ELSEVIER

The MDF discrimination measure: Fisher in disguise

Marco Loog[a,*], Robert P.W. Duin[b], Max A. Viergever[a]

[a]Image Sciences Institute, University Medical Center Utrecht, HP E01.334, P.O. Box 85500, 3508 GA Utrecht, The Netherlands
[b]Pattern Recognition Group, Faculty of Applied Sciences, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, The Netherlands

Abstract

Recently, a discrimination measure for feature extraction for two-class data, called the maximum discriminating (MDF) measure (Talukder and Casasent [Neural Networks 14 (2001) 1201–1218]), was introduced.

In the present paper, it is shown that the MDF discrimination measure produces *exactly* the same results as the classical Fisher criterion, on the condition that the two prior probabilities are chosen to be equal. The effect of unequal priors on the efficiency of the measures is also discussed.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: Discrimination measure; Fisher criterion; Fisher linear discriminant; Maximum discriminating features; Linear discriminant analysis; Fisher mapping; Prior probability

1. Introduction

In their article "A closed-form neural network for discriminatory feature extraction from high-dimensional data" (Talukder & Casasent, 2001), Talukder and Casasent develop a nonlinear maximum discriminating feature (MDF) neural network that is capable of extracting features from high-dimensional data. An important ingredient in the approach is the MDF discrimination measure they propose. Basically, maximizing this measure gives a subspace in which *all pairs* of samples from two different classes are separated in a certain optimal way.

Although the MDF measure is used for extracting nonlinear features, it essentially provides a linear feature subspace. Nonlinear features can be constructed by transforming the original features nonlinearly[1] before and/or after the actual feature extraction. The nonlinear pre- and postprocessing and the linear dimension reducing transformation jointly provide a nonlinear feature extraction method. Here, we focus on the linear feature extraction technique, the basic building block of the closed-form neural network.

Talukder and Casasent expect the MDF discrimination measure to be a better measure of separation than, for example, the well-known Fisher linear discriminant (FLD)[2] which is based on the Fisher criterion (Fukunaga, 1990; McLachlan, 1992). In this communication, however, we demonstrate that features extracted by using MDF are essentially equal to the solution produced by the Fisher criterion. More specific, the features obtained are *exactly* equal if equal prior probabilities are assumed.

We note that in Talukder (1999), Talukder already proved that, for the two-class case, the MDF and FLD approaches coincide if both classes are *normally distributed with equal covariance matrices*. Our result is stronger and shows that, given equal prior class probabilities, both feature extraction techniques *always* coincide, *independent of the underlying class distributions*.

1.1. Outline

Section 2 starts with a brief description of the MDF discrimination measure and gives the solution to

---

* Corresponding author. Tel.: +31-30-250-7772; fax: +31-30-251-3399.
E-mail address: marco@isi.uu.nl (M. Loog).
[1] In Talukder and Casasent (2001) polynomial transformations are considered, which are applied to the initial feature vectors prior to the linear feature extraction.

[2] Also known as linear discriminant analysis (LDA) or Fisher mapping.

optimizing this measure. Furthermore, it presents the Fisher criterion, whose maximization gives rise to the FLD. Section 3.1 then establishes the equivalence of MDF and FLD in the two-class case. While Section 3.1 merely deals with the case in which the prior probabilities are taken to be equal, Section 3.2 discusses the case in which classes do not necessarily have equal priors. In both cases, however, no assumption on the underlying distributions is made. It also presents an illustration, based on an example coming from Talukder and Casasent (2001), of the difference in feature spaces obtained by MDF and Fisher. Section 4 provides the conclusions and the final discussion. To keep the communication readable, the central proof is given in Appendix A.

In the following, we use the notation and terminology similar to the one in Talukder and Casasent's article.

2. The MDF discrimination measure and FLD

The goal is to linearly reduce an H-dimensional feature space, in which two-classes reside, to an M-dimensional feature space. The linear mapping describing this linear dimension reduction is denoted by $\underline{\Phi}_M$, which is an $H \times M$ matrix consisting of M H-dimensional vectors $\underline{\phi}_1, ..., \underline{\phi}_M$, i.e. $\underline{\Phi}_M = [\underline{\phi}_1 \underline{\phi}_2 ... \underline{\phi}_M]$.

