Generalized Linear Discriminant Analysis based on Trace Ratio Criterion for Feature Extraction in Classification

Guoqi Li[†], Changyun Wen, Fellow, IEEE, Guangshe Zhao, Luping Shi,

*G. Li and L. Shi are with the Department of Advanced Concepts and Nanotechnology (ACN), Data Storage Institute, A*STAR, 5, Engineer Drive, Singapore, 117608 (e-mail: LI_Guoqi@dsi.a-star.edu.sg (G. Li), SHI_Luping@dsi.a-star.edu.sg (L. Shi)).*
C. Wen is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: ecywen@ntu.edu.sg (C. Wen)).
G. Zhao is with School of Aerospace, Xi'an Jiaotong University, Shaanxi, Xi'an, 710049, P. R. China (e-mail: zhaogs@mail.xjtu.edu.cn).

Abstract

It is noted that linear discriminant analysis (LDA) can only extract limited features in classification problems. For example, in two-class classification, only one feature can be extracted by LDA. In this paper, a generalized linear discriminant analysis based on trace ratio criterion algorithm (GLDA-TRA) is provided to overcome the problem and an algorithm is derived. With 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 sub-space orthogonal to the space spanned by the previous extracted features.

Keywords: Feature extraction; Trace ratio; Class separability measure, Dimensionality reduction;

1. Introduction

Linear discriminant analysis (LDA) [1] [2] [3] has been proposed as a class separatory measure, which has been intensively used to reduce dimensionality of a classification problem as well as to improve the generalization capability of a pattern classifier. Generally speaking, LAD 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]. However, LDA method can only extract very 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.

In this paper, a generalized LDA based on trace ratio criterion is provided to overcome such a problem and an 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. Actually, it is shown that such a generalized eigenvalue problem is the same as the generalized eigenvalue problem of LDA. Then, the learning data is 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 precess 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 sub-space orthogonal to the space spanned by the previous extracted features. Finally the extracted features are shown to give a sequence of trace ratio criteria corresponding to these features is monotonically decreasing in magnitude.

2. Problem Formulation

Let $(x, y) \in R^d \times \mathcal{Y}$ be a sample, where R^d denotes an $d-$dimensional feature space and $\mathcal{Y} = \{1, 2, ..., C\}$ is a label set. Let x_{ij} denote the $i-$th sample in the $j-$th class. The within-class scatter matrix S_W and the between-class scatter matrix S_B [5] are respectively defined as

$$\begin{array}{ll} S_W & = \frac{1}{n}\sum_{j=1}^{C}\sum_{i=1}^{n_i}(x_{ij}-\mathbf{m}_j)(x_{ij}-\mathbf{m}_j)' \\ S_B & = \frac{1}{n}\sum_{j=1}^{C}\sum_{i=1}^{n_i}(\mathbf{m}_j-\mathbf{m})(\mathbf{m}_j-\mathbf{m})' \end{array} \tag{1}$$

where n_i is the number of samples in the $i-$th class and n is the total number of samples, and $\mathbf{m}_j$ is the sample mean of the $j-$th 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 non-negative definite matrix. The LAD method extracts features, i.e., 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 [7].

$$W_{opt} = arg\ max_W\ \frac{|W' S_B W|}{|W' S_W W|} = [w_1\ w_2\ ...w_m] \tag{2}$$

where $|.|$ is the determinant of a matrix, and $\{w_i | i = 1, ..., m\}$ is the set of generalized eigenvectors of S_B and S_W corresponding to the m largest generalized eigenvalues $\{\lambda_i | i = 1, ..., m\}$ such that $S_B w_i = \lambda_i S_W w_i,\ i = 1, ..., m$. 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 $w_1, ..., w_i$ are extracted with i starting from 0. We now use the trace ratio criterion function in [6] and [5] to formulate the above feature extraction problem as the following optimization problem.

