Improving Feature Extraction by Replacing the Fisher Criterion by an Upper Error Bound [*]

K. Luebke [†] C. Weihs

June 2005

Universität Dortmund
Fachbereich Statistik

Abstract

A lot of alternatives and constraints have been proposed in order to improve the Fisher criterion. But most of them are not linked to the error rate, the primary interest in many applications of classification. By introducing an upper bound for the error rate a criterion is developed which can improve the classification performance.

Keywords
Fisher criterion, Linear discriminant analysis, Feature extraction

1 Introduction

Linear feature extraction is a valuable preprocessing step in a classification problem. It can be useful for visualization of data or to avoid problems connected with overfitting and unstable estimates as well to save storage space. One of the most well-known selection criteria for a projection is the Fisher criterion. It evaluates the between-class variance relative to the within-class

[*] **Preprint.** Paper to appear in Pattern Recognition

[†] e-mail: luebke@statistik.uni-dortmund.de

variance. In the literature different constraints are introduced in order to get the best value in the Fisher criterion (see for example [2]). But the Fisher criterion is just an heuristic substitute for the main goal, minimal error rate. So a good value in the criterion need not lead to a good error rate [4]. As it is impossible in most cases to calculate the error rate directly an upper bound in the projection space is used in order to obtain an improved feature extraction criterion.

2 Criteria for Feature Extraction

Suppose that there are K known classes in a d dimensional problem with mean μ_i, covariance matrix Σ_i and a priori probability π_i of class i ($i = 1, 2, \ldots, K$). Then let $\mu_0 = \sum_{i=1}^{K} \pi_i \mu_i$ be the overall mean and

$$\begin{aligned} \Sigma_B &= \sum_{i=1}^{K} \pi_i (\mu_i - \mu_0)(\mu_i - \mu_0)^T, & (1) \\ \Sigma_W &= \sum_{i=1}^{K} \pi_i \Sigma_i, & (2) \\ \Sigma_T &= \Sigma_B + \Sigma_W & (3) \end{aligned}$$

are the within-class, between-class and total covariance matrices. The Fisher criterion for a set of so-called discriminant vectors $\gamma_j, j = 1, 2, \ldots, r$, is given by

$$J_F(\gamma_j) = \frac{\gamma_j^T \Sigma_B \gamma_j}{\gamma_j^T \Sigma_W \gamma_j}. \qquad (4)$$

Optimal γ_j are found by maximizing (4) under different constraints: The constraint used in classical linear discriminant analysis (CDA) is

$$\gamma_i^T \Sigma_W \gamma_j = 0, \quad \forall i \neq j, \quad i, j = 1, 2, \ldots, r. \qquad (5)$$

In the feature extraction literature different alternatives to (5) are used (see [2] and the references therein): For example the Foley-Sammon linear discriminant analysis (FSDA) uses

$$\gamma_i^T \gamma_j = 0, \quad \forall i \neq j, \quad i, j = 1, 2, \ldots, r \qquad (6)$$

or in uncorrelated linear discriminant analysis (UCDA) the constraint

$$\gamma_i^T \Sigma_T \gamma_j, \quad \forall i \neq j, \quad i, j = 1, 2, \ldots, r \qquad (7)$$

is used. It is shown in [2] that with (5) and (7) the resulting vectors are equivalent and that there are at most $r \leq \min(d, K-1)$ different discriminant vectors γ with $J_F(\gamma) > 0$.

Allocation to the classes in a Linear Discriminant Analysis (LDA) is done by

$$a(y) = \arg\max_i \pi_i |2\pi\Sigma_W|^{-0.5} \exp(-0.5(y-\mu_i)^T(\Sigma_W)^{-1}(y-\mu_i)), \quad (8)$$

for any feature vector y and corresponding means and covariances derived from observed data x. In case of a Nearest Mean (NM) Classifier in (8) all π_i are set to $\pi_i = \frac{1}{K}$ and Σ_W is set to $\Sigma_W = I$.

