Optimal Bayes Classification of High Dimensional Data in Face Recognition

Wissal Drira and Faouzi Ghorbel
GRIFT Research Group, CRISTAL Laboratory, National School of Computer Sciences, University of Manouba, Manouba, Tunisia

Keywords: Face Classification, Bayes, Feature Extraction, Reduction Dimension, L^2 Probabilistic Dependence Measure.

Abstract: In the supervised context, we intend to introduce a system which is composed of a series of novel and efficient algorithms that is able to realize a non parametric Bayesian classifier for high dimension. The proposed system tries to search for the best discriminate sub space in the mean of the minimum of the probability error of classification which is computed by using a modified kernel estimate of the conditional probability density functions. Therefore, Bayesian classification rule is applied in the reduced sub space. Such heuristic consists of four tasks. First, we maximize a novel estimate of the quadratic measure of the probabilistic dependence in order to realize multivariate extractors resulting from a number of different initializations of a given numerical optimizing procedure. Second, an estimation of the miss classification error is computed for each solution by the kernel estimate of the conditional probability density functions with the optimal band-with parameter in the sense of the Mean Integrate Square Error (MISE) which is obtained with the Plug in algorithm. Third, the sub space which presents the minimum of the miss classification values is thus chosen. After that, the Bayesian classification rule is operated in the reduced sub space with the optimal MISE of the modified kernel estimate. Finally, different algorithms will be applied to a base of images in grayscale representing classes of faces, showing its interest in the case of real data.

1 INTRODUCTION

One of the main goals of the discriminate analysis is to prepare the classification procedure in a relatively reduced space which is defined by optimizing a given criterion. Different classification rules have been applied in literature. The k means and the Bayesian are considered as ones of the well known classifiers. The first is very useful in practice since it could be working well in high dimension and does not need any assumptions about the form of the conditional distributions of the observation. It presents also a low algorithmic complexity and implementation facilities. Unfortunately, when one of the conditional distributions has a multimodal form (or we are in the Heterosedastic condition), the k mean does not lead to a satisfactory solution. In order to minimize the probability of error, the optimal solution could be reached by the application of the Bayesian classifier when it is possible to do. Despite its simplicity of implementation and low requirement at the sample size training, in the case of high dimensions, the bayes classifier cannot lead to precise results due to the fact that very few analytical expressions of random vectors are known outside of classical cases such as Gaussian vectors, uniform, Gamma, Beta ... These are close to covering the most frequent cases encountered in various applications. The nonparametric approach seems consequently more realistic to model in the concrete situations. However, it is well established in the statistical literature that the convergence of nonparametric probability density function (pdf) estimators in the sense of the conventional criteria, generally requires a sample size that increases exponentially with the dimension of the features space. Kernel estimate and the based orthogonal functions are non parametric and have strong requirements in terms of sample sizes with the same order to those required by the method of the histogram. To unblock this limitation, the approach known by linear discriminate analysis (LDA) is preferred. The LDA method has a certain number of advantages in the practice. We mention here its

Drira W. and Ghorbel F..
Optimal Bayes Classification of High Dimensional Data in Face Recognition.
DOI: 10.5220/0004345106410646
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods* (BTSA-2013), pages 641-646
ISBN: 978-989-8565-41-9

simplicity of implementation and its low algorithmic complexity. However, relatively recent works have identified major limitations coming from at least two reasons: in the first the instability of the results that may come from the numerical inversion method of the within class matrix in the case of high dimension feature observation vector D, the second is the convergence towards an unsatisfactory result of reduction dimension that can be interpreted by the fact that the scatter matrices are expressed only from the statistical moments of order less than or equal to 2 of the conditional distributions.

In order to overcome these limitations, the probabilistic distances were introduced in the context of two classes. Patrick Fischer distance was estimated between the two conditional probability density functions. Its estimations based on kernel method have been proposed for the implementation of this discriminate analysis qualified as probabilistic. The L^2-probabilistic measure of dependence was discussed in the literature (Devijver and Kittler, 1982) with the objective of generalizing the notion of Patrick Fischer distance to the multiclass case. In our recent work (Drira and Ghorbel, 2012), an estimator of the L^2 probabilistic dependence measure based on orthogonal functions has been introduced leading to a global multivariate extractor. In the implementation of the classification in the reduced space, a second optimization phase operated on different results of the numerical maximization of the estimate of the probabilistic dependence measure is representing an attempt to achieve a minimization of probabilities error and consequently a trial of the approximation of the Bayes classification.

