Journal of Machine Learning Research 9 (2008) 517-519 Submitted 10/07; Revised 12/07; Published 3/08

Comments on the Complete Characterization of a Family of Solutions to a Generalized *Fisher* Criterion

Jieping Ye JIEPING.YE@ASU.EDU
Department of Computer Science and Engineering
Arizona State University
Tempe, AZ 85287, USA

Editor: Xiaotong Shen

Abstract

Loog (2007) provided a complete characterization of the family of solutions to a generalized *Fisher* criterion. We show that this characterization is essentially equivalent to the original characterization proposed in Ye (2005). The computational advantage of the original characterization over the new one is discussed, which justifies its practical use.

Keywords: linear discriminant analysis, dimension reduction, linear transformation

1. Generalized Fisher Criterion

For a given data set consisting of n data points $\{a_i\}_{i=1}^n$ in $\mathbb{R}^d$, a linear transformation $G \in \mathbb{R}^{d\times\ell}$ ($\ell < d$) maps each a_i for $1 \le i \le n$ in the d-dimensional space to a vector $\tilde{a}_i$ in the ℓ-dimensional space as follows:

$$G : a_i \in \mathbb{R}^d \to \tilde{a}_i = G^T a_i \in \mathbb{R}^\ell.$$

Assume that there are k classes in the data set. The within-class scatter matrix S_w, the between-class scatter matrix S_b, and the total scatter matrix S_t involved in linear discriminant analysis are defined as follows (Fukunaga, 1990):

$$\begin{aligned} S_w &= \sum_{i=1}^k (A_i - c_i e^T)(A_i - c_i e^T)^T, \\ S_b &= \sum_{i=1}^k n_i (c_i - c)(c_i - c)^T, \\ S_t &= \sum_{i=1}^k (A_i - ce^T)(A_i - ce^T)^T, \end{aligned}$$

where A_i denotes the data matrix of the i-th class, $c_i = A_i e/n_i$ is the centroid of the i-th class, n_i is the sample size of the i-th class, $c = Ae/n$ is the global centroid, and e is the vector of all ones with an appropriate length. It is easy to verify that $S_t = S_b + S_w$.

In Ye (2005), the optimal transformation G is computed by maximizing a generalized *Fisher* criterion as follows:

$$G = \arg\max_{G\in\mathbb{R}^{m\times\ell}} \operatorname{trace}\left(\left(G^T S_t G\right)^+ G^T S_b G\right), \tag{1}$$

where M^+ denotes the pseudo-inverse (Golub and Van Loan, 1996) of M and it is introduced to overcome the singularity problem when dealing with high-dimensional low-sample-size data.

1.1 Equivalent Transformation

Two linear transformations G_1 and G_2 can be considered equivalent if there is a vector v such that $G_1^T(a_i - v) = G_2^T(a_i - v)$, for $i = 1, \cdots, n$. Indeed, in this case, the difference between the projections by G_1 and G_2 is a mere shift.

Definition 1.1 *For a given data set $\{a_1, \cdots, a_n\}$, two transformations G_1 and G_2 are equivalent, if there is a vector v such that*

$$G_1^T(a_i - v) = G_2^T(a_i - v), \text{ for } i = 1, \cdots, n.$$

2. Characterization of Solutions to the Generalized *Fisher* Criterion

Let $S_t = U\Sigma U^T$ be the orthogonal eigendecomposition of S_t (note that S_t is symmetric and positive semi-definite), where $U \in \mathbb{R}^{d\times d}$ is orthogonal and $\Sigma \in \mathbb{R}^{d\times d}$ is diagonal with nonnegative diagonal entries sorted in nonincreasing order. Denote Σ_r as the r-th principal submatrix of Σ, where $r = \text{rank}(S_t)$. Partition U into two components as $U = [U_1, U_2]$, where $U_1 \in \mathbb{R}^{d\times r}$ and $U_2 \in \mathbb{R}^{d\times(d-r)}$. Note that $r \leq n$, and for high-dimensional low-sample-size data, U_1 is much smaller than U_2.