$$\begin{array}{l} w_{i+1} = argmax_{w_{i+1}} F(W) = \frac{1}{2} \frac{tr(W' S_B W)}{tr(W' S_W W)} \\ subject\ to \quad W = W_{i+1} = [W_i\ w_{i+1}] \\ \qquad\qquad\quad w_{i+1}\ is\ orthonormal\ to\ span\{W_i\} \end{array} \tag{3}$$

where $W_i = [w_1\ w_2\ ...\ w_i]$ 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

$$F(W) = \frac{1}{2} \frac{w_1 S_B w_1 + ... + w_i S_B w_1 + w_{i+1} S_B w_{i+1}}{w_1 S_W w_1 + ... + w_1 S_W w_1 + w_{i+1} S_W w_{i+1}} \tag{4}$$

with $tr(.)$ 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 (3). As in (2), the numerator is the determinant of matrix $W' S_B W$. Implicity there is a constraint that $w_1, ..., w_i$ cannot be the same. Because when $w_1 = ... = w_i$, the numerator $|W' S_B W|$ would be zero. But in (3), the numerator is the trace of $W' S_B W$, which does not exclude the possibility that $w_1, ..., w_i$ are the same. With the constraint in (3), such a possibility can be avoided.*

3. Proposed Algorithm and Analysis

In this section, we present and then 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 (3). When the current extracted features become $W_i = [w_1, ..., w_i]$, let $span\{W_i\}$ denote a space spanned by the linear combination of all the columns of W_i and $span\{W_i\}^\perp$ 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 (3) until all m ($m \leq d$) features are extracted.

We first present the algorithm in subsection A. In subsection B, we will show how this algorithm is derived and analyze its properties.

3.1. Proposed Algorithm

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

Definition 1. *A number λ is called a generalized eigenvalue of matrix B with respect to A if λ satisfies that*

$$Bx = \lambda Ax \tag{5}$$

for a nonzero vector x, where A is a positive definite symmetrical matrix. When $A = I$, λ is a normal eigenvalue of B, which is a special case of (5).

Now the algorithm is given as below.

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 the available learning data with W_0 as an empty matrix.

Step i ($i = 1, ..., m$):

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 $w_i = \frac{(G_i^{-1})' x_i}{\|(G_i^{-1})' x_i\|_2}$ and $W_i = [W_{i-1} \; w_i]$.

b. Update stage:

$$\begin{array}{ll} S_{W_{i+1}} & = (I - W_i W_i^+) S_W (I - W_i W_i^+) + \mu W_i W_i^+ S_W W_i W_i^+ \\ S_{B_{i+1}} & = (I - W_i W_i^+) S_B (I - W_i W_i^+) \end{array} \tag{6}$$

where u is a sufficiently small positive number.

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

Remark 2.

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

$$\begin{array}{ll} S_{W_{i+1}} & = (I - w_i w_i^+) S_{W_i} (I - w_i w_i^+) + \mu \sum_{j=1}^{i} w_j w_j^+ S_W w_j w_j^+ \\ S_{B_{i+1}} & = (I - w_i w_i^+) S_{B_i} (I - w_i w_i^+) \end{array} \tag{7}$$

Later in Lemma 3.2, it will be shown that, S_{W_i} for $i = 1, ..., m$ 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 Theorem 3.5.*

3. *Suppose that i features $(i < d)$ have been extracted. Then theorem 3.8 shows that $w_1, ..., w_i, w_{i+1}$ are found to make the trace ratio criterion function in (3) attain its maximum value.*

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

3.2. Derivation and Analysis of GLDA-TRA

All the lemmas established in this subsection show how the algorithm is derived and are also useful to obtain the theorems which give the properties of the algorithm.

Lemma 3.1. *For an arbitrary full column rank matrix $W_i \in R^{d\times i}$, $W_iW_i^+ = W_i(W_i'W_i)^{-1}W_i'$ and $I - W_iW_i^+$ 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*$\{W_i\}$. Then there exists a vector $x \in R^{i\times 1}$ such that $v = W_ix$. It can be obtained that $W_iW_i^+v = W_i(W_i'W_i)^{-1}W_i')W_ix = W_ix = v$ and $(I - W_iW_i^+)v = (I - W_i(W_i'W_i)^{-1}W_i')v = (I - W_i(W_i'W_i)^{-1}W_i')W_ix = 0$. This lemma holds. □

Lemma 3.2. *$S_{W_{i+1}}$ in (6) for $i = 1, ..., m$ are positive definite for a sufficiently small positive μ.*