So this article is structured as following. A first section will include a brief reminder of the orthogonal probabilistic dependence measure estimate. A second section will be discussing the optimal bayes classifier method based on the orthogonal estimate of the probabilistic distance for reduction dimension presented in previous work (Drira and Ghorbel, 2012). In the third section, we will present a face classification process based on these different algorithms of the orthogonal estimate of the dependence probabilistic measure.

2 FORMULATION

Linear discriminate analysis LDA (Fisher, 1936) is probably the most well-known approach to supervised linear dimension reduction (LDR). The LDA is a very robust and effective technique but still has some limitations. Various techniques were introduced to improve it, among them we can find discriminate analysis based on Chernoff criterion like the Approximate Chernoff Criterion ACC (Loog et al., 2001) which takes the Chernoff distance in the original space into consideration to minimize the error rate in the transformed space, another method based on the information theory called Information Discriminate Analysis IDA (Nenadic, 2007) based on a numerical optimization of an information-theoretic objective function.

These methods are defined from first and second order statistical moments of the conditional class random variable or based on some hypothesis on the law distribution type such as the Heterosedastic condition. Therefore, in complex situations as the multimodal conditional distributions, these methods couldn't describe completely statistical dispersions.

In order to avoid this limitation, distances between the conditional probability density functions weighted by the prior probabilities have been suggested by Patrick and Fischer. They have been introduced in (Patrick and Fischer, 1969) a global discriminate analysis based on the L^2 dependence probabilistic measure. It issues directly the discriminate plane which optimizes the orthogonal estimate of L^2 measure of dependence. The multivariate estimator in the reduced space of this L^2 probabilistic dependence measure represented by the linear transformation W of R^D, introduced in previous work (Drira and Ghorbel, 2012), expressed in function of W as follows:

$$\hat{I}_P(W) = \frac{1}{(\sum_{k=1}^{K} n_k)^2} \sum_{k=1}^{K} \left(\sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \left[\widetilde{K}_{m_{nk}}(WV_i^k, WV_j^k) \right] + \sum_{p=1}^{K} \sum_{i=1}^{n_p} \frac{1}{n_p} \left[\sum_{l=1}^{K} \sum_{j=1}^{n_l} \frac{1}{n_l} \left[\widetilde{K}_{\min(m_{np},m_{nl})}(WV_i^p, WV_j^l) \right] \right] - 2 \sum_{i=1}^{n_k} \sum_{l=1}^{K} \sum_{j=1}^{n_l} \frac{1}{n_l} \left[\widetilde{K}_{\min(m_{nk},m_{nl})}(WV_i^k, WV_j^l) \right] \right)$$

Where $\{V_i^k,\ i = 1, \dots, n_k,\ k = 1, .., K\}$ denotes a supervised learning sample distributed according the random vector conditional of the class k of dimension D, n_k is the size of the observation relative to the class k and K denotes the number of class. $\langle W|V \rangle$ represents the inner product of two vectors V and W of the space R^D.

$\widetilde{K}_{m_N}(v, V_j)$ is the multivariate generalized kernel associated to an orthogonal functions basis:

$$\widetilde{K}_{m_N}(v, V_i) = \widetilde{K}_{m_N}\left((x^1, \dots, x^d), (X_i^1, \dots, X_i^d)\right) = \prod_{l=1}^{d} K_{m_N}(x^l, X_i^l)$$

And $K_{m_N}(x, X_j)$is the scalar generalized kernel associated to an orthogonal functions basis $e_l(x)$. For a trigonometric system defined on [-π, π] corresponds to a complete basis in the Hilbert space $L^2([-\pi, \pi])$, the estimation of the kernel K_{m_N} gives the Dirichlet kernel given by the following expression:

$$\widehat{K}_{m_N}(x, y) = \frac{\sin\left[\left(\frac{2m_N+1}{2}\right)(x-y)\right]}{2\pi \sin\left[\frac{x-y}{2}\right]} \;;\; x \in [-\pi, \pi]$$

So, the discriminate analysis is obtained numerically by maximizing the estimator of the L^2 probabilistic dependence measure relatively to W, defined by:

$$W^* = Arg \max_{W \in \mathbb{R}^{D\,x\,d}} \widehat{I_2}(W)$$

Where W is an orthonormal matrix i.e. $W\,W^T = I_D$ where I_D is the identity matrix of size D.

This functional maximization problem cannot be solved analytically. Numerical optimization methods are then implemented. The number of parameters involved in this optimizing problem is d x D.

