ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Probabilistic Fisher Discriminant Analysis

Charles Bouveyron[1] and Camille Brunet[2]

1- University Paris 1 Panthéon-Sorbonne – Laboratoire SAMM, EA 4543
90 rue de Tolbiac – 75013 PARIS - FRANCE

2- University of Evry - IBISC TADIB, FRE CNRS 3190
40 rue Pelvoux CE 1455 – 91020 EVRY - FRANCE

Abstract. Fisher Discriminant Analysis (FDA) is a powerful and popular method for dimensionality reduction and classification which has unfortunately poor performances in the cases of label noise and sparse labeled data. To overcome these limitations, we propose a probabilistic framework for FDA and extend it to the semi-supervised case. Experiments on real-world datasets show that the proposed approach works as well as FDA in standard situations and outperforms it in the label noise and sparse label cases.

1 Introduction

Fisher Discriminant Analysis (FDA) [1, 2], also known as LDA by misnomer, is a commonly used method for linear dimensionality reduction in supervised classification. FDA aims to find a linear subspace that well separates the classes in which a linear classifier (usually LDA) can be learned. In this paper, FDA will therefore refer to the strategy which first finds a discriminative subspace and then classify the data in this subspace using LDA. FDA is a popular method which works very well in several cases. However, FDA does have some very well-known limitations. In particular, FDA produces correlated axes and its prediction performances are very sensitive to outliers, unbalanced classes and label noise. Moreover, FDA has not been defined in a probabilistic framework and its theoretical justification can be obtained only under the homoscedastic assumption on the distribution of the classes.

Many authors have proposed different ways to deal with these problems and a first probabilistic framework has been proposed by Hastie *et al.* [3] by considering the different classes as a mixture of Gaussians with common covariance matrices. In 1998, Kumar *et al.* [4] have rewritten the Fisher's problem through a probabilistic framework constrained on the vector mean and the covariance matrix of the latent space. More recently, Ioffe [5] has proposed a probabilistic approach for LDA and in the same year, Yu *et al.* [6] have adapted the framework of probabilistic principal component analysis (PPCA) developed by Tipping *et al.* [7] in a supervised context and have found that the maximum likelihood of their approach is equivalent to the one of FDA in the homoscedastic context. Besides, Zhang *et al.* [8] have presented an extension of the Yu's work by considering the heteroscedastic case in a supervised and semi-supervised context which implies that the linear transformation is different for each class.

In this paper, we propose a supervised classification method, called Probabilistic Fisher Discriminant Analysis, based on a Gaussian parametrization of the

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

data in a latent orthonormal discriminative subspace with a low intrinsic dimension. This probabilistic framework enables to relax the homoscedastic assumption on the class covariance matrices and allows its use in the semi-supervised context. Numerical experiments show that PFDA improves predictive effectiveness in the cases of label noise and semi-supervised contexts.

2 Probabilistic Fisher Discriminant Analysis

2.1 The probabilistic model

Let us consider a complete training dataset $\{(y_1, z_1), ..., (y_n, z_n)\}$ where $y_1, ..., y_n$ are independent realizations of an observed random vector $Y \in \mathbb{R}^p$ and $z_i \in \{1, \ldots, K\}$ indicates the class label of y_i. Let us first assume that there exists a latent subspace $\mathbb{E}$ of dimension $d < p$ such that $\mathbf{0} \in \mathbb{E}$ and that $\mathbb{E}$ best discriminates the classes. Moreover, $\{(x_1, \ldots, x_n)\} \in \mathbb{E}$ denote the actual data in the latent space $\mathbb{E}$ which are presumed to be independent unobserved realizations of a random vector $X \in \mathbb{E}$. Finally, let us assume that Y and X are linked through a linear relationship of the form:

$$Y = UX + \varepsilon. \tag{1}$$

where U is a $p \times d$ orthonormal matrix, $X \in \mathbb{E}$ is the latent random vector and ε is a noise term. We further assume that:

$$X|_{Z=k} \sim \mathcal{N}(\mu_k, \Sigma_k) \quad \text{and} \quad \varepsilon|_{Z=k} \sim \mathcal{N}(\mathbf{0}, \Psi_k). \tag{2}$$