2.1. MDF discrimination measure

Following Talukder and Casasent, the best transformation vectors $\underline{\phi}_m (m \in \{1, ..., M\})$ are those that maximize the MDF discrimination measures E_D (Talukder & Casasent, 2001)

$$E_D = \sum_{m=1}^{M} \frac{\underline{\phi}_m^T \underline{R}_{12} \underline{\phi}_m}{\underline{\phi}_m^T (\underline{C}_1 + \underline{C}_2) \underline{\phi}_m}, \tag{1}$$

where $\underline{C}_1 := E[\mathbf{x}_1 \mathbf{x}_1^T] - E[\mathbf{x}_1]E[\mathbf{x}_1]^T$ and $\underline{C}_2 := E[\mathbf{x}_2 \mathbf{x}_2^T] - E[\mathbf{x}_2]E[\mathbf{x}_2]^T$ are the covariance matrices of class 1 and 2, respectively, $\underline{R}_{12} := E[(\mathbf{x}_1 - \mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)^T]$ is a vector-outer-product difference matrix, and $\mathbf{x}_k$ is a random (feature) vector from class k.

The mapping $\underline{\Phi}_M$ that best separates the two-classes according to the MDF criterion in Eq. (1) must satisfy the generalized eigenvalue decomposition (Talukder & Casasent, 2001)

$$[\underline{C}_1 + \underline{C}_2]^{-1} \underline{R}_{12} \underline{\Phi}_M = \underline{\Phi}_M \underline{\Lambda},$$

and so, the M best MDF basis functions $\underline{\phi}_m (m \in \{1, ..., M\})$ are the M dominant eigenvectors of

$$[\underline{C}_1 + \underline{C}_2]^{-1} \underline{R}_{12}. \tag{2}$$

2.2. FLD and the Fisher criterion

Another way of discriminatory feature extraction is to determine the FLD through optimizing the so called Fisher criterion[3]. This criterion equals (Fukunaga, 1990; McLachlan, 1992)

$$\text{trace}((\underline{\Phi}_M^T (p_1 \underline{C}_1 + p_2 \underline{C}_2) \underline{\Phi}_M)^{-1} (\underline{\Phi}_M^T \underline{B} \underline{\Phi}_M)),$$

where $\underline{B} := E[\underline{\mu}_i \underline{\mu}_i^T] - E[\underline{\mu}_i]E[\underline{\mu}_i]^T$ is the between-class covariance matrix, $\underline{\mu}_i$ is the mean of class i, and p_i is the prior probability of class i.

The matrix $p_1 \underline{C}_1 + p_2 \underline{C}_2$ in the Fisher criterion equals the average within-class covariance matrix, and so, in determining the FLD, one maximizes the ratio of the between-class covariance over the average within-class covariance in the lower-dimensional space. The solution to this optimization problem is, as in the MDF case, obtained by solving a generalized eigenvalue problem (see (Fukunaga, 1990))

$$[p_1 \underline{C}_1 + p_2 \underline{C}_2]^{-1} \underline{B} \underline{\Phi}_M = \underline{\Phi}_M \underline{\Lambda}.$$

The M basis functions $\underline{\Phi}_m$ ($m \in \{1, ..., M\}$) that maximize this criterion are the M dominant eigenvectors of the matrix

$$[p_1 \underline{C}_1 + p_2 \underline{C}_2]^{-1} \underline{B}. \tag{3}$$

3. The equivalence of MDF and FLD

Looking at the foregoing section, we see that the MDF measure maximizes the mean squared separation between *all* samples in class 1 and class 2, while the Fisher criterion maximizes the mean squared separation between the two-class means. These objectives, however, turn out to be the same.

3.1. Establishing the equivalence

Establishing the equivalence of the FLD and MDF approach to linear dimension reduction is based on the following relation between the matrices $\underline{B}$ and $\underline{R}_{12}$.

$$\underline{B} = p_1 p_2 (\underline{R}_{12} - \underline{C}_1 - \underline{C}_2). \tag{4}$$

We refer to Appendix A for a proof of Eq. (4). Now, consider the matrix in Eq. (3), and take both prior probabilities p_1 and p_2 to be equal, i.e. set $p_1 = p_2 = \frac{1}{2}$. Then, using Eq. (4), we see that obtaining the dominant eigenvectors of $[(\frac{1}{2})\underline{C}_1 + (\frac{1}{2})\underline{C}_2]^{-1} \underline{B}$ is equivalent to determining the dominant eigenvectors of $[(\frac{1}{2})\underline{C}_1 + (\frac{1}{2})\underline{C}_2]^{-1} 1/4(\underline{R}_{12} - \underline{C}_1 - \underline{C}_2)$, which in turn comes down to determining the dominant eigenvectors of $[\underline{C}_1 + \underline{C}_2]^{-1} \underline{R}_{12}$. To see this, let $\underline{\phi}$ be an arbitrary eigenvector of $[\underline{C}_1 + \underline{C}_2]^{-1} \underline{R}_{12}$ with associated eigenvalue λ. For this