3 BAYES OPTIMAL OF FEATURE EXTRACTION BASED ON ORTHOGONAL SERIES ESTIMATE

There exists a sort of equivalence between the distance of Patrick Fischer and the probability of misclassification. This equivalence is expressed by the possibility to surround the probability of error with this distance close to constants. Thus, from a theoretical point of view, the d-dimension reduction by the orthogonal estimate of the Patrick-Fischer distance presented in (Drira and Ghorbel, 2012) can be interpreted as a dimension reduction equivalent to that which could be obtained by minimizing the probability of misclassification. The corresponding discriminate analysis prepares the implementation of the Bayes classifier in a non-parametric frame and large dimensions. Since Bayesian classifier role can be expressed according to the conditionals probability densities as:

$$g^*(x) = Arg \max_k \pi_k f_k(x)$$

The previous expression has to be estimated with a non parametric method as the histogram, the orthogonal basis one or the kernel estimate. For a given precision, the convergence of the corresponding theorems needs a large size N of the supervised training sample$\{(X_i, Y_i)\}_{i\in[1..N]}$ which have to increase exponentially with the dimension of the feature space. Thus the application of the Bayesian rule can be obtained from the modified kernel estimate which is defined as:

$$\widehat{f_k}(x) = \frac{1}{N h_N^d} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h_N}\right) I_{[Y_i=k]}(x)$$

Where *K* represents a probability density function, I notices the indicator function of the set $[Y_i = k]$ and h_N is the smoothing parameter. The adjustment of the kernel method described in (Drira and Ghorbel, 2012) which is applied to its modified version, contributes to achieve the goal through its application in the reduced space supposing that its dimension is below the value 3. This tends to promote conditions of respect of convergence theorems of the kernel estimator. This convergence could be obtained in case of real data whose cost of collecting supervised observations still affordable.

On the other hand, the search for an absolute maximum of the objective function of the estimate L^2 probabilistic dependence measure cannot be obtained analytically, a numerical optimization process is therefore required. Solutions obtained from numerical optimization are often dependent on initialization and lead in most cases to local maxima. By multiplying different initializations of the numerical optimization process, we obtain a family of M d-reducers dimension characterized by a finite set of transformations $\{W_m^*, m=1,..,M\}$. At each of these d-reducers, the corresponding misclassification rate is then estimated from a supervised sample test that we denote by $\{(V'_j, Y'_j), j = 1, .., N'\}$.

In accordance with the notations introduced in (Drira and Ghorbel, 2012), the misclassification rate is given by the following quantity:

$$\frac{card\ \left\{ Y'_j \neq Arg\left[\max_k \hat{n}_k \sum_{i=1}^{N} K\left(\frac{V'_j - V_i}{h_N}\right) I_{[Y_i=k]}(V'_j)\right], j = 1, \dots, N' \right\}}{N'}$$

The d-reducer retained is the one who realizes the lowest misclassification rates. This minimum is well approaching the theorical probability of error. Through this series of non-parametric procedures, we tried to approach the Bayes classifier in a multivariate frame and that not necessarily for Gaussian distributions, since, in each phase any hypothesis about the type of law has been issued.

Therefore the proposed system of algorithms consists on the three following tasks:

- First we estimate the L^2 measure of the probabilistic dependence with orthogonal basis in order to realize different multivariate extractors resulting from a number of initializations of the used numerical optimizing procedure.
- Second, an estimation of the miss classification error is computed for each solution by the modified kernel estimate of the conditional probability density functions in the context of the optimal smoothing parameter. Such parameter is obtained by minimizing the Mean Integrate Square Error (MISE). The Plug in algorithm tries to provide this solution.
- Third, the sub space which presents the minimum of the miss classification values is therefore chosen.

In each task, the numerical optimization procedure does not necessarily give optimal solutions. So we cannot be sure that we are realizing Bayesian classifier at each time.

4 FACE CLASSIFICATION PROCESS

In face recognition, a huge number of classifiers were introduced in the litterature having various success rate according to the application type. While feature extraction is compulsory phase of a pattern recognition system, this approche is offering a convenient departure of the the classification of the feature vectors, which oriented the researches in the face recognition domain to introduce a large number of face feature extraction methods.

In this study, we have employed the "BioID" dataset (Jesorsky et al., 2001), composed of 1521 images in gray level of 23 faces of frontal view, for each face image of this database 20 feature points are displayed. For any used algorithm, facial recognition is accomplished in four step process: the acquisition, the face detection, the feature extraction and finally the classification. In this paragraph we will describe the details of all the steps used in our work to accomplish the face classification process.

As a first step, we selected the Adaboost to detect the face and characteristic features location knowing that the most of the existing methods for facial feature extraction assume that at least coarse location of the face is detected. Then, after this operation, the computational complexity of the facial feature extraction can be significantly reduced.

