

Nonparametric Discriminant Analysis Applied to Medical Diagnosis

Mayer E. Aladjem

Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev,
P.O.B. 653, 84105 Beer-Sheva, ISRAEL, e-mail: aladjem@bgu.ac.il

Abstract

We present an application of our method for discriminant analysis [3] to the diagnosis of the neurological diseases haemorrhages and infarction due to ischaemia. The method searches for the discriminant directions which maximize the Patrick-Fisher (PF) distance between the projected class-conditional densities. It is a nonparametric method, in the sense that the densities are estimated from the data. Since the PF distance is a highly nonlinear function, we use a recursive optimization procedure for searching the directions corresponding to several large local maxima of the PF distance. The application to the medical dataset indicates the potential of our method for finding a sequence of directions with significant class separation.

1. Introduction

We discuss discriminant analysis of two classes which is carried out by the linear mapping $\tau = r^T x$, $x \in R^n$, $\tau \in R^1$, $n \geq 2$, with x an arbitrary n -dimensional observation, and r a direction vector (having unit length). The vector r maximizes the *Patrick-Fisher (PF) distance* [3,5], which unfortunately is not an unimodal function with respect to r and has more than one maximum. In most applications the optimal solution, called the *PF discriminant vector*, is searched for along the gradient of the PF distance, hoping that with a good starting point the procedure will converge to the global maximum or at least to a practical one. We use an *extended Fisher (ExF) discriminant analysis* [1,2] for choosing the starting point for the optimization procedure. Since the observed maximum of the PF distance can be merely a local maximum we apply a recursive procedure for searching several large local maxima of the PF distance.

In Section 2 we describe a normalization of the data which is required by the recursive procedure. Section 3 presents the PF distance and the ExF criterion and the computation of the discriminant vectors related to them. The recursive procedure for optimization of PF distance is presented in Section 4. Section 5 includes the results and analyses of the application to medical datasets, which is the original contribution of this work.

2. Normalized data

Suppose we are given a set of N_d design (training) observations $(z_1, l_1), (z_2, l_2), \dots, (z_{N_d}, l_{N_d})$ in n -dimensional sample space $z_j \in R^n$, ($n \geq 2$), $j=1, 2, \dots, N_d$.

We discuss the two-class problem and the label $l_j \in \{\omega_1, \omega_2\}$ shows that z_j belongs to one of the classes ω_1 or ω_2 . These labels imply a decomposition of the design set $Z_d = \{z_1, z_2, \dots, z_{N_d}\}$ into two subsets corresponding to the unique classes. Let the decomposition be $Z_d = Z_{d1} \cup Z_{d2}$, where the subset Z_{di} contains N_{di} observations properly associated with the class labeled by ω_i for $i=1, 2$. We perform an eigenvalue-eigenvector decomposition $S_z = RDR^T$ of the pooled sample covariance matrix S_z with R and D $n \times n$ matrices; R is orthonormal and D a diagonal. We then define the normalization matrix $A = D^{-1/2}R^T$. In the remainder of the paper, all operations are performed on the *normalized design samples* $X_{di} = \{x: x = A(z - m_z), z \in Z_{di}\}$ with m_z the mean vector of the design sample Z_d . The pooled sample covariance matrix estimated over the normalized sample $\{X_{d1} \cup X_{d2}\}$ becomes the identity matrix $AS_zA^T = I$.

3. Discriminant criteria

The *Patrick-Fisher (PF) distance* ([5], pp.277-280) is:

$$D_{PF}(r, h) = \left(\int_{-\infty}^{+\infty} \left[\frac{N_{d1}}{N_d} \hat{p}_r(\zeta | \omega_1) - \frac{N_{d2}}{N_d} \hat{p}_r(\zeta | \omega_2) \right]^2 d\zeta \right)^{1/2} \quad (1)$$

with $\hat{p}_r(\zeta | \omega_i)$ the Parzen estimators with Gaussian kernels of the class-conditional densities of the projections $\zeta = r^T x$ of the observations x of the normalized data. The PF discriminant vector maximizes $G_{PF}(r, h)$ for the fixed value of the smoothing parameter h (standard deviation of the Gaussian kernel). From experience, a suitable value is $h=0.1$. We carry out the optimization with respect to r by a sequential quadratic programming method available as a routine E04UCF in the NAG Mathematical Library. The primary goal is to find the global maximum of $G_{PF}(r, h)$. By a naive use of the optimization algorithm, the computed value for the observed $\max\{G_{PF}(r, h)\}$ can be merely a local maximum. The solution depends strongly on the starting point (vector) of the local optimizer. On the other hand, in some data structures more than one direction with significant (interesting) class separations exist. We use an extended Fisher discriminant vector as a starting point because of its adaptation to the data structure under variations of a control parameter (see the following text). In order to search for several large local maxima we apply a procedure for recursive optimization of $G_{PF}(r, h)$ (Section 4).