---

[3] The Fisher criterion can be defined in several different ways, however, they all lead to the same linear subspace as a solution. For specific examples of different definitions we refer to Fukunaga (1990).

vector, the following holds:

$$[\tfrac{1}{2}\underline{C}_1 + \tfrac{1}{2}\underline{C}_2]^{-1}\tfrac{1}{4}(\underline{R}_{12} - \underline{C}_1 - \underline{C}_2)\underline{\phi}$$

$$= [\tfrac{1}{2}\underline{C}_1 + \tfrac{1}{2}\underline{C}_2]^{-1}\tfrac{1}{4}\underline{R}_{12}\underline{\phi} - [\tfrac{1}{2}\underline{C}_1 + \tfrac{1}{2}\underline{C}_2]^{-1}\tfrac{1}{4}(\underline{C}_1 + \underline{C}_2)\underline{\phi}$$

$$= \tfrac{1}{2}[\underline{C}_1 + \underline{C}_2]^{-1}\underline{R}_{12}\underline{\phi} - \tfrac{1}{2}\underline{I}\underline{\phi} = \tfrac{1}{2}\lambda\underline{\phi} - \tfrac{1}{2}\underline{\phi} = (\tfrac{1}{2}\lambda - \tfrac{1}{2})\underline{\phi}.$$

That is, the vector $\underline{\phi}$ is an eigenvector of $[(\tfrac{1}{2})\underline{C}_1 + (\tfrac{1}{2})\underline{C}_2]^{-1}\underline{B}$ with associated eigenvalue λ *if and only if* $\underline{\phi}$ is also an eigenvector of $[\underline{C}_1 + \underline{C}_2]^{-1}\underline{R}_{12}$ with associated eigenvalue $(\tfrac{1}{2})\lambda - (\tfrac{1}{2})$. From this, it directly follows that the FLD approach and the MDF approach give the same eigenvectors with the same ordering based on the associated eigenvalues.

This establishes the equivalence of the MDF discrimination measure and the Fisher criterion. Both matrices provide the same dominant eigenvectors, and so the optimization of these measures results in two equivalent linear transformations $\underline{\Phi}_M$ mapping the feature vectors to the same linear subspace.

3.2. Unequal priors

In Section 3.1, both class priors were taken to be equal. For the general two-class case in which prior probabilities may differ, the solutions to the MDF discrimination measure and the Fisher criterion do not necessarily coincide. Because of the result of Talukder (1999) mentioned in Section 1, and the result presented in this communication, we expect that the difference in performance of both approaches might be appreciable only if the prior probabilities differ significantly. However, even a large difference in prior probabilities does not necessarily lead to an appreciable difference in performance, as is illustrated in the following.

We present an example, taken from Talukder and Casasent (2001), with a two-dimensional data configuration. For this data, Fisher extracts a single feature that is bad for discriminatory and classification purposes (Fig. 1). Optimizing the Fisher criterion produces a one-dimensional subspace that is close to vertical (the dashed line in Fig. 1). Projecting the data to this single dimension results in some overlap between the two-classes. In addition, the single MDF feature is also extracted. As our figure shows both approaches do not differ visibly, although the priors are very different (0.35 and 0.65, respectively). (See the caption to Fig. 1. Compare this figure also to Figure 2 in Talukder and Casasent (2001).)

4. Conclusion and discussion

Talukder and Casasent's approach to nonlinear feature extraction given in Talukder and Casasent (2001) is, as
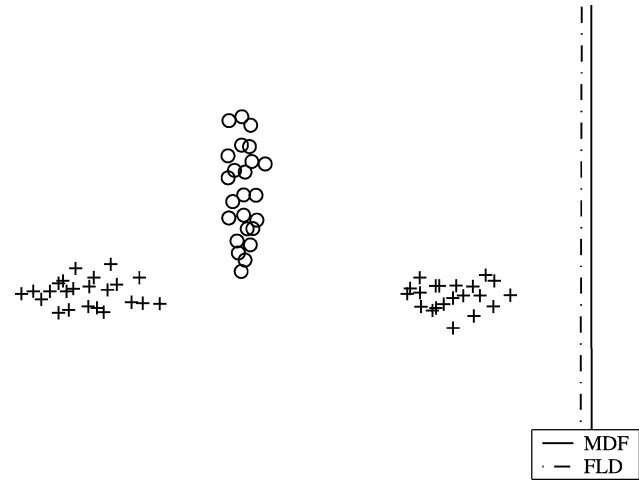

