X)W$. □

Lemma 7. *Let $w \in R^{d\times 1}$ and the trace ratio function $f(w) = (1/2)(\mathrm{tr}(w'S_W w)/\mathrm{tr}(w'S_B w))$. The gradient $\nabla f(w) = df(w)/dw = (wS_W \cdot \mathrm{tr}(w'S_B w) - wS_B \cdot \mathrm{tr}(w'S_W w))/(\mathrm{tr}(w'S_B w))^2$.*

Proof. Let $g_1(w) = \mathrm{tr}(w'S_W w)$ and $g_2(w) = \mathrm{tr}(w'S_B w)$. From Lemma 6, $\nabla f(w) = df(w)/dw = ((dg_1(w)/dw)g_2(w) - (dg_2(w)/dw)g_1(w))/(g_2(w))^2 = (wS_W \cdot \mathrm{tr}(w'S_B w) - wS_B \cdot \mathrm{tr}(w'S_W w))/(\mathrm{tr}(w'S_B w))^2$. □

Define $R(w) = \mathrm{tr}(w'S_{B_i}w)/\mathrm{tr}(w'S_{W_i}w)$, and we have the following results mentioned in the Calculation stage of step i.

Lemma 8. *Assume that the maximum eigenvalue of S_i is λ^* with an eigenvector x^*. Then the unit vector $w^* = (G_i^{-1})'x^*/\|(G_i^{-1})'x^*\|$ is an extracted feature ensuring that $R(w)$ attains its maximum value which is equal to λ^*.*

Proof. Note that $f(w) = (1/2)(1/R(w)) = (1/2)(\mathrm{tr}(w'S_{W_i}w)/\mathrm{tr}(w'S_{B_i}w))$. Thus maximizing $R(w)$ is equivalent to minimizing $f(w)$. From Lemma 7, we have

$$\begin{aligned}\nabla f(w(k)) &= \left.\frac{df(w)}{dw}\right|_{w=w(k)} \\ &= \left.\frac{wS_{W_i}\cdot \mathrm{tr}\left(w'S_{B_i}w\right) - wS_B\cdot \mathrm{tr}\left(w'S_{W_i}w\right)}{(\mathrm{tr}(w'S_{B_i}w))^2}\right|_{w=w(k)}.\end{aligned} \tag{11}$$

Then one can minimize $f(w)$ by using an iterative method. Let $w(k+1) - w(k) = -\eta\nabla f(w(k))$, where k denotes the kth iteration and η is a small positive constant. Then $f(w(k+1)) = f(w(k)) - \eta(\nabla f(w(k)))'(\nabla f(w(k))) + o(\|w(k+1) - w(k)\|_2^2) \le f(w(k))$. Thus $\{f(w(k))\}$ is a nonincreasing positive sequence and its limit, denoted as $f^*(w)$, exists. We have $\lim_{k\to\infty} f(w(k+1)) = f(w(k)) = f^*(w)$. This implies that $\lim_{k\to\infty}\eta(\nabla f(w_k))'(\nabla f(w(k))) = 0$. As $\eta > 0$, then $\{w(k)\}$ converges to an accumulation point $\overline{w}$ that satisfies $\nabla f(w)|_{w=\overline{w}} = 0$. This gives

$$\begin{aligned}&\left.\frac{df(w)}{dw}\right|_{w=\overline{w}} \\ &= \frac{\overline{w}S_{W_i}\cdot \mathrm{tr}\left(\overline{w}'S_{B_i}\overline{w}\right) - \overline{w}S_{B_i}\cdot \mathrm{tr}\left(\overline{w}'S_{W_i}\overline{w}\right)}{\left[\mathrm{tr}\left(\overline{w}'S_{B_i}\overline{w}\right)\right]^2} = 0.\end{aligned} \tag{12}$$

Let $\lambda = \mathrm{tr}(\overline{w}'S_{B_i}\overline{w})/\mathrm{tr}(\overline{w}'S_{W_i}\overline{w})$. From (12), we have $S_{B_i}\overline{w} = \lambda S_{W_i}\overline{w}$. That is, λ is the generalized eigenvalue obtained from

$$S_{B_i}w = \lambda S_{W_i}w. \tag{13}$$

Note that each eigenvector w in (13) is an accumulation point of $f(w)$, since it satisfies (12). Now we hope to convert the generalized eigenvalue problem of (13) to a normal eigenvalue problem. This is achieved by Cholesky decomposition of S_{W_i}, which is given as $S_{W_i} = G_iG_i'$, where G_i is a full column rank lower triangular matrix. By doing this, (13) becomes

$$S_{B_i}w = \lambda G_iG_i'w. \tag{14}$$

Defining $x = G_i'w$ and substituting x into (14), it can be obtained that $S_ix = \lambda x$, where $S_i = G_i^{-1}S_{B_i}(G_i^{-1})'$ is a symmetrical positive definite matrix. Let the maximum eigenvalue and its corresponding eigenvector of S_i be λ^* and x^*. Finally the optimal $w^* = (G_i^{-1})'x^*/\|(G_i^{-1})'x^*\|$ is a unit vector and $\lambda^* = \max(R(w)) = \mathrm{tr}(w^{*\prime}S_Bw^*)/\mathrm{tr}(w^{*\prime}S_Ww^*)$. □