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

$$\begin{aligned} x'S_{W_{i+1}}x \quad &= x'(I - W_iW_i^+)S_W(I - W_iW_i^+)x \\ &\quad + \mu x'(W_iW_i^+)S_W(W_iW_i^+)x \\ &= ((I - W_iW_i^+)x)'S_W(I - W_iW_i{}^+)x \\ &\quad + \mu(W_iW_i{}^+x)'S_WW_iW_i{}^+x \\ &= x_1'S_Wx_1 + \mu x_2'S_Wx_2 > 0 \end{aligned} \tag{8}$$

for a sufficiently small positive number μ. □

Lemma 3.3. *For matrices $W \in R^{d\times m}$, $X \in R^{d\times d}$, define $g(W) = tr(W'XW)$. Then $\frac{dg}{dW} = (X' + X)W$.*

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

$$\begin{aligned} g(W) \quad &= tr(W'XW) \\ &= tr(\textstyle\sum_{i=1}^d w_1'x_iw_{i1} + ... + \sum_{i=1}^d w_k'x_iw_{ik} + ... + \sum_{i=1}^d w_m'x_iw_{i1}) \\ &= \textstyle\sum_{i=1}^d \sum_{j=1}^d w_{j1}x_{ji}w_{i1} + ...+ \\ &\quad \textstyle\sum_{i=1}^d \sum_{j=1}^d w_{jk}x_{ji}w_{ik} + \sum_{i=1}^d \sum_{j=1}^d w_{jm}x_{ji}w_{im} \end{aligned} \tag{9}$$

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

$$\begin{aligned} \tfrac{dg(W)}{dw_{j_0k_0}} \quad &= \textstyle\sum_{i=1}^d x_{j_0i}w_{ik_0} + \sum_{j=1}^d w_{jk_0}x_{jj_0} \\ &= \textstyle\sum_{i=1}^d x_{j_0i}w_{ik_0} + \sum_{i=1}^d x_{ij_0}w_{ik_0} \end{aligned} \tag{10}$$

So (9) gives $\frac{dg}{dW} = (X' + X)W$. □

Lemma 3.4. *Let $w \in R^{d\times 1}$ and the trace ratio function $f(w) = \frac{1}{2}\frac{tr(w'S_W w)}{tr(w'S_B w)}$. The gradient $\triangledown f(w) = \frac{df(w)}{dw} = \frac{wS_W \cdot tr(w'S_B w) - wS_B \cdot tr(w'S_W w)}{(tr(w'S_B w))^2}$.*

Proof. Let $g_1(w) = tr(w'S_W w)$ and $g_2(w) = tr(w'S_B w)$. From Lemma 3.3, $\triangledown f(w) = \frac{df(w)}{dw} = \frac{\frac{dg_1(w)}{dw}g_2(w) - \frac{dg_2(w)}{dw}g_1(w)}{(g_2(w))^2} = \frac{wS_W \cdot tr(w'S_B w) - wS_B \cdot tr(w'S_W w)}{(tr(w'S_B w))^2}$ □

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

Lemma 3.5. *Assume that the maximum eigenvalue of S_i is λ^* with an eigenvector x^*. Then the unit vector $w^* = \frac{(G_i^{-1})'x^*}{\|(G_i^{-1})'x^*\|}$ is an extracted feature ensuring that $R(w)$ attains its maximum value which is equal to λ^*.*

Proof. Note that $f(w) = \frac{1}{2}\frac{1}{R(w)} = \frac{1}{2}\frac{tr(w'S_{W_i}w)}{tr(w'S_{B_i}w)}$. Thus maximizing $R(w)$ is equivalent to minimizing $f(w)$. From Lemma 3.4, we have

$$\triangledown f(w(k)) = \frac{df(w)}{dw}|_{w=w(k)} = \frac{wS_{W_i} \cdot tr(w'S_{B_i}w) - wS_B \cdot tr(w'S_{W_i}w)}{(tr(w'S_{B_i}w))^2}|_{w=w(k)} \tag{11}$$