However, (4) is not linked to the misclassification probability. Indeed, maximization of (4) leads to maximization of scatter between the class means which might lead to the problem that classes which are already well separated are separated most in the projection while distances between classes which are relatively close are minimized and so misclassification between these classes is more frequent than necessary [3, p. 93].

3 Optimal Separation Criterion

As in linear feature extraction the generated features are sums of the original features it can be shown that under mild assumptions [1] the new features $Y = X\Gamma$ with data matrix $X \in \mathbb{R}^{n \times d}$ and feature extraction matrix $\Gamma = (\gamma_1, \gamma_2, \cdots, \gamma_r) \in \mathbb{R}^{d \times r}$ are normally distributed. With Lemma 1 this can be utilized to get an upper bound for the Bayes error rate for classification problems after feature extraction. Let

$$\delta(i,j)^2 = (\mu_i - \mu_j)^T \Sigma_W^{-1} (\mu_i - \mu_j) \quad \forall i \neq j, \quad i,j = 1,2,\ldots,K \quad (9)$$

be the squared Mahalanobis distance between the classes i and j.

Lemma 1: Given normally distributed data with means $\mu_1, \ldots, \mu_K$, equal covariance matrices $\Sigma_1 = \cdots = \Sigma_K$, and equal a priori probabilities $\pi_1 = \cdots = \pi_K$ the error rate (err) of (8) is bounded by

$$err \leq \frac{K-1}{K} \sum_{i=1}^{K} \Phi\left(-\frac{1}{2} \min_{j=1,\ldots,K, j\neq i} \delta(i,j)\right), \quad (10)$$

where Φ is the (cumulative) distribution function of the standard normal distribution.
The proof of Lemma 1 is given in the appendix.

The assumption of equal a priori probabilities is not necessary but enables are more simple form of (10).

For multi-class problems this error bound can be calculated quite fast without the use of resampling methods. Moreover it enables a new criterion for feature extraction. Let

$$\delta(i,j|\Gamma)^2 := \left((\mu_i - \mu_j)^T\Gamma\right)\left(\Gamma^T\Sigma_W\Gamma\right)^{-1}\left(\Gamma^T(\mu_i - \mu_j)\right) \tag{11}$$

be the squared Mahalanobis distance in the projected space then the Optimal Separation (OS) criterion can be defined as follows:

$$J_{OS}(\Gamma) := \frac{K-1}{K}\sum_{i=1}^{K}\Phi\left(-\frac{1}{2}\min_{j=1,\dots,K,j\neq i}\delta(i,j|\Gamma)\right). \tag{12}$$

Note that in (12) it is no longer possible to calculate the discriminant vectors stepwise like in Fisher's criterion but the whole projection matrix Γ must be evaluated. The projection matrix that minimize (12) is called the Optimal Separation Projection (OSP).

Unfortunately in order to minimize (12) stochastic optimization methods must be used as the derivative of J_{OS} to Γ does not exist everywhere (caused by the use of $\min(\cdot)$) and also there are local minima which must be overcome. For example Simulated annealing can used to obtain the optimal Γ in terms of (12) and to avoid the aforementioned problems.

4 Experimental Results

The data set consists of 13 economic variables with quarterly observations from 1961/2 to 2002/4 of the German business cycle. The German business cycle is classified in a four phase scheme: upswing, upper turning point, downswing and lower turning point. The last complete business cycle ends in 1994/1, which is the last observation for training. The remaining observations are used as test data in order to calculate the error rate. The training data was used for checking the Fisher (4) as well as the Optimal Separation (12) criterion where optimization of (4) is done under the CDA (5) as well as under the FSDA (6) constraint. Table 1 shows the good performance of the FSDA method for the Fisher criterion and also the bad performance of the Optimal Separation Projection method in terms of the Fisher criterion but the good performance especially for $r = 2$ in terms of the OS criterion. But this is expected as the OSP method aims at minimizing this criterion.