where $\mu_k \in \mathbb{E}$ and $\Sigma_k \in \mathbb{R}^{d \times d}$ are respectively the mean and the covariance matrix of the kth class in the latent space. In conjunction with equation (1), these assumptions imply that the marginal probability distribution of Y is $Y|_{Z=k} \sim \mathcal{N}(m_k, S_k)$ where $m_k = U\mu_k$ and $S_k = U\Sigma_k U^t + \Psi_k$ are respectively the mean and covariance matrix of class k in the observation space. We also define $\pi_k = P(Z = k)$ as the prior probability of class k. Let us introduce $W = [U, V]$ a $p \times p$ matrix which satisfies $W^t W = WW^t = \mathbf{I}_p$ and for which the $p \times (p - d)$ matrix V is the orthonormal complement of U defined above. Finally, we assume that the noise covariance matrix Ψ_k satisfies the conditions $V\Psi_k V^t = \beta_k \mathbf{I}_{p-d}$ and $U\Psi_k U^t = \mathbf{0}_d$, such that $\Delta_k = W^t S_k W = \mathrm{diag}(\Sigma_k, \beta_k \mathbf{I}_{p-d})$. Then, given these assumptions, the log-likelihood of the dataset is:

$$\begin{aligned} \mathcal{L}(\theta) = -\frac{1}{2} \sum_{k=1}^{K} \Big[& -2\log(\pi_k) + \mathrm{trace}(\Sigma_k^{-1} U^t C_k U) + \log(|\Sigma_k|) \\ & + (p-d)\log(\beta_k) + \frac{1}{\beta_k}\Big(\mathrm{trace}(C_k) - \sum_{j=1}^{d} u_j^t C_k u_j\Big) + \gamma \Big]. \end{aligned} \tag{3}$$

where C_k is the empirical covariance matrix of the k^{th} class, u_j is the jth column vector of U and $\gamma = p\log(2\pi)$ is a constant term.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

This model will be referred to by $[\Sigma_k\beta_k]$ in the sequel. In order to obtain more parsimonious models, Σ_k or β_k can be constrained between and within the classes and it is possible to decline 12 different models. A detailed description of those models in the unsupervised context can be found in [9]. Besides, in this paper, 4 models will be considered: the general model $[\Sigma_k\beta_k]$, the $[\Sigma_k\beta]$ model which assumes an isotropic variance in the residual space ($\forall k, \beta_k = \beta$), the $[\alpha_{kj}\beta_k]$ model which supposes a diagonal covariance matrix in the latent space in each class ($\Sigma_k = \text{diag}(\alpha_{k1}, \ldots, \alpha_{kd})$) and finally the $[\alpha_{kj}\beta]$ model which assumes both a diagonal covariance matrix in the latent space and a common variance in the residual space.

2.2 Parameter estimation

Conversely to the probabilistic approaches reviewed in Section 1, the probabilistic model presented above is very general and there is no explicit solution for the likelihood maximization with respect to U. Therefore, we propose to estimate the linear transformation U and the model parameters in two different steps. Firstly, the estimate $\hat{U}$ of the latent subspace orientation U is obtained through the optimization of the Fisher criterion with respect to the orthogonality of its column vectors,

$$\max_{U} \quad \text{tr}\left((U^tSU)^{-1}U^tS_BU\right) \quad \text{wrt} \quad U^tU = \mathbf{I_d}, \tag{4}$$

where $S_B = \frac{1}{n}\sum_{k=1}^{K} n_k(m_k - \bar{y})(m_k - \bar{y})^t$ and $S = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^t$ are respectively the between and the total covariance matrices with $m_k = \frac{1}{n_k}\sum_{i=1}^{n}\mathbf{1}_{\{z_i=k\}}y_i$, $n_k = \sum_{i=1}^{n}\mathbf{1}_{\{z_i=k\}}$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Secondly, given $U = \hat{U}$ and in conjunction with equation (4), the maximization of the log-likelihood (3) conduces to the following estimates of the model parameters:

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{n}\mathbf{1}_{\{z_i=k\}}\hat{U}^ty_i, \quad \hat{\Sigma}_k = \hat{U}^tC_k\hat{U}, \quad \hat{\beta}_k = \frac{\text{tr}(C_k) - \sum_{j=1}^{d}\hat{u}_j^tC_k\hat{u}_j}{p-d}. \tag{5}$$

Finally, the intrinsic dimension d is set to the rank of $S_B \leq K-1$ (see [2]).

2.3 Classification of new observations

In the discriminant analysis framework, new observations are usually assigned to a class using the *maximum a posteriori* rule which assigns a new observation $y \in \mathbb{R}^p$ to the class for which y has the highest posterior probability $P(Z = k|Y = y)$. Maximizing the posterior probability is equivalent to minimizing the cost function $\Gamma_k(y) = -2\log(\pi_k\phi(y; m_k, S_k)$ which is for our model equal to:

$$\begin{aligned}\Gamma_k(y_i) = \quad & \left\|UU^t(y_i - m_k)\right\|_{\vartheta_k}^2 + \frac{1}{\beta_k}\left\|(y_i - m_k) - UU^t(y_i - m_k)\right\|^2 \\ & + \log(|\Sigma_k|) + (p-d)\ \log(\beta_k) - 2\log(\pi_k) + p\ \log(2\pi),\end{aligned} \tag{6}$$

where $\vartheta_k = [U, \mathbf{0}_{p-d}]\,\Delta_k^{-1}\,[U, \mathbf{0}_{p-d}]^t$ and $\|.\|_{\vartheta_k}$ is a norm on the latent space spanned by $[U, \mathbf{0}_{p-d}]$ such that $||y||_{\vartheta_k}^2 = y^t\vartheta_k y$.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