Then one can minimize $f(w)$ by using an iterative method. Let $w(k+1) - w(k) = -\eta\triangledown f(w(k))$ where k denotes the $k-$th iteration and η is a small positive constant. Then $f(w(k+1)) = f(w(k)) - \eta(\triangledown f(w(k)))'(\triangledown f(w(k))) + o(\|w(k+1) - w(k)\|_2^2) \leq f(w(k))$. Thus $\{f(w(k))\}$ is a non-increasing 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 $\lim_{k\to\infty} \eta(\triangledown f(w_k))'(\triangledown f(w(k))) = 0$. As $\eta > 0$, then $\{w(k)\}$ converges to an accumulation point $\bar{w}$ that satisfies $\triangledown f(w)|_{w=\bar{w}} = 0$. This gives

$$\frac{df(w)}{dw}|_{w=\bar{w}} = \frac{\bar{w}S_{W_i} \cdot tr(\bar{w}'S_{B_i}\bar{w}) - \bar{w}S_{B_i} \cdot tr(\bar{w}'S_{W_i}\bar{w})}{[tr(\bar{w}'S_{B_i}\bar{w})]^2} = 0 \tag{12}$$

Let $\lambda = \frac{tr(\bar{w}'S_{B_i}\bar{w})}{tr(\bar{w}'S_{W_i}\bar{w})}$. From (12), we have $S_{B_i}\bar{w} = \lambda S_{W_i}\bar{w}$. 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 $S_{W_i} = G_i G_i'$ where G_i is a full column rank lower triangular matrix. By doing this, (13) becomes

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

Defining $x = G_i' w$ and substituting x into (14), it can be obtained that $S_i x = \lambda x$ 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 $w^* = \frac{(G_i^{-1})' x^*}{\|(G_i^{-1})' x^*\|}$ is a unit vector and $\lambda^* = max(R(w)) = \frac{tr(w^{*\prime} S_B w^*)}{tr(w^{*\prime} S_W w^*)}$. □

Lemma 3.6. *Suppose that w_i is a feature extracted at step i. Then the trace ratio $\frac{1}{2} \frac{tr(w_i' S_{B_{i+1}} w_i)}{tr(w_i' S_{W_{i+1}} w_i)} = 0$ in the direction of w_i at step $i+1$.*

Proof. Note that $tr(w_i' S_{B_{i+1}} w_i) = w_i (I - W_i W_i^+) S_B (I - W_i W_i^+) w_i$. We have $tr(w_i' S_{W_{i+1}} w_i) > 0$ since $S_{W_{i+1}}$ is positive definite by Lemma 3.2. Also, $(I - W_i W_i^+) w_i = 0$ as $(I - W_i W_i^+) w_i$ is a projection operator which projects a vector onto $span\{W_i\}^{\perp}$ based on Lemma 3.1. Thus, $\frac{1}{2} \frac{tr(w_i' S_{B_{i+1}} w_i)}{tr(w_i' S_{W_{i+1}} w_i)} = 0$. □

Remark 3. *As seen in Lemma 3.6, if the feature w_i has been extracted at step i, then w_i is supposed to make the trace ratio $\frac{1}{2} \frac{tr(w S_{B_i} w)}{tr(w S_{W_i} w)}$ 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 $\frac{1}{2} \frac{tr(w S_{B_{i+1}} w)}{tr(w' S_{W_{i+1}} w)} = 0$ at $w = w_i$. This means that the algorithm needs to find w_{i+1} which can maximize $\frac{1}{2} \frac{tr(w S_{B_{i+1}} w)}{tr(w' S_{W_{i+1}} w)}$, and obviously, w_{i+1} must be different from w_i.*

Theorem 3.7. *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 $i-$th 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, it can be obtained that

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

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

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

Thus, it can be concluded that $w_i \in span\{W_i\}$ while $w_{i+1} \in V_{i+1} \subseteq span\{W_i\}^\perp$. Similarly, $w_i \in V_i \subseteq span\{W_{i-1}\}^\perp$. As $W_i = [w_1, ..., w_i]$, $span\{W_i\}$ increases as i increases. Thus, V_i decreases as i increases. In addition, we have $V_1 = R^d = V_2 \oplus span\{w_1\} = V_3 \oplus span\{w_1\} \oplus span\{w_2\} = ... = V_m \oplus span\{w_1\} \oplus ... \oplus span\{w_{m-1}\}$. Thus, $F(w_1) \geq F(w_2) \geq ... \geq F(w_m)$. □

Theorem 3.8. *The feature w_{i+1} extracted by GLDA-TRA is the optimal feature maximizing $F(W_{i+1})$ in (3) when $W_i = [w_1 \ ... \ w_i]$ is extracted.*