In Loog (2007), a complete family of solutions $\mathcal{S}$ to the maximization problem in Eq. (1) is given as (We correct the error in Loog (2007) by using U instead of U^T.)

$$\mathcal{S} = \left\{ U \begin{pmatrix} \Lambda Z \\ Y \end{pmatrix} \in \mathbb{R}^{d\times\ell} \;\middle|\; Z \in \mathbb{R}^{\ell\times\ell} \text{ is nonsingular }, \; Y \in \mathbb{R}^{(n-r)\times\ell} \right\},$$

where $\Lambda \in \mathbb{R}^{r\times\ell}$ maximizes the following objective function:

$$F_0(X) = \text{trace}\left(\left(X^T\Sigma_r X\right)^{-1} X^T (U_1^T S_b U_1) X \right).$$

In Ye (2005), a family of solutions $\tilde{\mathcal{S}}$ is given as

$$\tilde{\mathcal{S}} = \left\{ U \begin{pmatrix} \Lambda Z \\ 0 \end{pmatrix} \in \mathbb{R}^{d\times\ell} \;\middle|\; Z \in \mathbb{R}^{\ell\times\ell} \text{ is nonsingular } \right\}.$$

The only difference between these two characterizations of solutions is the matrix Y in $\mathcal{S}$, which is replaced by the zero matrix in $\tilde{\mathcal{S}}$. We show in the next section the equivalence relationship between these two characterizations.

3. Equivalent Solution Characterizations

Consider the following two transformations G_1 and G_2 from $\mathcal{S}$ and $\tilde{\mathcal{S}}$ respectively:

$$G_1 = U \begin{pmatrix} \Lambda Z \\ Y \end{pmatrix} \in \mathcal{S}, \quad G_2 = U \begin{pmatrix} \Lambda Z \\ 0 \end{pmatrix} \in \tilde{\mathcal{S}}.$$

Recall that $U = [U_1, U_2]$, where the columns of U_2 span the null space of S_t. Hence,

$$0 = U_2^T S_t U_2 = \sum_{i=1}^{n} U_2^T (a_i - c) \cdot (U_2^T (a_i - c))^T,$$

and $U_2^T (a_i - c) = 0$, for $i = 1, \cdots, n$, where c is the global centroid. It follows that

$$G_1^T (a_i - c) = Z^T \Lambda^T U_1^T (a_i - c) + Y^T U_2^T (a_i - c) = Z^T \Lambda^T U_1^T (a_i - c) = G_2^T (a_i - c),$$

for $i = 1, \cdots, n$. That is, G_1 and G_2 are equivalent transformations. Hence, the two solution characterizations $\mathcal{S}$ and $\tilde{\mathcal{S}}$ are essentially equivalent.

Remark 3.1 *The analysis above shows that the additional information contained in $\mathcal{S}$ is the null space, U_2, of S_t, which leads to an equivalent transformation. In $\tilde{\mathcal{S}}$, the null space U_2 is removed, which can be further justified as follows. Since $S_t = S_b + S_w$, we have*

$$0 = U_2^T S_t U_2 = U_2^T S_b U_2 + U_2^T S_w U_2.$$

It follows that $U_2^T S_b U_2 = 0$, as both S_b and S_w are positive semi-definite. Thus, the null space U_2 does not contain any discriminant information. This explains why the null space of S_t is removed in most discriminant analysis based algorithms proposed in the past.

4. Efficiency Comparison

In $\mathcal{S}$, the full matrix U is involved, whose computation may be expensive, especially for high-dimensional data. In contrast, only the first component $U_1 \in \mathbb{R}^{d \times r}$ of U is involved in $\tilde{\mathcal{S}}$, which can be computed efficiently for high-dimensional low-sample-size problem by directly working on the Gram matrix instead of the covariance matrix.

In summary, we show that $\mathcal{S}$ and $\tilde{\mathcal{S}}$ are equivalent characterizations of the solutions to the generalized *Fisher* criterion in Eq. (1). However, the latter one is preferred in practice due to its relative efficiency for high-dimensional low-sample-size data.

References

K. Fukunaga. *Introduction to Statistical Pattern Classification*. Academic Press, San Diego, California, USA, 1990.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996.

M. Loog. A Complete Characterization of a Family of Solutions to a Generalized Fisher Criterion. *Journal of Machine Learning Research*, 8:2121–2123, 2007.

J. Ye. Characterization of a Family of Algorithms for Generalized Discriminant Analysis on Undersampled Problems. *Journal of Machine Learning Research*, 6:483–502, 2005.