Fig. 1. Comparison of feature extracted by means of the MDF measure (solid line) and the Fisher criterion (dashed line) based on two-class data taken from Talukder and Casasent (2001) (Figure 2). In this example the priors differ. The circles have a prior probability of 0.35, while the crosses' prior equals 0.65. The difference between the obtained subspaces by means of the MDF method and FLD method is indiscernible. The difference between the first eigenvectors is very close to the null vector: $\underline{\phi}_1^{\text{MDF}} - \underline{\phi}_1^{\text{Fisher}} \approx [-0.0037, 0.0000]^{\text{T}}$.

a whole, interesting in its own respect. Furthermore, the theoretical issues they additionally discussed concerning their transformation and other nonlinear transforms is valuable, and provides insight into the behavior of the discussed transformations.

However, we noted and demonstrated that the MDF discrimination measure they introduce is exactly equal to the Fisher criterion, associated with the well-known FLD, on the condition that both priors are chosen to be equal. Hence both feature extraction techniques can provide the same features. Apart from the equality of the prior probabilities, no other assumptions were made and as such our result generalizes the equivalence result in Talukder (1999), Appendix A.

Although this exact equivalence is only shown to hold when the prior probabilities of the classes are taken to be equal, there is no reason to support the belief that the MDF approach outperforms the FLD approach. On the contrary, because the Fisher criterion can take differences in priors into account, it is expected to perform better than the MDF measure. Moreover, when using the Fisher criterion, one can choose the priors to be equal in which case FLD and MDF coincide. The extra degree of freedom in the Fisher criterion, provided by the choice of priors, gives one the opportunity to improve the performance of the FLD, which is not possible with the MDF discrimination measure. As such, the MDF measure can be considered a specific instance of the Fisher criterion.

Appendix A. Proof of Eq. (4)

The basis for the observation that $\underline{B}$ equals $p_1p_2(\underline{R}_{12} - \underline{C}_1 - \underline{C}_2)$ (Eq. (4)) comes from a reformulation of a theorem from Loog (1999). This theorem gives an alternative view on determining and representing covariance matrices based on pairwise differences between feature vectors (see also (Loog, Duin, & Haeb-Umbach, 2001)). The (re)formulation is as follows: Let $\underline{C} := E[\mathbf{x}\mathbf{x}^T] - E[\mathbf{x}]E[\mathbf{x}]^T$ be the covariance matrix for the random vector $\mathbf{x}$, then $\underline{C}$ can be written as half the expectation over all outer products of pairwise differences between the independent identically distributed (i.i.d.) random vectors $\mathbf{x}$ and $\mathbf{y}$, i.e.

$$\underline{C} = \tfrac{1}{2}E[(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T]. \qquad (5)$$

The proof is straightforward (cf. (Loog, 1999; Loog et al., 2001)). Expanding Eq. (5) gives $(\tfrac{1}{2})E[(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T] = (\tfrac{1}{2})E[\mathbf{x}\mathbf{x}^T + \mathbf{y}\mathbf{y}^T - \mathbf{y}\mathbf{x}^T - \mathbf{x}\mathbf{y}^T]$. Because $\mathbf{x}$ and $\mathbf{y}$ are i.i.d., this equals $\tfrac{1}{2}(E[\mathbf{x}\mathbf{x}^T] + E[\mathbf{y}\mathbf{y}^T] - E[\mathbf{x}]E[\mathbf{x}]^T - E[\mathbf{y}]E[\mathbf{y}]^T) = \tfrac{1}{2}(2E[\mathbf{x}\mathbf{x}^T] - 2E[\mathbf{x}]E[\mathbf{x}]^T)$, which in turn equals $E[\mathbf{x}\mathbf{x}^T] - E[\mathbf{x}]E[\mathbf{x}]^T =: \underline{C}$.