The *ExF discriminant vector* maximizes the extended Fisher discriminant criterion proposed by us [1,2]:

$$G(r, \beta) = [(1-\beta)r^T B r + \beta r^T S^{(-)} r] [r^T S_w r]^{-1} \quad (2)$$

with r direction vector, β ($0 \leq \beta \leq 1$) control parameter; $B = (m_1 - m_2)(m_1 - m_2)^T$ the sample between-class scatter matrix with m_i the class-conditional sample mean vectors; $S^{(-)} = S_1 - S_2$ or $S_2 - S_1$ with S_i the class-conditional sample covariance matrices for $i=1,2$; S_w the pooled within-class sample covariance matrix. All matrices are computed for the sphered design samples x_{di} , $i=1,2$. In (2) the symbol $S^{(-)}$ implies two forms of the criterion $G(r, \beta)$. The ExF discriminant vector r_β , which maximizes $G(r, \beta)$, is the eigenvector corresponding to the largest eigenvalue of the matrices $S_w^{-1}[(1-\beta)B + \beta(S_1 - S_2)]$ and $S_w^{-1}[(1-\beta)B + \beta(S_2 - S_1)]$. An appropriate value of the control parameter β is not known in advance. We choose the value of β that maximizes the PF distance along r_β . For this purpose we choose a grid of values in the interval ($0 \leq \beta \leq 1$), calculate the PF distance at each value and then choose the value with the largest PF distance as the value of β . From experience, a suitable size of optimization grid is 21 values (uniform grid with step 0.05).

4. Recursive optimization of the PF distance

In our previous work [3] we proposed a recursive procedure for optimization of the PF distance. The idea is to obtain a PF discriminant vector and then to transform the data along it into data with greater class overlap and to iterate for obtaining the next PF discriminant vector. A short description of the data transformation which increases the class overlap follows. Assume that U is an orthonormal $n \times n$ matrix with the PF discriminant vector r as the first row. Then applying the linear transformation $t = Ux$ results in a rotation such that the new first coordinate is $\tau_1 = r^T x$. We denote other coordinates as $\tau_2, \tau_3, \dots, \tau_n$ ($t = [\tau_1, \tau_2, \dots, \tau_n]^T$). Let $p_r(\tau_1 | \omega_i)$, $i=1,2$ be the class-conditional densities along r and $m_{r|\omega_i}$, $\sigma_{r|\omega_i}^2$ their means and variances. We require a transformation that takes the class-conditional densities along r to normal densities, but leaves all other coordinates $\tau_2, \tau_3, \dots, \tau_n$ unchanged. Let q be a vector function with components q_1, q_2, \dots, q_n that carries out this transformation: $\tau_1 = q_1(\tau_1)$ takes $p_r(\tau_1 | \omega_i) = N(m_{r|\omega_i} - \Delta m_i, \sigma_{r|\omega_i}^2 \pm \Delta \sigma^2)$ for $i=1,2$ with $\Delta \sigma^2$, Δm_1 , Δm_2 user-supplied parameters and $\tau_i = q_i(\tau_i)$, $i=2,3, \dots, n$ each given by identity transformation. The function q_1 is obtained by the percentile transformation method [3]. Finally, we define the transformed data $x' = U^T q(Ux)$.

We make trials while progressively increasing the values of $\Delta \sigma^2$ in the interval ($0 \leq \Delta \sigma^2 \leq 1$). In order to preserve the normalization of the data we compute Δm_1 and Δm_2 using the properties of the normalized data (zero unconditional mean and unconditional variance equal to one). Starting from $\Delta \sigma^2 = 0$ (and $\Delta m_i = 0$, $i=1,2$) we make

minimal changes of the data in the sense of the minimal relative entropy distance measure between the original and transformed class-conditional distributions [4]. Using $\sigma_{r|\omega_i}^2 \pm \Delta \sigma^2 = 1$ (and $m_{r|\omega_i} - \Delta m_i = 0$, $i=1,2$) we totally remove the classification structure along r ($p_r(\tau_1 | \omega_1) = p_r(\tau_1 | \omega_2)$). This causes (usually) a larger change of the class-conditional distributions of x' .