Theorem 9. *Suppose that w_i is a feature extracted at step i; the trace ratio $(1/2)(\mathrm{tr}(w_i'S_{B_{i+1}}w_i)/\mathrm{tr}(w_i'S_{W_{i+1}}w_i)) = 0$ at step $i+1$.*

Proof. Note that $\mathrm{tr}(w_i'S_{B_{i+1}}w_i) = w_i(I - W_iW_i^+)S_B(I - W_iW_i^+)w_i$. We have $\mathrm{tr}(w_i'S_{W_{i+1}}w_i) > 0$ since $S_{W_{i+1}}$ is positive definite by Lemma 5. Also, $(I - W_iW_i^+)w_i = 0$ as $(I - W_iW_i^+)w_i$ is a projection operator which projects a vector onto $\mathrm{span}\{W_i\}^\perp$ based on Lemma 4. Thus, $(1/2)(\mathrm{tr}(w_i'S_{B_{i+1}}w_i)/\mathrm{tr}(w_i'S_{W_{i+1}}w_i)) = 0$. □

Remark 10. As seen in Theorem 9, if the feature w_i has been extracted at step i, then w_i is supposed to make the trace ratio $(1/2)(\mathrm{tr}(wS_{B_i}w)/\mathrm{tr}(wS_{W_i}w))$ attain its maximum in this step. While at step $i+1$, after S_{W_i} and S_{B_i} are updated to $S_{W_{i+1}}$ and $S_{B_{i+1}}$, respectively, we have $(1/2)(\mathrm{tr}(wS_{B_{i+1}}w)/\mathrm{tr}(w'S_{W_{i+1}}w)) = 0$ at $w = w_i$. This means that the algorithm needs to find w_{i+1} which can maximize $(1/2)(\mathrm{tr}(wS_{B_{i+1}}w)/\mathrm{tr}(w'S_{W_{i+1}}w))$, and, obviously, w_{i+1} must be different from w_i.

Theorem 11. *Sequence $\{F(w_i)\}_{i=1}^m$ produced by GLDA-TRA is a decreasing sequence.*

Proof. Assume that V_i is a space which the ith feature w_i belongs to. That is, w_i is the feature extracted from V_i such that $F(w_i)$ attains its maximum. Now we consider $w_i \in V_i$ at step i and $w_{i+1} \in V_{i+1}$ at step $i+1$ with the respective corresponding maximum eigenvalues λ_i and λ_{i+1}. As $S_{W_{i+1}}$ is positive definite; from Lemma 4 it can be obtained that

$$\begin{aligned}\left\|S_{W_{i+1}}w\right\|_2 &= \left\|\left(\left(I - W_iW_i^+\right)S_W\left(I - W_iW_i^+\right)\right.\right. \\ &\qquad \left.\left. + \mu W_iW_i^+S_WW_iW_i^+\right)w\right\|_2 \\ &= \left\|(I - W_iW_i^+)S_W(I - W_iW_i^+)w\right\|_2 \\ &\quad + \mu\left\|W_iW_i^+S_WW_iW_i^+w\right\|_2 \\ &\longrightarrow \left\|\left(I - W_iW_i^+\right)S_W\left(I - W_iW_i^+\right)w\right\|_2,\end{aligned} \tag{15}$$

when $\mu \to 0$. If $w \in \mathrm{span}\{W_i\}$, $(I - W_iW_i^+)w = 0$. Then, $\|S_{W_{i+1}}w\|_2 = \|(I - W_iW_i^+)S_W(I - W_iW_i^+)w\|_2 = 0$. If $w \in \mathrm{span}\{W_i\}^\perp$, $(I - W_iW_i^+)w = w$ and then $\|S_{W_{i+1}}w\|_2 > 0$. Also, as λ_i and λ_{i+1} correspond to the maximum eigenvalues at steps i and $i+1$, it is obvious that

$$\begin{aligned}\left\|\lambda_iS_{W_i}w_i\right\|_2 &\ge \mu\|w_i\|_2 > 0, \\ \left\|\lambda_{i+1}S_{W_{i+1}}w_{i+1}\right\|_2 &\ge \mu\|w_{i+1}\|_2 > 0.\end{aligned} \tag{16}$$