In the second step, we move on to the face normalization which is very important stage for the recognition algorithms. First we start with a geometric normalization resumed in performing a rotation of face to align the axis of the eye with the horizontal axis and then we recover a face image whose distance between centers of the eyes is fixed. The dimensions of the face image are retrieved from the distance between the obtained eyes centers. In this phase we set the position of the mouth center in the normalized image in order to get acceptable column normalization and to ensure that the different face parts (eyes, mouth and nose) are in the same position for all faces. We apply next an increase in the dynamics to the normalized image, which is based on a decrease in the center of the image histogram to achieve images with the same ranges of distribution of gray levels and an average alignment of these levels. Second we apply an illumination normalization using the histogram equalization to re-calibrate the grayscale image leading to better contrast and a gamma correction to reduce the gap between light and dark areas of the face using a nonlinear transformation of grayscale (Fig 1).

When the facial regions could be retrieved, the analysis will be focused on the facial features. The adopted method locates features coarsely by searching areas with low intensity among possible face regions. This approach involve basic computer vision operations such morphology and projection analysis. The morphological operators can fit to it regarding their easy and fast implementation added to their strong nature. The projections also could be computed easily and are convenient with real-time applications. Some of the dark side of the projection methods is that the gray scale information is deeply influenced by a variation of illumination conditions and noise. Due to that, projection curves are not smooth, which keeps them difficult to be analyzed automatically. And therefore will try to use the geometrical face model hand to hand with the projections and morphology to avoid such problems. First, we apply the projection analysis: the gray-level intensity for the facial features is too weak in the image compared to close neighbors. So, the position of facial features can be reached by projections of the image. The significant minima are issued from the retrieved horizontal projection. For these minima the vertical projection is calculated and significant minima are searched for again. The obtained results will be treated as feature candidates. Next and once we apply horizontal projection to the facial images and we got the base lines, a morphological operator will be used to find the eyes position. And since the position of eyes and intraocular distance are almost similar for most of

people, detecting the eyes is a compulsory component in automatic facial feature extraction systems; it will be helpful to define the rest of facial features. We first locate the possible eye regions by detecting the valley points in an image and because the human iris is darker than its surrounding, a valley exists on the eye region, it will be extracted thanks to the gray-scale morphological operators (Fig 3).

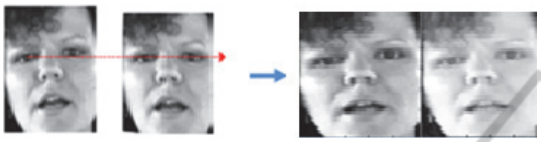

Figure 1: Geometric normalization and illumination normalization.

Finally, we apply method based on the geometrical face model. This model includes facial features like eyes, eyebrows, nostrils and mouth. The line passing through the eyes centers is called a base line. The geometric face model gets the configuration among eyes, nostrils, and mouth to locate facial features. Supposing that in most of the faces, the vertical distances between eyes and nose and between eyes and mouth are proportional to the horizontal distance between the two centers of eyes (Fig 2).

A Principal Component Analysis PCA is performed on the data sets after those who have all principal component's eigenvalue smaller than one millionth of the total variance and by this way problems related to near singular covariance matrices are kept away and all three transformations can be properly found out. The evaluation consists in dividing randomly the data set into k non overlapping folds of equal size, and for k times, each time choose one fold to be assigned as a test data and the others will be combined to issue the training data. The selection of the number of folds k relies on the bias-variance trade-off. In order to provide a good bias-variance compromise, by default we use "20-fold" CV which is widely accepted (Effron, 1983).

Next step, we search for the three Linear Reduction dimension transformations using the transformed train data and we reduce the dimensionality of this train to d where is in {1,2}.

In the last step, in the d-dimensional reduced feature space, we apply the nearest mean, the linear and Optimal Bayes classifier using the train data and on the other hand we classify the test data after transforming its instances in the same way as the train instances. The classification error is estimated on the test data.

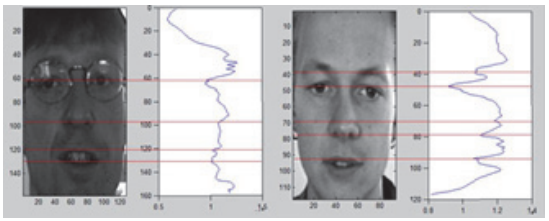

Figure 2: Horizontal projection of facial images from the BioID Face Database.