We have proposed (see [3]) the following computation procedure of the sequence of PF discriminant vectors:

Initialization: $x_1 = x_{d1}$, $x_2 = x_{d2}$.

Step 1: Reduction of class separation

1.1. Using the sample set $\{x_1 \cup x_2\}$, compute the ExF vector with the largest PF distance (Section 3).

1.2. Starting from the ExF vector, search for a convergence point by using a local maximizer (NAG routine E04UCF) of PF distance. The direction vector after convergence of the maximizer is a current PF vector. Save it.

1.3. Reduce the class separation along the PF vector and obtain a new set $\{x_1' \cup x_2'\}$. Assign the new set to be the current sample set, i.e. $x_1 = x_1'$, $x_2 = x_2'$.

1.4. Repeat above steps 1.1 - 1.3.

Step 2: Adjust (reoptimize) the PF vectors

Starting from PF vectors obtained in Step 1.2, search for the convergence points of the local optimizer of the PF distance into the original normalized data x_{d1} , x_{d2} . The direction vectors after convergence of the algorithm are the *adjusted PF vectors*. Save them. We stop the iterations if several PF vectors, with different class separations along them, are obtained.

In step 1.3., we carry out trials with progressive increasing values of $\Delta \sigma^2$. We examine the class-conditional distributions of the projected samples before and after reduction of class separation along the PF vector and we choose suitable values of $\Delta \sigma^2$ subjectively.

5. Application to Medical Data

The data concerns the medical diagnosis of the neurological disease cerebrovascular accident (CVA). It contains pathologo-anatomically verified CVA cases including a first class of 200 cases with haemorrhages and a second class of 200 cases with infarction due to ischaemia. Twenty numerical results from the neurological examination were recorded for each CVA case [1]. In order to eliminate the small pooled variances we used eight largest eigenvalues in the decomposition of S_z (see Section 2). Subsequently we reduced the dimensionality of the normalized data to eight.

5.1. Recursive Optimization

Following the recursive optimization procedure (Section 4), we computed the ExF discriminant vector for $\beta=0$, which implied PF distance 0.5046. Starting from it we ran the local optimizer of the PF criterion, which converged to a PF discriminant vector with PF distance of value 0.8013. The class-conditional densities along the obtained discriminant vectors are shown in Fig.1. The

class separation was increased significantly along the PF vector (Fig.1b.). Actually, this was the best result obtained in our study.

In order to monitor the progress of our procedure we specified the coordinates of the normalized data which implied the class discrimination along the PF vector. They are the 8th, 3th, 4th, 2th and 7th coordinates corresponding to the components of the PF vector with dominant values (see Fig.1b). We decided to monitor the results using plots spanned by the 3th and 4th, and 7th and 8th coordinate-axes. Fig.2 presents the projection of the samples on to these plots.

We iterated by a sequence of reductions of class separation in order to search for other directions with discriminant information. Following the procedure of Section 4 we started trials with small values of $\Delta\sigma^2$. We iterated with $\Delta\sigma^2=0.0, 0.0, 0.1, 0.2$ in the 1st to the 4th trials. The adjusted PF vectors for this sequence (see Step 2 of the procedure) converged to the result of the first trial (Fig.1b) and we detected no new local maximum of the PF distance. We decided to continue with stronger reductions of class separation (larger values of $\Delta\sigma^2$). We iterated with $\Delta\sigma^2=0.3, 0.4, 0.5, 0.5$ and observed the adjusted PF distances 0.4475, 0.3498, 0.4129, 0.5069 in the 5th to the 8th trials. Fig.3 shows the transformed data after the 5th reduction of class separation. The presented destructuring of the data (with respect to the original data Fig.2) directed the optimization procedure to a maximum different from the initial solution.