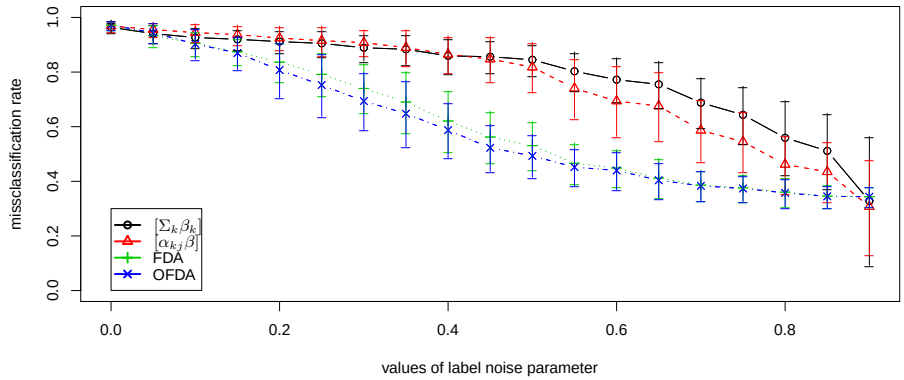

Fig. 1: Effect of label noise in the learning dataset on the prediction effectiveness.

1

2.4 Extension to the semi-supervised context

Let us consider now that $\{(y_i, z_i)\}_{i=1}^{n_\ell}$ where $n_\ell \leq n$ are the labeled data and there are $n - n_\ell$ unlabeled data referred to by $\{y_i\}_{i=n_\ell+1}^{n}$. The n_ℓ labeled observations are modeled by the probabilistic framework developed in Section 2.1 and the unlabeled data are modeled by a mixture model parametrized by π_k, the mixture proportion of the class k, and $\theta_k = (m_k, S_k)$, respectively its mean vector and its covariance matrix. Thus, the log-likelihood can be written as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n_\ell}\sum_{k=1}^{K} \mathbf{1}_{\{z_i=k\}} \log\left(\pi_k \phi(y_i;\theta_k)\right) + \sum_{i=n_\ell+1}^{n} \log(\sum_{k=1}^{K} \pi_k \phi(y_i;\theta_k)) \qquad (7)$$

In such a case, the direct maximization of $\mathcal{L}(\theta)$ is intractable and an iterative procedure have to be used. Therefore, we use the Fisher-EM algorithm proposed by [9] which alternates 3 steps: an E-step which computes the posterior probabilities, a F-step which estimates the linear transformation and a M-step which estimates the parameters of the mixture model.

3 Experiments

3.1 Robustness to label noise

This first experiment aims to study the robustness to label noise of PFDA and to compare it with traditional methods such as FDA and orthonormalized FDA (OFDA) [10]. Two different models of the PFDA approach are here considered, the $[\Sigma_k\beta_k]$ and $[\alpha_{kj}\beta]$ models and their robustness to label noise is compared to FDA and OFDA. For this experimentation, a wink to the work of Sir R. A. Fisher is given since we apply here PFDA to the iris dataset. This dataset consists of 3 classes of 50 observations corresponding to different species of iris (setosa, versicolor and virginica) which are described by 4 features relative to the

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

model		chironomus	wine	iris	usps358
PFDA	$[\Sigma_k \beta_k]$	96.1 ± 2.8	96.9 ± 1.4	92.2 ± 2.8	61.9 ± 4.3
	$[\Sigma_k \beta]$	97.0 ± 2.6	96.1 ± 1.6	$\mathbf{96.9 \pm 2.5}$	90.7 ± 0.5
	$[\alpha_{kj} \beta_k]$	$\mathbf{97.1 \pm 2.1}$	$\mathbf{97.3 \pm 1.5}$	92.7 ± 4.1	66.2 ± 5.4
	$[\alpha_{kj} \beta]$	96.8 ± 3.8	96.3 ± 1.8	96.6 ± 1.8	91.5 ± 1.0
SELF		92.5 ± 2.5	93.9 ± 3.0	93.1 ± 6.9	$\mathbf{92.3 \pm 0.8}$
FDA		87.7 ± 7.2	93.4 ± 2.9	95.8 ± 2.9	44.1 ± 3.4
OFDA		85.6 ± 8.3	90.9 ± 4.5	96.2 ± 2.5	40.8 ± 4.5