Thus, it can be concluded that $w_i \in \mathrm{span}\{W_i\}$ while $w_{i+1} \in V_{i+1} \subseteq \mathrm{span}\{W_i\}^\perp$. Similarly, $w_i \in V_i \subseteq \mathrm{span}\{W_{i-1}\}^\perp$. As

$W_i = [w_1, \ldots, w_i]$, span$\{W_i\}$ increases as i increases. Thus, V_i decreases as i increases. In addition, we have $V_1 = R^d = V_2 \oplus \text{span}\{w_1\} = V_3 \oplus \text{span}\{w_1\} \oplus \text{span}\{w_2\} = \cdots = V_m \oplus \text{span}\{w_1\} \oplus \cdots \oplus \text{span}\{w_{m-1}\}$. Thus, $F(w_1) \geq F(w_2) \geq \cdots \geq F(w_m)$. □

Theorem 12. *The feature w_{i+1} extracted by GLDA-TRA is the optimal feature maximizing $F(W_{i+1})$ in (4) when $W_i = [w_1 \ \cdots \ w_i]$ is extracted.*

Proof. Note that $w_{i+1} \in \text{span}\{W_i\}^\perp$ as w_{i+1} is orthonormal to span$\{W_i\}$. Consider an arbitrary $\widetilde{w}_{i+1} \neq w_{i+1} \in \text{span}\{W_i\}^\perp$. Construct $W_{i+1} = [W_i \ w_{i+1}]$ and $\widetilde{W}_{i+1} = [W_i \ \widetilde{w}_{i+1}]$. Note that W_i is an orthonormal matrix. Then $w_{i_0}' w_{j_0} = 0$ as long as $i_0 \neq j_0$ and we have

$$\begin{aligned} &F(W_{i+1}) - F(\widetilde{W}_{i+1}) \\ &= \frac{w_{i+1}S_B w_{i+1} + \cdots + w_{i+1}S_B w_{i+1} + w_{i+1}S_B w_{i+1}}{w_{i+1}S_W w_{i+1} + \cdots + w_{i+1}S_W w_{i+1} + w_{i+1}S_W w_{i+1}} \\ &\quad - \frac{w_{i+1}S_B w_{i+1} + \cdots + w_{i+1}S_B w_{i+1} + \widetilde{w}_{i+1}S_B \widetilde{w}_{i+1}}{w_{i+1}S_W w_{i+1} + \cdots + w_{i+1}S_W w_{i+1} + \widetilde{w}_{i+1}S_W \widetilde{w}_{i+1}} \\ &= \left(w_{i+1}S_B w_{i+1}\widetilde{w}_{i+1}S_W \widetilde{w}_{i+1} \right. \\ &\quad - \widetilde{w}_{i+1}S_B \widetilde{w}_{i+1} w_{i+1} S_W w_{i+1}\big) \\ &\quad \times \left(\left(w_1 S_B w_1\right)^2 + \cdots + \left(w_i S_B w_1\right)^2 \right. \\ &\quad \left. \left. + \left(w_{i+1}S_B w_{i+1}\right)\left(\widetilde{w}_{i+1}S_W \widetilde{w}_{i+1}\right)\right)^{-1}. \end{aligned} \tag{17}$$

Denote that $S_B = (I - W_i W_i^+)S_B + W_i W_i^+ S_B$ and $S_W = (I - W_i W_i^+)S_W + W_i W_i^+ S_W$. Then, $W_i W_i^+ w_{i+1} = 0$ and $(I - W_i W_i^+)w_{i+1} = w_{i+1}$. Thus, we have

$$\begin{aligned} &w_{i+1}S_B w_{i+1} \\ &= w_{i+1}\left(\left(I - W_i W_i^+\right)S_B + W_i W_i^+ S_B\right) w_{i+1} \\ &= w_{i+1}\left(\left(I - W_i W_i^+\right)S_B \left(I - W_i W_i^+\right)\right) w_{i+1} \\ &= w_{i+1} S_{B_{i+1}} w_{i+1}, \end{aligned} \tag{18}$$

where $S_{B_{i+1}} = ((I - W_i W_i^+)S_B(I - W_i W_i^+))$. Similarly, it can be obtained that

$$\begin{aligned} w_{i+1}S_W w_{i+1} &= w_{i+1} S_{W_{i+1}} w_{i+1}, \\ \widetilde{w}_{i+1}S_B \widetilde{w}_{i+1} &= \widetilde{w}_{i+1} S_{B_{i+1}} \widetilde{w}_{i+1}, \\ \widetilde{w}_{i+1}S_W \widetilde{w}_{i+1} &= \widetilde{w}_{i+1} S_{W_{i+1}} \widetilde{w}_{i+1}, \end{aligned} \tag{19}$$