The 8th reduction of class separation implied an "interesting" result. We analysed the discriminant information gained by the PF vector at this trial with respect to the best result shown in Fig.1b. For this purpose we projected the original data on to the plot spanned by the latter vectors. The projections of the samples on to this plot are shown in Fig.4. We detected a cluster (the encircled area in Fig.4) of large (nearly full) overlap of the classes. The other clusters define areas with a dominant number of observations from one of the classes. We found that the two-dimensional presentation (Fig.4) gains less class overlap compared with the one-dimensional projection (Fig.1b). We concluded that the PF vector at the 8th reduction of class separation adds "new" discriminant information to the best solution and consequently the obtained two-dimensional mapping may be a suitable choice for allocation of the CVA cases. We would mention that an assessment of the probability of misclassification and rejection by "extra" (test) samples (holdout and bootstrap methods) is outside the scope of this paper.

5.2. Optimization with no reduction of class separation

The goal of this experiment was to examine the usefulness of the reduction of class separation for detecting the sequence of large local maxima of PF distance. For this purpose we maximized the PF distance of the original

data (with no reductions of class separation) starting the local optimizer from the ExF vectors used in the trials of the previous experiment (ExF vectors for $\beta=0.0, 0.5, 0.7, 1.0$). After convergence of the local search we observed the PF distances 0.8013 for $\beta=0.0, 0.5, 0.7$ and 0.4436 for $\beta=1.0$. The values $\beta=0.0, 0.5, 0.7$ implied the best solution obtained in Section 5.1 (Fig.1b). The value of $\beta=1.0$ implied the local maximum of value 0.4436, which was less than the values of 0.4475 and 0.5069 found by our procedure. We concluded that the reduction of class separation was useful, because it directed the local search to values of the PF distance which were larger than the values found for the original data.

6. Conclusion

We have presented an application of our procedure for recursive optimization [3]. It succeeded in finding the sequence of directions with significant information for discrimination of the CVA cases. Our procedure was more effective than the initialization by the Fisher discriminant vectors which was used in the past.

Our method implements Friedman's [4] procedure for recursive optimization, called "structure removal". Just like Friedman's procedure we transform the densities along the found discriminant directions into normal densities. The main difference of our proposal from Friedman's procedure consists in the choice of the density which is transformed. Friedman's algorithm transforms the mixture (unconditional) density to normal density while our algorithm processes the class-conditional densities separately. The latter arises from the goal of our method. It is a tool for discriminant analysis (analysis of samples previously grouped into classes) while Friedman's procedure is oriented to cluster analysis of unclassified samples.

Acknowledgments

This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben Gurion University of the Negev, Israel.

References

- [1] M.E.Aladjem, "PNM: A program for parametric and nonparametric mapping of multidimensional data", *Computers in Biology and Medicine*, vol. 21, pp. 321-343, 1991.
- [2] M.E.Aladjem, "Multiclass discriminant mappings", *Signal Processing*, vol.35, pp.1-18, 1994.
- [3] M.E.Aladjem, "Two-class pattern discrimination via recursive optimization of Patrick-Fisher distance", *Proc. of the 13th International Conference on Pattern Recognition*, vol.2, pp.60-64, 1996.
- [4] J.H.Friedman, "Exploratory projection pursuit", *Journal of the American Statistical Association*, vol. 82, pp. 249-266, 1987.
- [5] P.A.Devijver and J.Kittler, *Pattern Recognition: A Statistical Approach*, Prentice-Hall International, Inc., London, 1982.

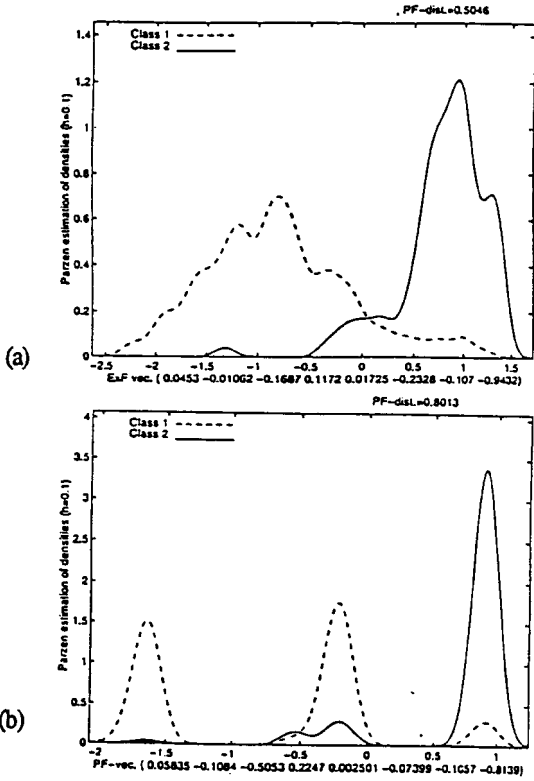


Fig.1. Maximal class separation with no reduction of class separation: (a) along ExF vector ($\beta=0.0$), (b) along PF vector.

