DIMENSION REDUCTION BY AN ORTHOGONAL SERIES ESTIMATE OF THE PROBABILISTIC DEPENDENCE MEASURE

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

Keywords: Feature extraction, Probabilistic dependence measure, Kernel estimate, Fourier series.

Abstract: Here, we intend to introduce a new estimate of the L^2 probabilistic dependence measure by Fourier series for 2-dimensional reduction. Its performance is compared to the Fischer Linear Discriminate Analysis (LDA) and the Approximate Chernoff Criterion (ACC) in the mean of classification probability error.

1 INTRODUCTION

One of the well studied problems in the statistical pattern recognition field is feature selection. It is well known that pattern recognition systems usually need many features to improve its performances. The definition of suitable representation requires generally a serious study in order to consider only pertinent features. A large training and test samples are therefore constructed in order to design and evaluate the performances. In practice, the training sample serves to estimate conditional probability density functions of each class. The convergences of the different estimates have been well studied in the statistical literature according to the size N of the training sample. It is well established that the sample size which is needed to estimate the probability density function as the histogram, the kernel or the orthogonal series, has to increase exponentially with the dimension D of the feature vector. In order to undergo over such limitation, the discriminate analysis for reducing the dimensionality is generally required. The convergence of such algorithms could be possible since the estimation of the criteria is realized in the reduced d-space (d<<D). However, when one of the conditional distributions is not Gaussian, the popular Fisher discriminate analysis based on the scatter matrices (Fisher, 1936, Loog et al., 2001) could give a not well reduced d-space. Later, Patrick and Fisher have proposed a non parametric solution based on probability density function distances. In the same work, they have introduced a kernel estimate for the Patrick-Fisher distance (Patrick and Fisher, 1969).

A. Hillion (1988) has applied a Gaussian kernel estimate for a scalar feature extraction in order to classify image by its texture. They showed experimentally the better performance of their method relatively to the Fisher one in the mean of the probability error. Thus, we intend to introduce new estimators of the dependence probabilistic measure by using the density orthogonal estimate

This estimator could be extended to the multivariate reduction. The optimization of smoothing parameters will be discussed in the present work.

2 L^2-PDM ESTIMATE

A rectangular matrix W from a D-dimensional feature space to a d-dimensional reduced space is obtained by optimizing criteria which are defined according to the between scatter and the within scatter matrices. Note that those matrices are defined from first and second order statistical moments of the conditional random vectors of observation. For multimodal conditional distributions the feature extraction algorithms based on these scatter matrices called linear discriminate analysis (LDA) could not give the best classifier in the sense of the probability error. In order to dispose of this limitation, distances between the conditional probability density functions weighted by the prior probabilities have been suggested in the literature. An estimate of such distance has been proposed for scalar extraction in

Drira W., Neji W. and Ghorbel F. (2012).
DIMENSION REDUCTION BY AN ORTHOGONAL SERIES ESTIMATE OF THE PROBABILISTIC DEPENDENCE MEASURE.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 314-317
DOI: 10.5220/0003845403140317

the context of binary classification. The d-multivariate reduction has been extended by the pursuit procedure. We propose here to extend this procedure to the 2-dimensional reduction by using the orthogonal series probability density estimate.

We begin by defining a new orthogonal series estimate for the scalar reduction which will serve to derive the multivariate reduction.

The orthogonal probability density estimate assumes that the density function belongs to Hilbert space which is having an orthogonal basis functions $e_l(x)$.

According to a sample $\{X_i, i = 1..N\}$ of a random variable *X*, the estimate can be written as:

$$\hat{f}(x) = \frac{1}{N}\sum_{j=1}^{N} K_{m_N}(x, X_j) \quad (1)$$

$$\text{Where} \quad K_{m_N}(x, X_j) = \sum_{l=1}^{m_N} e_l(x) e_l(X_j) \quad (2)$$

and m_N is a sequence of integer which is similar to the well known smoothing parameter for kernel density function estimator. Thus, the L^2-Probabilistic Dependence Measure (L^2-PDM) estimate $\hat{I}_2$ could be deduced and expressed according a supervised sample ${X_i}^k$ as following:

$$\begin{aligned}\hat{I}_2 &= \frac{1}{(\sum_{k=1}^{K} N_k)^2}\sum_{k=1}^{K}\left(\sum_{i,j=1}^{N_k}[K_{m_k}({X_i}^k, {X_j}^k)]\right. \\ &+ \sum_{r,l=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{N_l}[K_{m_N}({X_i}^r, {X_j}^l)] \\ &\left. - 2\sum_{l=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{N_l}[K_{m_k}({X_i}^k, {X_j}^l)]\right)\end{aligned} \quad (3)$$