where $S_{W_{i+1}} = ((I - W_i W_i^+)S_W(I - W_i W_i^+))$. From Lemma 8, it is noted that $\text{tr}(wS_{B_{i+1}}w)/\text{tr}(wS_{W_{i+1}}w)$ attains its maximum if and only if w is parallel to the eigenvector corresponding to the maximum generalized eigenvalue of $S_{B_{i+1}}$ with respect to $S_{W_{i+1}}$. Since w_{i+1} is parallel to the generalized eigenvector, as long as $w_{i+1} \neq \widetilde{w}_{i+1}$, we have

$$\frac{\text{tr}\left(w_{i+1}S_{B_{i+1}} w_{i+1}\right)}{\text{tr}\left(w_{i+1}S_{W_{i+1}} w_{i+1}\right)} > \frac{\text{tr}\left(\widetilde{w}_{i+1}S_{B_{i+1}} \widetilde{w}_{i+1}\right)}{\text{tr}\left(\widetilde{w}_{i+1}S_{W_{i+1}} \widetilde{w}_{i+1}\right)}. \tag{20}$$

That is to say, $w_{i+1}S_B w_{i+1}\widetilde{w}_{i+1}S_W \widetilde{w}_{i+1} > \widetilde{w}_{i+1}S_B \widetilde{w}_{i+1} w_{i+1} S_W w_{i+1}$. So we have $F(W_{i+1}) > F(\widetilde{W}_{i+1})$. Thus, this theorem holds. □

Remark 13. Our main theoretical conclusions of this paper are shown in Theorems 9–12. The implication of Theorem 9 is summarized in Remark 10. In Theorem 11, the decreasing of the trace ratio sequence $\{F(w_i)\}$ means that the separability of the data set on the each newly extracted features decreases compared with the previous extracted features. This is the main result and contribution of this paper. Note that the trace ratio cost function is to evaluate whether a feature is good or not. By employing our proposed method, one can obtain m ($m \leq d$ and m can be greater than $C - 1$) new extracted orthogonal features in which the w_i is better than or at least equal to w_{i+1}. Theorem 12 gives us more concrete conclusions: if we already have i features denoted as $w_1, \ldots, w_i$, the $i + i$th feature w_{i+1} extracted by GLDA-TRA is the optimal feature in the space orthogonal to the space spanned by $w_1, \ldots, w_i$. We present and prove Lemmas 4–8 as they are needed in proving Theorems 9–12. The theorems will be illustrated and verified in Section 4.

4. Simulation and Discussion

To simply illustrate our idea, two experiments are done. The first one is on Iris data set and the second one is on an artificial data set.

Example 14. Iris data set is a standard data set to verify the performance of classification algorithms. There are 150 data points belonging to three classes ($C = 3$): Setosa, Versicolor, and Virginica, respectively. Each class has 50 samples with four features ($d = 4$): Sepal length, Sepal width, Petal length, and Petal width.

One can construct S_W and S_B based on the Iris data set. It can be checked that $\text{rank}(S_W) = 4$ and $\text{rank}(S_B) = C - 1 = 2$, so LDA can only extract at most two features, while by employing GLDA-TRA to the Iris data set, we can extract m ($m \leq 4$) features one by one, which are denoted as $W = [w_1 \ w_2 \ \cdots \ w_m]$. When $m = 4$, the obtained $W = [w_1 \ w_2 \ w_3 \ w_4]$ is given as

$$W = \begin{bmatrix} -0.0828 & 0.1717 & 0.6691 & -0.7183 \\ -0.4612 & -0.0012 & 0.6222 & 0.6325 \\ 0.4446 & -0.8434 & 0.3004 & 0.0270 \\ 0.7634 & 0.5090 & 0.2735 & 0.2886 \end{bmatrix}. \tag{21}$$

Obviously, $W'W = I$. Figure 1 shows the distribution of the three classes of the data points after projecting them onto each extracted feature w_1, w_2, w_3, and w_4. As seen in Figure 1, the separability of the data decreases in the directions of w_1, w_2, w_3, and w_4. One can also notice that the sequence $\{F(w_i)\}_{i=1}^4$ is strictly decreasing as shown in Figure 2.

When using the extracted features to do classification via support vector machine (SVM), both LDA and GLDA-TRA can obtain good results in this example as the data points are well separated even if they are projected to one-dimensional space. When we extract one feature,

Table 1: Comparison of LDA and GLDA-TRA in the Iris data set.