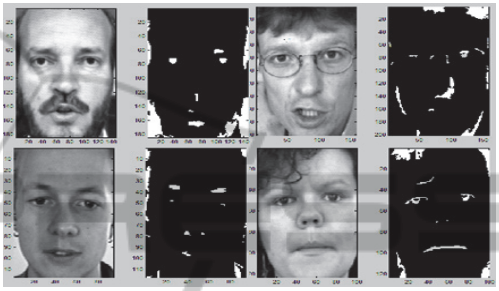

Figure 3: Morphological operator for some examples of the BioID Database.

In table 1, we present the observed Mean Square Error MSE for the data Set BioID using the three different algorithms. The estimated MSE: using features defined by Eigen faces representation is noted "FULL"; using the previous features based on principal component analysis is noted "PCA" and using the features database is noted "D". We note that the average error rates of the 2D L^2-PDM / Bayes combination compare favorably to those of other techniques. This advantage explains the link between the probabilistic dependence measure and the probability error of Bayes.

These MSE rates seem to correlate with the classification problem: In case of using the K Nearest Neighbors classifier, we can see that both 2D L^2PMD and the LDA are ranked as better result than IDA with a better advance for 2D L^2PMD. For the linear classifiers, the optimal results were provided by IDA and 2D L^2PMD showing the best overall performance far away from the best performances of the LDA technique. We should note that the feasibility of LDA is seriously limited by the constraint d < K (number of classes), note also that this approach cannot be realizable in the case of reduction features defined by Eigen faces representation without applying the ACP method.

Table 1: Observed MSE for the data Set BioID using the three different Classifier, the **K** **N**earest **N**eighbors, the **l**inear and the **o**ptimal **Bayes** classifier. Optimal observed MCE per classifier is typeset in bold.

		FULL	ACP	D
LDA	**K NN**	-	**0.3000**	0.3105
	L	-	0.3366	0. 3366
	Bayes	-	0.3793	0.3910
IDA	**K NN**	0.3512	0.3679	0.3492
	L	0.3360	**0.2814**	**0.2692**
	Bayes	0.3119	0.3119	0.2805
2D L^2PMD	**K NN**	**0.3103**	**0.2963**	**0.3007**
	L	**0.3015**	**0.2815**	**0.2605**
	Bayes	**0.2165**	**0.2165**	**0.2300**

5 CONCLUSIONS

In this paper, we introduced a pattern classification system for supervised classification composed of a series of novel and efficient algorithms which is able to realize a non parametric Bayesian classifier for high dimension. The proposed system is aiming to search for the best discriminate sub space in the mean of the minimum of the probability error of classification which is computed by using a modified kernel estimate of the conditional probability density functions. Therefore, Bayesian classification rule is applied in the reduced sub space, with the optimal MISE of the modified kernel estimate. Thus, the performance of the suggested system was compared to the other process based on classical algorithms in the real dataset in face classification. In the future works, we intend to evaluate the effectiveness of this process by studying the classification accuracy of a Bayesian classifier in term of probability of error based on hermit basis.

REFERENCES

Devijver, P. A., Kittler, J., 1982. *Pattern recognition: A statistical approach*, Prentice-Hall, Englewood Cliffs, NJ.

Drira, W., Neji, W. and Ghorbel F., 2012. Dimension reduction by an orthogonal series estimate of the probabilistic dependence measure. *The International Conference on Pattern Recognition Applications and Methods ICPRAM 2012,* Portugal, February, 2012.

Drira, W., Ghorbel F., 2012. Non Parametric Feature Discriminate Analysis for High Dimension. *The 16 International Conference on Image Processing, Computer Vision, and Pattern Recognition IPCV*, Las Vegas USA, July 2012.

Effron, B., 1983. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J. Am. Statistical Assoc.*

Fisher, R. A., 1936.The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, vol. 7, 179-188.

Hillion, A., Masson, P., Roux, C., 1988. A nonparametric approach to linear feature extraction; Application to classification of binary synthetic textures. *9th International Conference on Pattern Recognition, ICPR*, 1036-1039.

Jesorsky, O., Kirchberg, K., Frischholz, R. 2001.Robust Face Detection Using the Hausdorff Distance. *In Proc. Third International Conference on Audio and Video based Person Authentication - AVBPA 2001*, pp. 90-95, https://support.bioid.com/downloads/facedb/index.php.

Loog, M., Duin R. P. W., Haeb-Umbach R., 2001. Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria. *IEEE transaction on PAMI*, vol. 23, n° 7.

Nenadic Z., 2007. Information Discriminant Analysis: Feature Extraction with an Information-Theoretic Objective. *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 29, n° 8.

Patrick, E. A., Fischer, F. P., 1969. Non parametric feature selection. *IEEE Trans. On Inf. Theory*, vol. IT-15, 577-584.