Now, using that the sum of the between-class covariance matrix $\underline{B}$ and the average within-class covariance matrix $p_1\underline{C}_1 + p_2\underline{C}_2$ equals the total covariance matrix $\underline{T}$ (i.e. the covariance matrix over all data points irrespective of their class), we see that the following holds

$$
\begin{aligned}
\underline{B} &= \underline{T} - p_1\underline{C}_1 - p_2\underline{C}_2 \\
&= \frac{1}{2}E[(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T] - \frac{p_1}{2}E[(\mathbf{x}_1 - \mathbf{y}_1)(\mathbf{x}_1 - \mathbf{y}_1)^T] \\
&\quad - \frac{p_2}{2}E[(\mathbf{x}_2 - \mathbf{y}_2)(\mathbf{x}_2 - \mathbf{y}_2)^T] \\
&= \frac{1}{2}(E[(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T] - (p_1p_1 + p_2p_1)E[(\mathbf{x}_1 - \mathbf{y}_1) \\
&\quad \times (\mathbf{x}_1 - \mathbf{y}_1)^T] - (p_1p_2 + p_2p_2)E[(\mathbf{x}_2 - \mathbf{y}_2)(\mathbf{x}_2 - \mathbf{y}_2)^T]).
\end{aligned}
\qquad (6)
$$

Furthermore, the matrix $E[(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T]$, in which the random vectors $\mathbf{x}$ and $\mathbf{y}$ go over both classes, can be split up in terms of random vectors $\mathbf{x}_1$, $\mathbf{y}_1$, $\mathbf{x}_2$, and $\mathbf{y}_2$ coming from class 1 and class 2, respectively:

$$
\begin{aligned}
&E[(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T] \\
&= p_1p_1E[(\mathbf{x}_1 - \mathbf{y}_1)(\mathbf{x}_1 - \mathbf{y}_1)^T] + p_2p_2E[(\mathbf{x}_2 - \mathbf{y}_2) \\
&\quad \times (\mathbf{x}_2 - \mathbf{y}_2)^T] + p_1p_2E[(\mathbf{x}_1 - \mathbf{y}_2)(\mathbf{x}_1 - \mathbf{y}_2)^T] + p_2p_1 \\
&\quad \times E[(\mathbf{x}_2 - \mathbf{y}_1)(\mathbf{x}_2 - \mathbf{y}_1)^T].
\end{aligned}
\qquad (7)
$$

Finally, substituting Eq. (7) into Eq. (6) and rearranging terms gives the following

$$
\begin{aligned}
\underline{B} &= \frac{1}{2}(p_1p_1E[(\mathbf{x}_1 - \mathbf{y}_1)(\mathbf{x}_1 - \mathbf{y}_1)^T] \\
&\quad + p_2p_2E[(\mathbf{x}_2 - \mathbf{y}_2)(\mathbf{x}_2 - \mathbf{y}_2)^T] \\
&\quad + p_1p_2E[(\mathbf{x}_1 - \mathbf{y}_2)(\mathbf{x}_1 - \mathbf{y}_2)^T] \\
&\quad + p_2p_1E[(\mathbf{x}_2 - \mathbf{y}_1)(\mathbf{x}_2 - \mathbf{y}_1)^T] \\
&\quad - (p_1p_1 + p_2p_1)E[(\mathbf{x}_1 - \mathbf{y}_1)(\mathbf{x}_1 - \mathbf{y}_1)^T] \\
&\quad - (p_1p_2 + p_2p_2)E[(\mathbf{x}_2 - \mathbf{y}_2)(\mathbf{x}_2 - \mathbf{y}_2)^T]) \\
&= p_1p_2E[(\mathbf{x}_1 - \mathbf{y}_2)(\mathbf{x}_1 - \mathbf{y}_2)^T] - \frac{p_1p_2}{2}E[(\mathbf{x}_1 - \mathbf{y}_1) \\
&\quad \times (\mathbf{x}_1 - \mathbf{y}_1)^T] - \frac{p_1p_2}{2}E[(\mathbf{x}_2 - \mathbf{y}_2)(\mathbf{x}_2 - \mathbf{y}_2)^T] \\
&= p_1p_2(\underline{R}_{12} - \underline{C}_1 - \underline{C}_2),
\end{aligned}
$$

which proofs Eq. (4).

References

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. New York: Academic Press.

Loog, M. (1999). *Approximate pairwise accuracy criteria for multiclass linear dimension reduction: Generalisations of the Fisher criterion. Number 44 in WBBM report series*, Delft: Delft University Press.

Loog, M., Duin, R. P. W., & Haeb-Umbach, R. (2001). Weighted pairwise linear discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(7), 762–766.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley.

Talukder, A. (1999). *Nonlinear feature extraction for pattern recognition applications*. PhD Thesis, Carnegie Mellon University.

Talukder, A., & Casasent, D. (2001). A closed-form neural network for discriminatory feature extraction from high-dimensional data. *Neural Networks*, *14*, 1201–1218.