Name	Number of extracted features	Accuracy rate	Name	Accuracy rate
LDA	1	98%	GLDA-TRA	98%
LDA	2	98%	GLDA-TRA	98%
LDA	3	Not available	GLDA-TRA	98.6667%
LDA	4	Not available	GLDA-TRA	98%

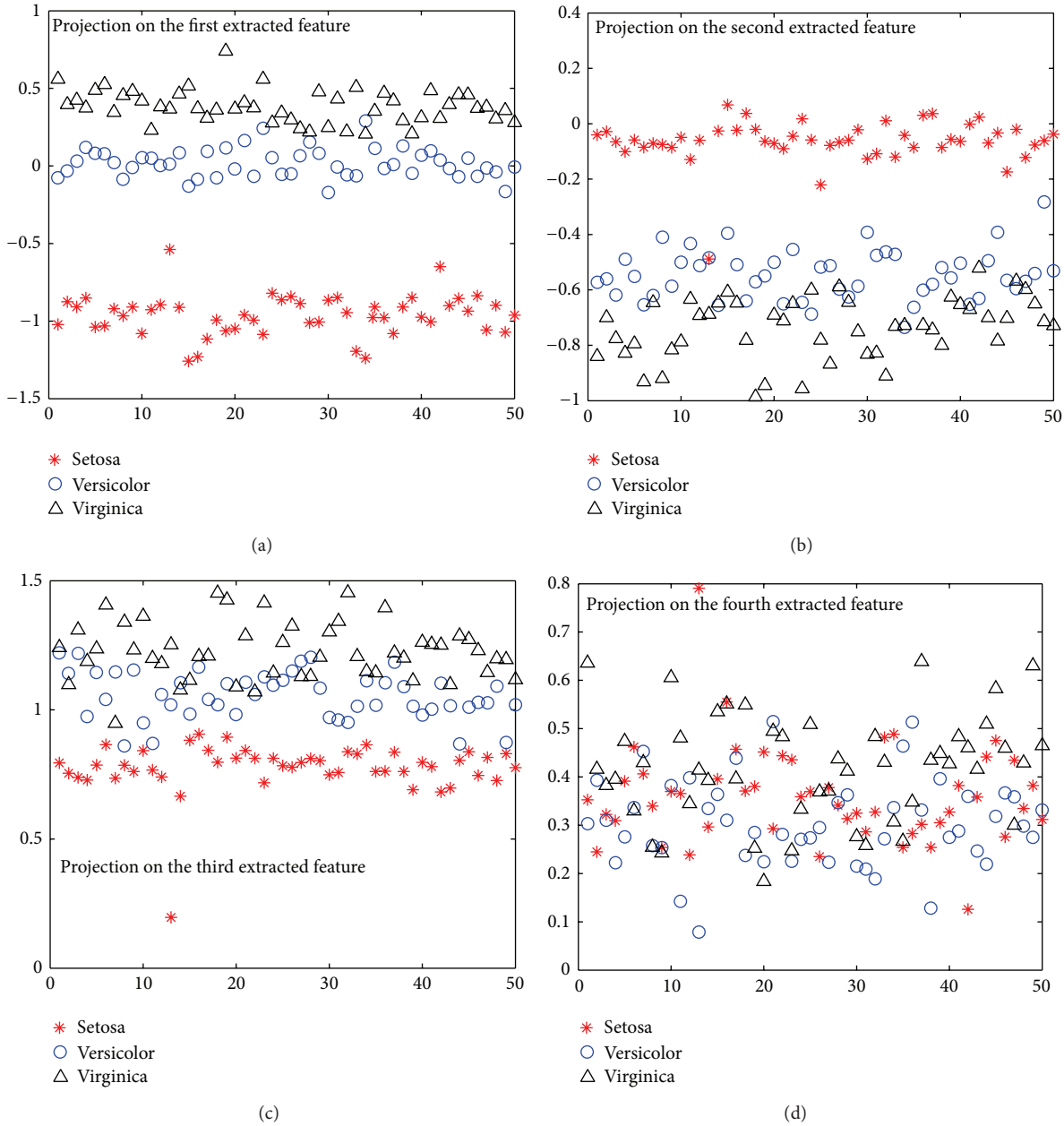

Figure 1: Illustration of projecting data points onto the extracted feature in order.

the accuracy is 98% (147/150) for both LDA and GLDA-TRA, since they extract the same feature when $m = 1$ (see Remarks 1 and 3). When we extract m features with our methods for $m = 2, 3$, and 4, the accuracies are, respectively, 98% (147/150) when $m = 2$, 98.6667% (148/150) when $m = 3$, and 98% (147/150) when $m = 4$. More details are shown in Table 1 and even when the data points are well separable, it is still observed that GLDA-TRA can be better than LDA in this case.