Proof. Note that $w_{i+1} \in span\{W_i\}^\perp$ as w_{i+1} is orthonormal to $span\{W_i\}$. Consider an arbitrary $\tilde{w}_{i+1} \neq w_{i+1} \in span\{W_i\}^\perp$. Construct $W_{i+1} = [W_i \ w_{i+1}]$ and $\tilde{W}_{i+1} = [W_i \ \tilde{w}_{i+1}]$. 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

$$\begin{aligned}&F(W_{i+1}) - F(\tilde{W}_{i+1})\\ &= \tfrac{w_1S_Bw_1+...+w_iS_Bw_1+w_{i+1}S_Bw_{i+1}}{w_1S_Ww_1+...+w_1S_Ww_1+w_{i+1}S_Ww_{i+1}} - \tfrac{w_1S_Bw_1+...+w_iS_Bw_1+\tilde{w}_{i+1}S_B\tilde{w}_{i+1}}{w_1S_Ww_1+...+w_1S_Ww_1+\tilde{w}_{i+1}S_W\tilde{w}_{i+1}}\\ &= \tfrac{w_{i+1}S_Bw_{i+1}\tilde{w}_{i+1}S_W\tilde{w}_{i+1}-\tilde{w}_{i+1}S_B\tilde{w}_{i+1}w_{i+1}S_Ww_{i+1}}{(w_1S_Bw_1)^2+...+(w_iS_Bw_1)^2+(w_{i+1}S_Bw_{i+1})(\tilde{w}_{i+1}S_W\tilde{w}_{i+1})}\end{aligned} \tag{17}$$

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

$$\begin{aligned} w_{i+1}S_Bw_{i+1} &= w_{i+1}((I-W_iW_i^+)S_B+W_iW_i^+S_B)w_{i+1} \\ &= w_{i+1}((I-W_iW_i^+)S_B(I-W_iW_i^+))w_{i+1} \\ &= w_{i+1}S_{B_{i+1}}w_{i+1} \end{aligned} \tag{18}$$

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

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

where $S_{W_{i+1}} = ((I-W_iW_i^+)S_W(I-W_iW_i^+))$. From Theorem (3.5), it is noted that $\frac{tr(wS_{B_{i+1}}w)}{tr(wS_{W_{i+1}}w)}$ 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 $S_{W_{i+1}}$. Then, as long as $w_{i+1} \neq \tilde{w}_{i+1}$, we have

$$\frac{tr(w_{i+1}S_{B_{i+1}}w_{i+1})}{tr(w_{i+1}S_{W_{i+1}}w_{i+1})} > \frac{tr(\tilde{w}_{i+1}S_{B_{i+1}}\tilde{w}_{i+1})}{tr(\tilde{w}_{i+1}S_{W_{i+1}}\tilde{w}_{i+1})} \tag{20}$$

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

4. Simulation and Discussion

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

Example 1. *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 Wideth, Petal length and Petal width.*

One can construct S_W *and* S_B *based on the Iris data set. It can be checked that* $rank(S_W) = 4$ *and* $rank(S_B) = C - 1 = 2$*, 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* $W =$

$[w_1\ w_2\ \dots\ w_m]$. *When* $m = 4$, *the obtained* $W = [w_1\ w_2\ w_3\ w_4]$ *is given as*

$$W = \begin{bmatrix} -0.2062 & 0.2978 & 0.8610 & -0.3570 \\ -0.5688 & -0.1773 & 0.2425 & 0.7656 \\ 0.4668 & -0.7879 & 0.3999 & 0.0376 \\ 0.6450 & 0.5090 & 0.1997 & 0.5338 \end{bmatrix}$$

