

DIMENSIONALITY REDUCTION USING PROBABILISTIC DISTANCE MEASURES FOR PATTERN CLASSIFICATION

Faycel El Ayeb and Faouzi Ghorbel

CRISTAL Laboratory, GRIFT Research Group, National School of Computer Sciences, University of Manouba 2010, Tunisia
faycel.elayeb@cristal.rnu.tn , faouzi.ghorbel@ensi.rnu.tn

ABSTRACT

In this paper, we propose a dimensionality reduction method for classification purpose. It is based on a new estimation of a Patrick-Fischer (PF) distance for the two class case and for the multi-class one using orthogonal functions. Simulation studies and applications to the handwritten digits classification are presented. A comparison between the performances in terms of classification accuracy of the method proposed and those of the standard one based on the Fisher Linear Discriminant Analysis (FLDA) is made by evaluating a non parametric Bayesian classifier applied on the projected data on the reduced space given by these two methods. Experiment results indicate that the proposed method provides a better performance than the FLDA one.

1. INTRODUCTION

In pattern classification applications with high dimensional dataspace, the training sample size is required to be large to obtain satisfactorily class description. Collecting such training sample is difficult, expensive and can be impossible. Another serious problem that occur when dealing with these kind of data concern their high demand on computation time. Thus, dimensionality reduction could be used as an important step before any classification process. In the next section, we remind the fundamentals of the FLDA method [6]. Discussion about some of its limits followed. In section 3 we derive equations for the new estimator of PF distance and a new one of the measure of probabilistic dependence [4]. Section 4 contains classification experiments results obtained by the application of the two methods on simulation data and on real world dataset. Finally, conclusion and Future work are given.

2. THE FLDA METHOD

2.1. Theoretical background

The FLDA is a widely used method for dimensionality reduction. It intend to reduce the dimension, so that in the

new space, the between class distances are maximized while the within class distances are minimizing. To that purpose, FLDA considers searching for orthogonal linear projection matrix w that maximizes the following so-called Fischer optimization criterion [5,6]

$$J(w) = \frac{tr(w^T S_b w)}{tr(w^T S_w w)} \quad (1)$$

S_w is the within class scatter matrix and S_b is the between class scatter one. They are given by

$$S_w = \sum_{k=1}^c \pi_k E_k \left[(X - \mu_k)(X - \mu_k)^T \right] \quad (2)$$

$$S_b = \sum_{k=1}^c \pi_k (\mu_k - \mu)(\mu_k - \mu)^T \quad (3)$$

where $\mu_k = E_k[X]$ is the conditional expectation of the multidimensional random vector X given the class k , μ corresponds to the mean vector over all classes and π_k denote the prior probability of the k^{th} class.

A well-known estimation of S_w and S_b based on a given supervised data samples could be written as

$$\hat{S}_w = \frac{1}{N} \sum_{k=1}^c \sum_{i=1}^{n_k} \left[(x_i^k - \hat{\mu}_k)(x_i^k - \hat{\mu}_k)^T \right] \quad (4)$$

$$\hat{S}_b = \frac{1}{N} \sum_{k=1}^c \hat{n}_k (\hat{\mu}_k - \hat{\mu})(\hat{\mu}_k - \hat{\mu})^T \quad (5)$$

where $\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i^k$, $\hat{\mu} = \frac{1}{c} \sum_{k=1}^c \hat{\mu}_k$, $\hat{\pi}_k = \frac{\hat{n}_k}{N}$ and $N = \sum_{k=1}^c \hat{n}_k$.

Here x_i^k represent the i^{th} sample from the k^{th} class and n_k is the sample size associated to each class k .

Because it's not practical to find an analytical solution w that verify the criteria J , one possible suboptimal solution is to choose w formed by the d first eigenvectors of $S_w^{-1} S_b$ those correspond to the d largest eigenvalues. In general, the value of d is chosen to be equal to the number of classes minus one.

After computation of w , the FLDA method proceeds to the projection of the original data onto the reduced space spanned by the vectors of w .

2.2. Limitations

Note that this method is based only on second order moment and thus it assumes that the different underlying distributions of classes are normally distributed. This restrictive assumption constitute a limitation to using FLDA and make it fail when dealing with many interesting real world datasets in which class distributions are non-gaussian.