Example 15. This radar data was a 34-dimensional data set ($x \in R^{34}$, $y \in \mathcal{Y} = \{1, 2\}$, $C = 2$, $d = 34$)

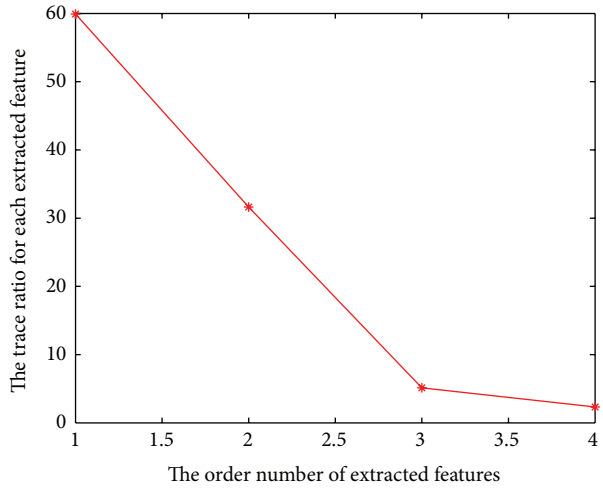

Figure 2: The trace ration for each extracted feature in order.

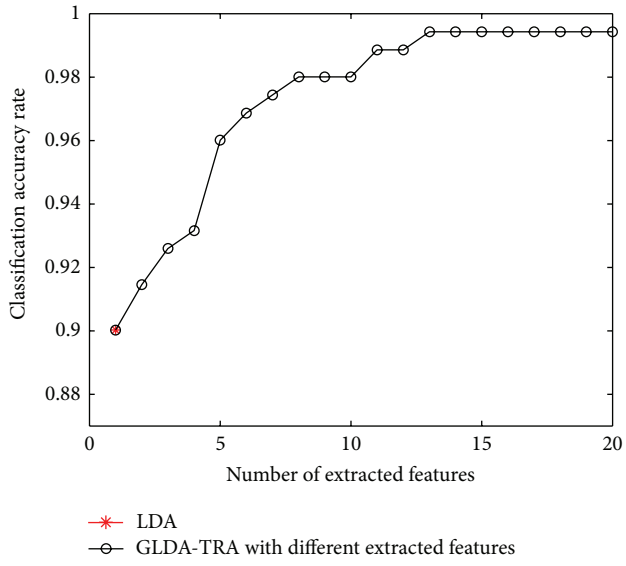

Figure 3: Classification accuracy rate for LDA and GLDA-TRA in Example 15.

collected by a system in Goose Bay, Labrador (http://archive.ics.uci.edu/ml/datasets/Ionosphere). This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere.

The comparisons of LDA and GLDA-TRA in this example are shown in Figure 3. It is seen that LDA and GLDA-TRA are the same when $m = 1$. But LDA cannot continue to extract more features while the proposed GLDA-TRA still extracts $m \leq d$ features one by one. When $m > 12$, the classification accuracy attains its maximum, which means that the dimension of the radar data set can be reduced to as low as 12 dimensions without losing any classification information.

In Theorem 11, it is shown that the trace ratio sequence $\{F(w_i)\}$ on each extracted feature in order is decreasing; this can also be verified in Figure 4, where both trace ratio and its logarithm are plotted. And from Theorem 12, we know that each newly extracted feature is the optimal feature that maximizes the trace ratio function in the subspace orthogonal to the space spanned by the previous extracted features, which is exactly what we claimed in the abstract.

Example 16. In the previous examples, both LDA and GLDA-TRA perform quite well even when only one feature is extracted. In this example, we will show that the data points are inseparable if they are projected to one-dimensional space. The classification problem is given by

$$(x, y) \in R^7 \times \mathcal{Y} = \begin{cases} \|x\|_2 = |1 + v_i| & \text{if } y = 1 \\ \|x\|_2 = |4 + v_i| & \text{if } y = 2, \end{cases} \tag{22}$$

where $y \in \mathcal{Y} = \{1, 2\}$ and v_i is a variable following normal distribution $N(0, 1)$. It can be known that most data points with label 1 locate around the surface of a sphere with radius $y = 1$ while data points with label $y = 2$ mostly locate around the surface of a sphere with radius 2.

To do the experiment, two hundred data points that are equally distributed in above two classes are generated. It is observed that LDA does not perform well if it is used to extract features in this problem. This is because LDA can only extract one feature and the data points are inseparable or well separable in an arbitrary one-dimensional feature space. By using GLDA-TRA, when we extract two features ($m = 2$), we obtain a projection matrix $W \in R^{34 \times 2}$. By employing $\tilde{x} = W^T x$, one can then visualize the data points by projecting the date onto the two extracted features $\tilde{x}$. Figure 5 shows the classification using SVM on this two-dimensional extracted feature space. With GLDA-TRA, the data points can be better separated if we extract more ($m > 1$) features. The comparisons of LDA and GLDA-TRA regarding the classification accuracy rate are shown in Table 2 and Figure 6.