Obviously, $W'W = I$. *Figure 1 shows the distribution of the three-class of*

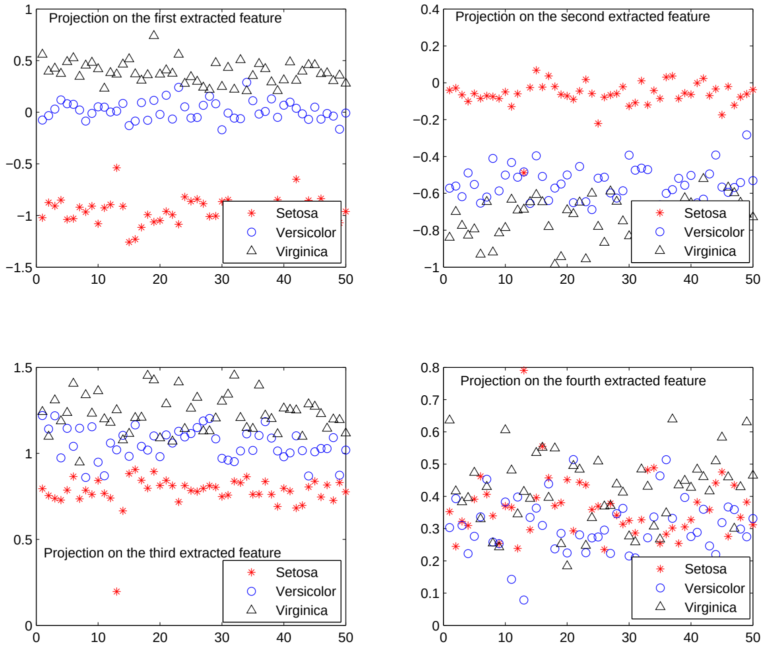

Figure 1: Illustration of projecting data points onto the extracted feature in order

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. *The data set can be best separated in the direction of the first extract feature. One can also note that the sequence* $\{F(w_i)\}_{i=1}^{4}$ *is strictly decreasing as shown in Figure 2.*

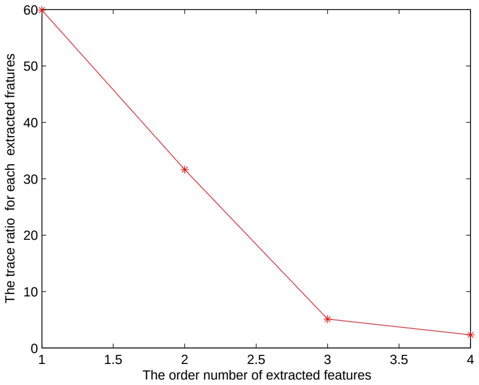

Figure 2: The trace ration for each extracted feature in order

In Example 1, both GLDA-TRA and LDA can obtain good results as the data points are well separated even if they are projected to one dimensional space. If the data are inseparable when the data are projected to one dimensional space, GLDA-TRA will definitely outperform LDA, as illustrated in the following example.

Example 2.

$$(x, y) \in R^3 \times \mathcal{Y} = \begin{cases} \|x\|_2 = |1 + v_i| & \text{if } y = 1 \\ \|x\|_2 = |2 + v_i| & \text{if } y = 2, \end{cases}$$

where $\mathcal{Y} = \{1, 2\}$ 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 1 while data points with label 2 mostly locate around the surface of a sphere with radius 2. Note that LDA does not perform well if it is used as a method to extract features in this problem. This is because LDA can only extract one feature and the data points are inseparable in an arbitrary one dimensional feature space. But with GLDA-TRA, the data points can be better separated if we extract m $(m > 1)$ features. For illustration, let $m = 2$. Then we apply the features extracted by both GLDA-TRA and LAD to do classification and it can be obtain that the classification errors for GLDA-TRA and LAD are about 11.5% and 35.5%, respectively.

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 sub-space 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.

References

[1] K. Fukunaga, “ Introduction to Statistical Pattern Recognition, Second edition, ” *Academic Press*, 1990.

[2] X. Li, W. Hu, H. Wang, and Z. Zhang, “ Linear discriminantanalysisusingrotationalinvariant L1 norm, ” *Neurocompting*, Vol.73, pp.2571-2579, 2010.

[3] S. Noushatha, G. H. Kumara, and P. Shivakumara, “ Diagonal Fisher linear discriminant analysis for efficient face recognition, ” *Neurocompting*, Vol.69, 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, pp. 228-233, 2001.

[5] L. Zhou, L. Wang, and C. Shen, “ Feature selection with redundancy-constrained class separability,” *IEEE Transactions on Neural Network*, vol.21, pp. 853-858, 2010.

[6] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, “Trace ratio criterion for feature selection,” *in Proceeding of 23rd AAAI Conference on Artifical Intellgence*, pp. 671-676, 2008.

[7] 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, p-p. 711-720, 1997.