Note also that the Fischer optimization criterion J cited above involve differences between class means. Therefore, the solution given by this criterion will be optimal only for certain cases in which classes has distributions that are unimodal and class means are well separated. In the other cases, when datasets contains multimodal class distributions or having at least two class means that are near to each other or equal, FLDA method will obviously fail.

In order to overcome the limitations mentioned above, we proposed a new method that uses a new estimation of the PF distance for the two class case and a new one of the measure of probabilistic dependence for the multiple class case based on orthogonal functions [4].

3. THE METHOD PROPOSED

3.1. The PF distance estimation based on Kernel estimator

The PF distance estimation based on the kernel estimator has been originally introduced by Patrick and Fischer in their famous paper [3]. Further, Hillion and al. applied it to texture classification [1]. Such estimator is obtained by substituting into the expression of PF distance the probability density function (PDF) with the kernel density estimation (KDE) [2]. After some computations the expression of the PF distance could be written as

$$\hat{d}_p(f_1, f_2) = \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_1} C_{i,j}^{1,1} + \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} C_{i,j}^{2,2} - 2 \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} C_{i,j}^{1,2} \right)^{1/2} \quad (6)$$

$$\text{where } C_{i,j}^{m,n} = \frac{\pi_m \pi_n}{n_i n_j} \left(\frac{1}{\sqrt{2\pi\sigma}} \right)^d \exp\left(-\frac{1}{4\sigma^2} |w(x_i^m - x_j^n)|^2 \right) \quad (7)$$

For dimensionality reduction purpose, the PF distance estimation based Kernel estimator is considered as the criterion function to be maximized with respect to a linear projection matrix w that transform original dataspace onto a d -dimensional subspace so that classes are most separated.

3.2. The proposed PF distance estimation based on orthogonal expansion estimator

The proposed PF distance estimator in [4] is obtained by replacing the PDF in the PF formula by an estimation based on orthogonal functions. After simplification due to

the orthogonality of the used system of functions, the PF distance can be expressed as following

$$\hat{d}_p(f_1, f_2) = \left(\frac{1}{N^2} \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_1} K(\langle x_i^1 | w \rangle, \langle x_j^1 | w \rangle) + \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} K(\langle x_i^2 | w \rangle, \langle x_j^2 | w \rangle) - 2 \operatorname{Re} \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K(\langle x_i^1 | w \rangle, \langle x_j^2 | w \rangle) \right) \right) \right)^{1/2} \quad (8)$$

where $K(x,y)$ is the kernel that correspond to the orthogonal system of functions $\{e_m(x)\}$ according to the following equation

$$K(x, y) = \sum_{m=1}^{k_n} e_m(x) e_m^*(y) \quad (9)$$

In the above equation k_n is the so-called truncation point and e_m^* denote the complex conjugate of e_m .

To reduce dimensionality, a linear projection matrix w that maximizes this PF distance estimator should be found. Since the equation of this new estimator is highly nonlinear according to the element of w and an analytical solution is often practically not feasible, we will resort to an optimization algorithm to compute a suboptimal projection matrix w .

The projection of the high dimensional dataset according to the determined w allows us to obtain the new samples coordinates onto the reduced space.

3.3. The proposed PF measure of probabilistic dependence estimation based on orthogonal expansion estimator

The PF measure of probabilistic dependence [4] is used to measure the class separability between several classes. It forms an extension of the PF distance for the multi-class case. In this case, we have to estimate a multivariate PDF. This is not a difficult task since the orthogonal functions in multidimensional case can be considered as the product of one-dimensional orthogonal functions [8]. By replacing into the equation of the PF measure of probabilistic dependence, the multivariate PDF with an estimator based on orthogonal functions we obtain a new measure of probabilistic dependence estimator based on orthogonal functions [4] that is defined as

$$\hat{I}_p(f, f_i) = \left(\frac{1}{N^2} \sum_{i=1}^c \sum_{k=1}^N \sum_{j=1}^N K(\langle x_j^i | w \rangle, \langle x_k^i | w \rangle) + \frac{1}{N} \sum_{i=1}^c \frac{1}{n_i} \sum_{k=1}^{n_i} \sum_{j=1}^{n_i} K(\langle x_j^i | w \rangle, \langle x_k^i | w \rangle) - \frac{2}{N^2} \operatorname{Re} \left(\sum_{i=1}^c \sum_{k=1}^N \sum_{j=1}^{n_i} K(\langle x_j^i | w \rangle, \langle x_k^i | w \rangle) \right) \right)^{1/2} \quad (10)$$

As with the new PF distance estimator proposed, we are invited to find a suboptimal solution w that maximize this new estimation of the measure of probabilistic dependence.