5. Conclusion

In this paper, a generalized linear discriminant analysis based on trace ratio criterion (GLDA-TRA) algorithm has been proposed. This is to overcome the problem that linear discriminant analysis (LDA) can only extract limited features in classification. It is shown that, in GLDA-TRA, a set of orthogonal features can be extracted one by one. Each newly extracted feature is the optimal feature that maximizes the trace ratio criterion function in the subspace orthogonal to the space spanned by the previous extracted features. Finally the extracted features are such that the trace ratio sequence of these features is decreasing in order. Experimental results also show the effectiveness of our proposed algorithm.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

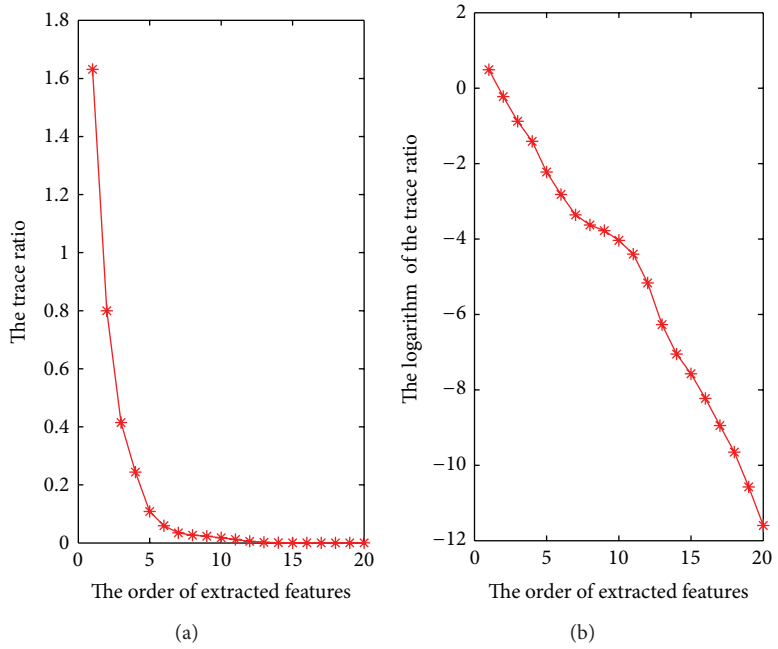

Figure 4: The trace ratio (a) and logarithm of the trace ratio (b) on each extracted feature in order.

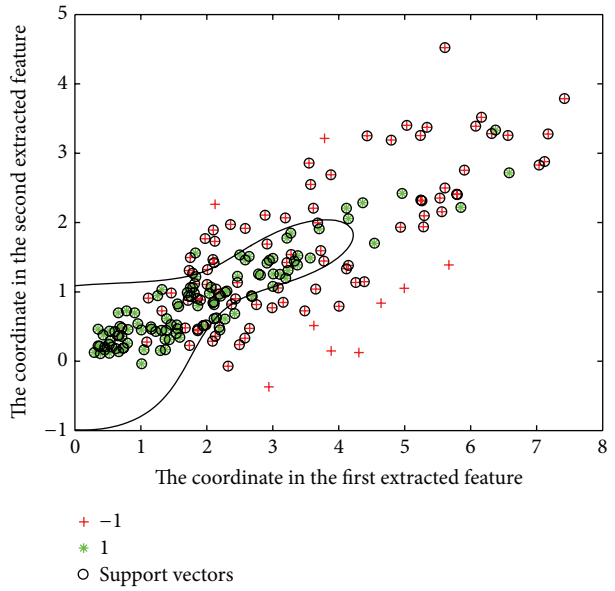

Figure 5: The data points projected onto the first two extracted features in Example 16.

Table 2: Comparison of LDA and GLDA-TRA in Example 16.

Name	Number of extracted features	Accuracy rate	Name	Accuracy rate
LDA	1	69%	GLDA-TRA	69%
LDA	2	Not available	GLDA-TRA	77.5%
LDA	3	Not available	GLDA-TRA	88%
LDA	4	Not available	GLDA-TRA	92.5%
LDA	5	Not available	GLDA-TRA	96%
LDA	6	Not available	GLDA-TRA	97%