In order to check the classification performance of these feature extraction methods two different classifiers are used: Linear Discriminant Analysis (8)

	Fisher Criterion			OS Criterion		
r	CDA	FSDA	OSP	CDA	FSDA	OSP
1	91.04	91.04	87.60	0.89	0.89	0.88
2	59.63	76.16	44.06	0.40	0.45	0.33
3	17.09	44.16	17.09	0.25	0.44	0.25

Table 1: Optimal Values for Fisher and Optimal Separation criterion

	LDA Classifier			NM Classifier		
r	CDA	FSDA	OSP	CDA	FSDA	OSP
1	0.26	0.26	0.26	0.67	0.67	0.56
2	0.41	0.33	0.15	0.48	0.45	0.26
3	0.22	0.33	0.22	0.19	0.44	0.19

Table 2: Error Rates for feature extraction methods with different classifiers

and Nearest Mean Classifier. One can see that with this data the OSP method performs best in terms of the error rate - even though it is worst in terms of the Fisher criterion. The best error rate which can be achieved is 0.15 with $r = 2$ with OSP and the LDA Classifier.

5 Conclusion

By the (reasonable) assumption of a normal distribution in the projected space an upper bound for the misclassification error can be proofed. With help of this error bound a criterion for feature extraction can be set up which can improve the classification performance of the extracted features compared to other feature extraction criteria used in pattern recognition. Still more research about the closeness of this bound as well as improved methods for calculation of the optimal feature extraction matrices are necessary.

Acknowledgments

This work has been supported by the Collaborative Research Center “Reduction of Complexity for Multivariate Data Structures” (SFB 475) of the German Research Foundation (DFG).

A Proof of Lemma 1

To proof the Lemma first note that in the case of 2 classes the misclassification probability under the assumptions of the Lemma is given by [3, page 61]:

$$\Phi\left(-\frac{1}{2}\left((\mu_1-\mu_2)^T\Sigma_W^{-1}(\mu_1-\mu_2)\right)^{\frac{1}{2}}\right). \tag{13}$$

So it is easy to verify that the probability of assigning an observation to class i when in fact it comes from class j is

$$err(j|i) \leq \Phi\left(-\frac{1}{2}\left((\mu_i-\mu_j)^T\Sigma_W^{-1}(\mu_i-\mu_j)\right)^{\frac{1}{2}}\right). \tag{14}$$

Let $err(j)$ be the probability of assigning an object from class j to a class $i \neq j$ then with (14) and the Bonferroni inequality $P(\bigcup_i A_i) \leq \sum_i P(A_i)$ it follows that

$$\begin{aligned} err(i) &= P(\bigcup_{j\neq i} a(x) = j|class(x) = i) \\ &\leq \sum_{j\neq i} P(a(x) = j|class(x) = i) = \sum_{j\neq i} err(j|i) \\ &\leq \sum_{j\neq i} \max_{j\neq i} err(j|i) = (K-1)\max_{j\neq i} err(j|i) \\ &\leq (K-1)\Phi\left(-\frac{1}{2}\min_{j\neq i}\delta(j,i)\right). \end{aligned}$$

The proof is finished by noting that $err = \sum_{i=1}^{K} \pi_i err(i)$ and $\pi_i = \frac{1}{K}$.□

References

[1] P. Diaconis, D. Freedman, Asymptotics of graphical projection pursuit, The Annals of Statistics 12 (1984) 793–815.

[2] Z. Jin, J.Y. Yang, Z.M. Tang, Z.S Hu, A theorem on the uncorrelated optimal discriminant vectors, Pattern Recognition 34(10)(2001) 2041–2047.

[3] G. J. McLachlan, Discriminant analysis and statistical pattern recognition, John Wiley & Sons, 1992.

[4] J. Yang, J.Y. Yang, D. Zhang, What's wrong with Fisher criterion? Pattern Recognition 35 (11) (2002) 2665–2668.