4. EXPERIMENTAL RESULTS

In this section, we intend to compare the performances of the proposed method with the FLDA technique both on simulated data and on real world dataset. To do that, we evaluate the classification accuracy of a nonparametric Bayesian classifier that is applied on the projected data onto the reduce space obtained by these two dimensionality reduction methods.

The classifier used is designed by replacing in the Bayes rule the PDE with its KDE. It assigns a given sample to the class with the highest KDE. We evaluate the classification accuracy by counting the number of misclassified samples obtained by the classifier over all classes of the projected data.

4.1. Experiments with simulated data

Note that throughout all the experiments described in the remainder of this paper, we choose the value of k_n to be equal to 3. For each class, we will generate 100 random samples with fourteen-dimensions. A unit covariance matrix will be used for all gaussian distributions and the Trigonometric system [4,7,8] will be chosen as an orthogonal system of functions.

The first experiment concerns the two-class case. Samples from the class 1 are drawn from a multivariate gaussian distribution with mean vector $\mu_{1=[3...3]}^T$.

For the class 2, samples are drawn from a mixture of two multidimensional gaussian distributions. The first distribution have a mean vector $\mu_{2=[1...1]}^T$, and the second one have a mean vector $\mu_{3=[7...7]}^T$.

The second experiment is related to the multi-class case. It considers three groups of data.

For the class 1, samples are drawn from a multivariate gaussian distribution with mean vector $\mu_{1=[3...3]}^T$.

For the class 2, samples are drawn from a mixture of two multidimensional gaussian distributions. The first distribution have a mean vector $\mu_{2=[1...1]}^T$ and the second one have a mean vector $\mu_{3=[11...11]}^T$.

For the class 3, samples are drawn from a multivariate gaussian distribution with mean vector $\mu_{4=[5...5]}^T$.

For each one of these experiments, we search for the projection matrices that map the generated data onto the optimal n-dimensional subspace (here we choose $n=1$ for the first experiment and $n=2$ for the second one) according to the proposed method and the FLDA procedure.

After finding the projection matrix for each case, simulated data are projected onto the reduced space. Then, the classifier based on the projected data is evaluated as indicated above.

Figure 1 and Figure 2 below shows the projection of the generated data for the second experiment onto the optimal two-dimensional subspace with according to the method proposed and to the FLDA one respectively.

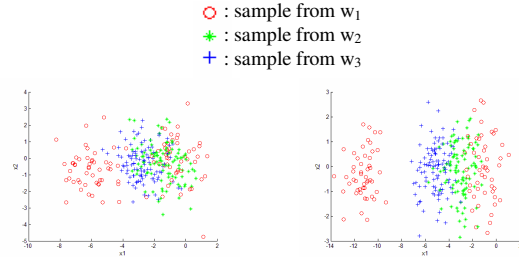


Figure 1: Projection of simulated data onto the reduced subspace using the FLDA method (3-class case).

Figure 2: Projection of simulated data onto the reduced subspace using the proposed method (3-class case).

Results given by these two figures, indicate that discrimination quality of the proposed method seems better than the one of the FLDA procedure.

Evaluation results of classification are summarized in Table 1.

Table 1. Experiments with simulated data

	FLDA method	Proposed method
Number of misclassified samples over 200 (1 st Experience)	36	19
Number of misclassified samples over 300 (2 nd Experience)	51	17

As can be seen on the Table 1, the classification accuracy of the proposed method is consistently better than the one of FLDA technique as well for the two-class classification problem as for the case of multi-class problem. The FLDA method fails to find an optimal subspace in which satisfactorily class separation is obtained since the original simulated data contain multimodal distribution (class 2). However, the proposed method succeeds to overcome the restriction of unimodality. The success of the proposed approach even when classes are multimodal can be explained by the fact that the proposed probabilistic distance function estimator accounts for higher order statistics and not just for the second order as in the Fisher criterion.