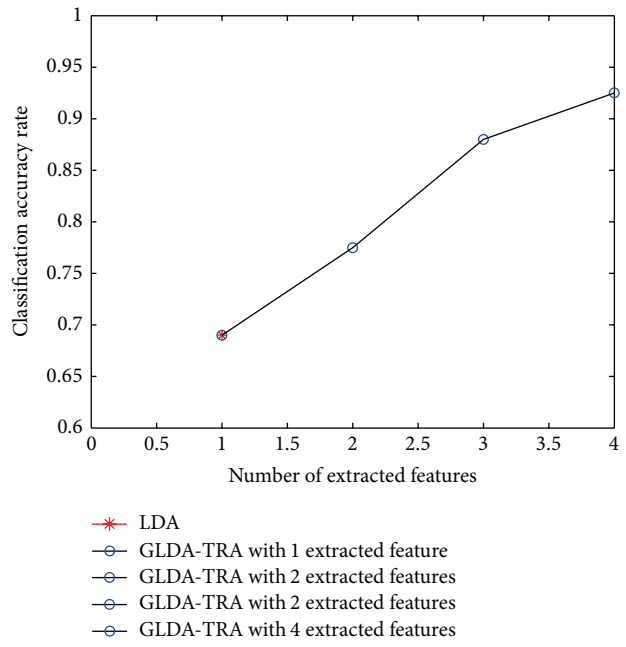

Figure 6: Classification accuracy rate for LDA and GLDA-TRA in Example 16.

References

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition, Second Edition*, Academic Press, Boston, Mass, USA, 1990.

[2] X. Li, W. Hu, H. Wang, and Z. Zhang, "Linear discriminant analysis using rotational invariant L1 norm," *Neurocompting*, vol. 73, no. 13–15, pp. 2571–2579, 2010.

[3] S. Noushath, G. Hemantha Kumar, and P. Shivakumara, "Diagonal Fisher linear discriminant analysis for efficient face recognition," *Neurocomputing*, vol. 69, no. 13–15, pp. 1711–1716, 2006.

[4] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[5] X. Shi, Y. Yang, Z. Guo, and Z. Lai, "Face recognition by sparse discriminant analysis via joint L21-norm minimization," *Pattern Recognition*, 2014.

[6] M. Masaeli, G. Fung, and J. G. Dy, "From transformation-based dimensionality reduction to feature selection," in *Proceedings of the International Conference on Machine Learning (ICML '10)*, pp. 751–758, June 2010.

[7] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[8] Z. Lai, Y. Xu, J. Yang, J. Tang, and D. Zhang, "Sparse tensor discriminant analysis," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3904–3915, 2013.

[9] L. Chen, H. M. Liao, M. Ko, J. Lin, and G. Yu, "New LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.

[10] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.

[11] D. Q. Dai and P. C. Yuen, "Regularized discriminant analysis and its application to face recognition," *Pattern Recognition*, vol. 36, no. 3, pp. 845–847, 2003.

[12] G. Li, C. Wen, G. Huang, and Y. Chen, "Error tolerance based support vector machine for regression," *Neurocomputing*, vol. 74, no. 5, pp. 771–782, 2011.

[13] G. Li, C. Wen, Z. G. Li, A. Zhang, Y. Feng, and K. Z. Mao, "Model based online learning with Kernels," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 3, pp. 356–369, 2013.

[14] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *The Annals of Statistics*, vol. 23, no. 1, pp. 73–102, 1995.

[15] M. Skurichina and R. P. W. Duin, "Regularisation of linear classifiers by adding redundant features," *Pattern Analysis and Applications*, vol. 2, no. 1, pp. 44–52, 1999.

[16] J. Ye, R. Janardan, C. H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 982–994, 2004.

[17] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *Journal of Machine Learning Research (JMLR)*, vol. 6, pp. 483–502, 2005.

[18] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, June 2007.

[19] Y. Liu, F. Nie, J. Wu, and L. Chen, "Efficient semi-supervised feature selection with noise insensitive trace ratio criterion," *Neurocomputing*, vol. 105, pp. 12–18, 2013.

[20] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, 2009.

[21] S. Cui and Y. C. Soh, "Linearity indices and linearity improvement of 2-D tetralateral position-sensitive detector," *IEEE Transactions on Electron Devices*, vol. 57, no. 9, pp. 2310–2316, 2010.

[22] T. T. Ngo, M. Bellalij, and Y. Saad, "The trace ratio optimization problem for dimensionality reduction," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 5, pp. 2950–2971, 2010.

[23] Z. Zhang, T. Chow, and M. Zhao, "Trace ratio optimization-based semi-supervised nonlinear dimensionality reduction for marginal manifold visualization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 5, pp. 1148–1161, 2013.

[24] L. Zhou, L. Wang, and C. Shen, "Feature selection with redundancy-constrained class separability," *IEEE Transactions on Neural Networks*, vol. 21, no. 5, pp. 853–858, 2010.

[25] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[26] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection," in *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp. 671–676, July 2008.