With N_k the number of instances for the label k. Now, we assume that the two dimensional conditional probability densities of the class k, belong to Hilbert space $L^2(R^2)$ which has $e_{l,k}(v)$ as an orthogonal basis. Their estimates could be written according a supervised sample ${V_i}^k$ as follow:

$$\hat{f}_{V^k}(v) = \frac{1}{N}\sum_{i=1}^{N} \tilde{K}_{m_N}(v, {V_i}^k) \quad (4)$$

$$\text{Where } \tilde{K}_{m_N}(v, V_i) = K_{m_N}(x, X_i).K_{m_N}(y, Y_i) \quad (5)$$

The 2D L^2-Probabilistic Dependence Measure (2D L^2-PDM) estimate can be expressed as:

$$\begin{aligned}\hat{I}_P &= \frac{1}{(\sum_{k=1}^{K} N_k)^2}\sum_{k=1}^{K}\left(\sum_{i,j=1}^{N_k}[\tilde{K}_{m_k}({V_i}^k, {V_j}^k)]\right. \\ &+ \sum_{r,l=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{N_l}[\tilde{K}_{m_k}({V_i}^r, {V_j}^l)] \\ &\left. - 2\sum_{l=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{N_l}[\tilde{K}_{m_k}({V_i}^k, {V_j}^l)]\right)\end{aligned} \quad (6)$$

3 EXPERIMENTAL RESULTS

Data sets used for tests were taken from the UCI Repository of machine learning databases (Murphy and Aha, 2004). We have chosen 10 real-world data sets that come from a variety of applications. These data sets, labeled (a) to (j), have a various numbers of classes and attributes and various sample sizes (Table 1). Instances with missing values were taken out of the data sets prior to the experiments. The number of test instances is given in table 1 as they were designated by their donors. For all other data sets, a k-fold Cross-Validation (CV) was used. The justification of the choice of k will be given later.

Table 1: The 10 data sets used in the experiments. Information is provided on initial dimensionality D, dimensionality after principal component analysis PC, number of classes K, number of total instances N and validation type.

Data set	Label	D	PC	K	N	Validation
Breast cancer	(a)	9	9	2	683	20-fold
liver disorders	(b)	6	6	2	345	20-fold
Diabetes	(c)	8	8	2	768	20-fold
diagnostic breast cancer	(d)	30	7	2	569	20-fold
Heart disease	(e)	13	13	5	298	10-fold
Iris	(f)	4	4	3	150	10-fold
Thyroid	(g)	21	21	3	7200	3428
Karhunen-Love	(h)	64	64	10	2000	200
Glass identification	(i)	10	8	7	214	10-fold
Breast Tissue	(j)	9	5	6	106	10-fold

3.1 The Experimental Setup

In order to determine properly all three transformations, problems related to near singular covariance matrices should be avoided. Such a problem can be solved by performing a PCA on the train set of every of the 10 data sets, where only the principal components with an eigenvalue bigger than one millionth of the total variance are kept.

For data sets (g) and (h), the transformation matrices **W** were estimated from the training data, which was then transformed to a subspace of appropriate dimension.

Table 2: Observed MCE for the 10 data sets (a) to (j) for the reduced dimensions **d=1**, Using the three mentioned classifiers Linear classifier (L), Quadratic classifier (Q) and Nearest Mean classifier (NM) and the three different reduction techniques indicated by LDA, ACC and PF/L^2-PDM. The estimated MCE using no reduction is below "FULL".