Table 1: Prediction accuracies and their standard deviations (in percentage) on the UCI datasets averaged on 25 trials.

length and the width of the sepal and the petal. Since the aim of this experiment is to evaluate the robustness to label noise, let us define τ the percentage of false labels in the learning set which varies between 0 and 0.9. At each trial, the iris dataset is randomly divided in 2 balanced samples: a learning set in which a percentage τ of the data is mislabeled and a test set on which the prediction performances of the 4 methods are evaluated. This process has been repeated 50 times for each value of τ in order to monitor both the average performances and their variances. Figure 1 presents the evolution of correct classification rate computed on the test set for the 4 methods according to τ. First of all, it can be observed that the FDA and OFDA methods are comparable since their curves are almost superimposed and their classification rates lower drastically and linearly until $\tau = 0.7$ where their prediction performances are comparable to those of a random classifier. Conversely, the $[\Sigma_k \beta_k]$ and $[\alpha_{kj} \beta]$ models of PFDA appear robust to label noise since their correct classification rates remain superior to 0.8 for a label noise up to $\tau = 0.6$. These improvements can be explained by the PFDA framework which takes into account an error term and this avoids to overfit the embedding space on the labeled data and remains generally enough to be robust on label noise contrary to FDA and OFDA.

3.2 Semi-supervised context

This second experiment will focus on comparing on 4 real-world datasets the efficiency of semi-supervised approaches with traditional supervised ones. Four different models of PFDA ($[\Sigma_k \beta_k]$, $[\Sigma_k \beta]$, $[\alpha_{kj} \beta_k]$ and $[\alpha_{kj} \beta]$) are compared with a recent semi-supervised local approach of FDA proposed by [11]. This approach, called SELF, aims to find a discriminative subspace by considering both global and class structures. Besides, the experiment includes the 2 supervised approaches previously seen in Section 3.1 (FDA and OFDA). The comparison has been made on 4 different benchmark datasets coming from the UCI machine repository: the chironomus, the wine, the iris and the usps358 datasets. They are all made of 3 classes, but the first three datasets contain respectively 148, 178 and 150 observations whereas the usps358 data contain 1756 individuals. Finally, the respective dimension of each dataset is as following: the irises are

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

described by 4 variables, the wines by 13, the chironomus by 17 and the usps358 by 256 variables. Moreover, each dataset has been randomly divided 25 times in 2 samples composed by a learning set and a test set containing 50% of the data each. Moreover, in the learning set, 30% of data are randomly selected to constitute the known labeled data. For the experiment, the algorithms are initialized on parameters estimated on the known labeled data and then, the modeling of the data has been made on the learning set of labeled and unlabeled data. Table 1 presents the average correct classification rate and the associated standard deviations obtained for the 5 studied methods. First of all, one can notice that the semi-supervised methods always improve the prediction accuracy and outperform FDA and OFDA. This can be explained by the fact that the supervised methods estimate the embedding space only on the labeled data and thus overfit it. Conversely, the semi-supervised methods use also unlabeled data in their estimation of the discriminative subspace which enables them to be often more effective. Furthermore, most of PFDA models present the best correct classification rates except for the usps358 dataset where the SELF algorithm reaches 92.3% of classification accuracy.

4 Conclusion

This paper has presented a probabilistic framework for FDA and has shown that the proposed PFDA method works as well as the traditional FDA method in standard situations and it improves clearly the modeling and the prediction when the dataset is subject to label noise or sparse labels. The practitioner may therefore replace FDA by PFDA for its daily use.

References

[1] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[2] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[3] T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixture. *Journal of the Royal Statistical Society, Series B*, 58(1):155–176, 1996.

[4] N. Kumar and A.G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication*, 26(4):283–297, 1998.

[5] S. Ioffe. Probabilistic linear discriminant analysis. *Computer Vision ECCV*, 2006.

[6] S. Yu, K. Yu, V. Tresp, H.P. Kriegel, and M. Wu. Supervised probabilistic principal component analysis. In *Proc. of the 12th ACM SIGKDD*, pages 464–473, USA, 2006.

[7] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.

[8] Yu Zhang and DY Yeung. Heteroscedastic Probabilistic Linear Discriminant Analysis with Semi-supervised Extension. *Lecture Notes in Computer Science*, 5782:602–616, 2009.

[9] C. Bouveyron and C. Brunet. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Preprint HAL no 00492406*, 2010.

[10] Y. Hamamoto, Y. Matsuura, T. Kanaoka, and S. Tomita. A note on the orthonormal discriminant vector method for feature extraction. *Pattern Recognition*, 24(7), 1991.

[11] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese. Semi-supervised local Fisher discriminant analysis for dimensionality reduction. *Machine Learning*, 78:35–61, 2009.