4.2. Experiment with a real dataset

In this experiment, we consider a sample set selected from the publicly available MNIST database containing binary images of handwritten digits. Our selected sample set contains three classes. Each one is formed by a 100 randomly selected digits. Figure 3 shows an example of selected digits.

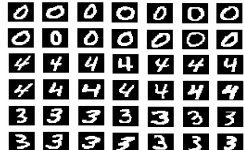


Figure 3: Some examples of characters selected from MNIST database.

Each digit from this selected sample set will be described by means of its Fourier descriptors (FD) [1,5]. The FD are calculated from the digit outline boundary and are chosen to be invariant regarding to the elementary geometrical transformations, such as translation, rotation and scaling. In the following experiment, we will sample each digit boundary to fourteen pixels. Thus, we are invited to compute fourteen FD to each digit. The set of the FD vectors obtained for the whole digits in our selected sample will form the shape descriptors dataset to be used by the proposed dimensionality reduction method and by the FLDA one.

As in the previous experiments, a non linear Bayesian classifier will be trained and evaluated based on data projected onto the optimal two-dimensional subspace.

Figure 4 and Figure 5 displayed below visualize the projection of the shape descriptors dataset onto two-dimensional subspace with according to the method proposed and to the FLDA one respectively.

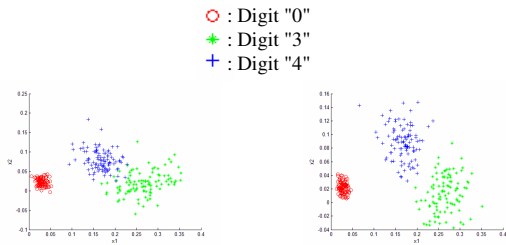


Figure 4: Projection of shape descriptors dataset onto the reduced subspace using the FLDA method (3-class case).

Figure 5: Projection of shape descriptors dataset onto the reduced subspace using the proposed method (3-class case).

Following the results showed in figure 4 and figure 5, the proposed method gives a well separation between the three classes. However, the FLDA method fail to separate between the class that correspond to the digit 3 (green asterisks) and the one that correspond to the digit 4 (blue cross).

Evaluation results of classification for this experiment are given by the Table2.

Table 2. Experiment with a real dataset

	Number of misclassified samples over 300
FLDA method	8
Proposed method	5

Results showed in Table2 indicate that the proposed method slightly outperforms the standard one based on Fischer criterion.

5. CONCLUSION AND FUTURE WORK

In this paper, a new method for dimensionality reduction is proposed. Its novelty lies on the using of a new estimation, based on orthogonal functions, of the Patrick-Fischer distance and a new one for the measure of probabilistic dependence.

The simulation studies and the real dataset experiments have shown that the suggested method increases the separability measure between the projected classes onto the reduced space and decreases the number of the misclassified samples.

In our future work, we will concentrate on the selection of a better parameter value of the truncation point k_n rather than choosing its value arbitrarily.

6. REFERENCES

- [1] A. Hillion, P. Masson and C. Roux, "Une méthode de classification de textures par extraction linéaire non paramétrique de caractéristiques", Colloque TIPI, vol. 5, no. 4, 1988.
- [2] E. Parzen, "On estimation of a probability density function and mode", The Annals of Mathematical Statistics, vol. 33, no. 3, pp. 1065-1076, 1962.
- [3] E. A. Patrick and F. P. Fischer, "Non parametric feature selection". IEEE Trans. Information Theory, vol. IT-15, no. 5, 1969.
- [4] F. Ghorbel, "Vers une approche mathématique unifiée des aspects géométriques et statistiques de la reconnaissance de formes planes", Doctorat thesis, University Rennes I, 1990.
- [5] F. Ghorbel and J. L. Bougrenette la Tocnaye, "Automatic Control of Lamellibranch Larva Growth Using Contour Invariant Feature Extraction", Pattern Recognition, vol. 23, no. 3/4, pp. 319-323, 1990.
- [6] K. Fukunaga, "Introduction to Statistical Pattern Classification", Academic Press, New York, 1990.
- [7] M. Zribi and F. Ghorbel, "An unsupervised and non parametric Bayesian classifier", Pattern Recognition Letters 24, pp. 97-112, 2003.
- [8] M. Tarter and R. Kronmal, "On Multivariate Density Estimation Based On Orthogonal Expansions", The Annals of Mathematical Statistics, vol. 41, vo. 2, pp. 718-722, 1970.