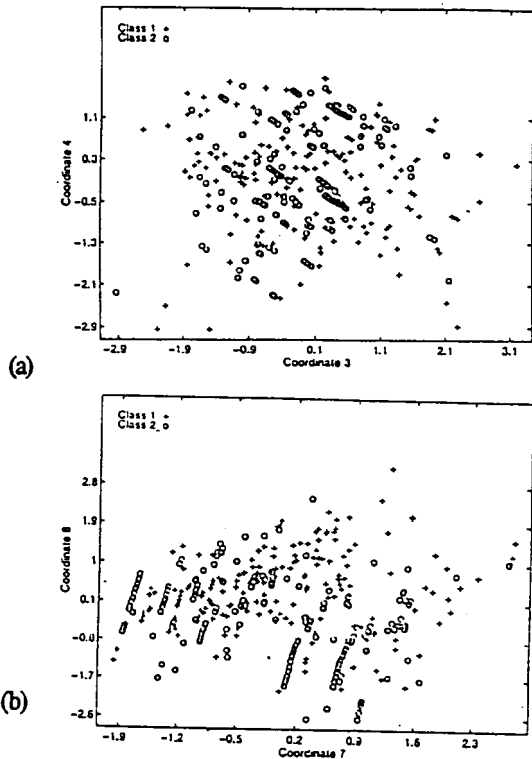


Fig.3. Transformed CVA data, after 5th reduction of class separation: (a) along 3th and 4th coordinate axes, (b) along 7th and 8th coordinate axes.

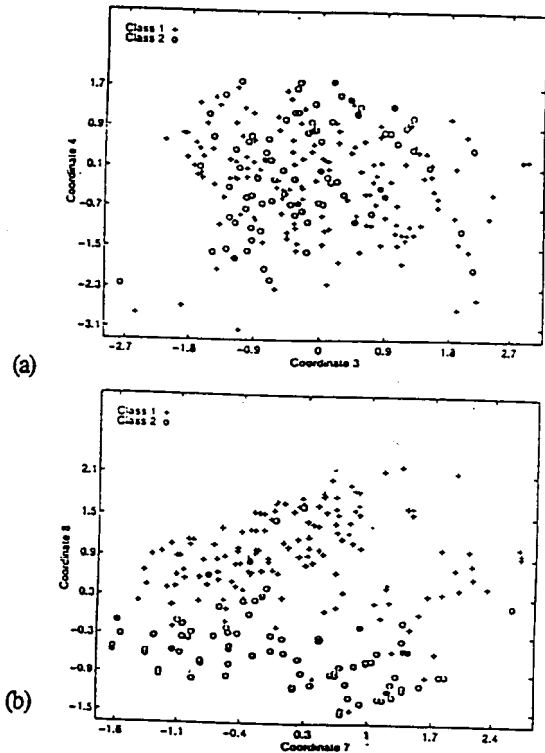


Fig.2. Normalized CVA data : (a) along 3th and 4th coordinate axes, (b) along 7th and 8th coordinate axes.

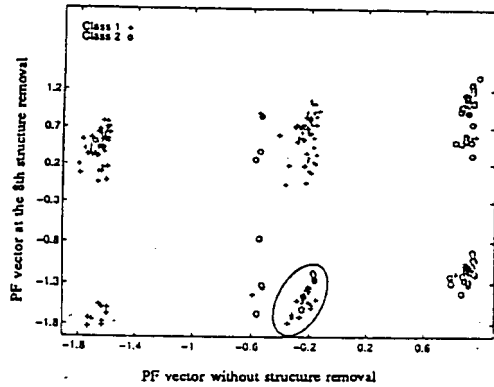


Fig.4. Projection of the CVA samples on to the plot spanned by the discriminant vectors with largest PF distances.