	FULL			LDA			ACC			PF/L^2-PDM		
	L	**Q**	**NM**	**L**	**Q**	**NM**	**L**	Q	**NM**	**L**	**Q**	**NM**
(a)	0.0513	0.0513	0.0339	0.0369	0.0369	**0.0367**	**0.0018**	**0.0018**	0.1709	**0.0039**	**0.0039**	**0.0340**
(b)	0.3913	0.3913	0.4253	0.3865	0.3865	**0.3672**	**0.0193**	**0.0193**	0.4023	**0.0236**	**0.0236**	0.4234
(c)	0.2599	0.2599	0.3644	0.2571	0.2571	**0.2415**	**0.0128**	**0.0128**	0.3054	**0.0161**	**0.0161**	0.5792
(d)	0.0560	0.0560	0.1119	0.1027	0.1027	**0.0557**	**0.0051**	**0.0051**	0.1395	**0.0060**	**0.0060**	**0.0559**
(e)	0.6654	0.6654	0.5999	0.4274	0.4274	0.6532	**0.0427**	**0.0427**	0.7404	**0.0451**	**0.0451**	**0.4198**
(f)	0.0187	0.0187	0.0633	0.1312	0.131	0.0383	0.0131	0.0131	0.0312	0.0428	0.0428	**0.0187**
(g)	0.9851	0.9851	0.5717	**0.0697**	**0.0697**	**0. 1307**	0.9798	0.9798	0.9766	**0.0749**	**0.0749**	**0.1759**
(h)	0.79	0.7950	0.765	0.73	0.73	0.71	0.7	0.7	0.77	**0.555**	**0.555**	**0.580**
(i)	0.5569	0.5569	0.1666	0.2600	0.2600	0.3028	0.0260	0.0260	0.5405	**0.0151**	**0.0151**	**0.1707**
(j)	0.2816	0.2816	0.4984	0.4135	0.4135	0.5363	**0.0413**	**0.0413**	**0.4075**	**0.0517**	**0.0517**	0.5653

Table 3: Observed MCE for the 10 data sets (a) to (j) for the reduced dimensions **d=2**, Using the three mentioned classifiers Linear classifier (L), Quadratic classifier (Q) and Nearest Mean classifier (NM) and the three different reduction techniques indicated by LDA, ACC and PF/L^2-PDM. The estimated MCE using no reduction is below "FULL".

	FULL			LDA			ACC			PF/L^2-PDM		
	L	**Q**	**NM**	**L**	**Q**	**NM**	**L**	Q	**NM**	**L**	**Q**	**NM**
(a)	0.0513	0.0513	0.0339	0.0369	0.0369	**0.0368**	0.0369	**0.0018**	0.1709	**0.0029**	**0.0029**	**0.0383**
(b)	0.3913	0.3913	0.4253	0.3865	0.3865	**0.3672**	0.3865	**0.0193**	0.4024	**0.0204**	**0.0204**	0.5032
(c)	0.2599	0.2599	0.3644	0.2571	0.2571	**0.2416**	0.2571	**0.0128**	0.3055	**0.0149**	**0.0149**	0.3914
(d)	0.0560	0.0560	0.1119	0.1027	0.1027	**0.0557**	0.1027	**0.0051**	0.1396	**0.0048**	**0.0048**	0.1047
(e)	0.6654	0.6654	0.5999	0.4274	0.4274	0.6532	0.4274	**0.0427**	0.7404	**0.0445**	**0.0445**	**0.4880**
(f)	0.0187	0.0187	0.0633	0.1312	0.1312	**0.0383**	0.1312	0.0131	**0.0312**	**0.0031**	**0.0031**	**0.0437**
(g)	0.9851	0.9851	0.5717	**0.0685**	**0.0685**	0.263	**0.0685**	0.9787	0.9766	**0.0749**	**0.0749**	**0.1759**
(h)	0.79	0.7950	0.765	0.545	0.545	0.525	0.545	0.615	0.61	**0.335**	**0.335**	**0.43**
(i)	0.5569	0.5569	0.1666	0.2600	0.2600	0.3028	0.2600	0.0260	0.5405	**0.0076**	**0.0076**	**0.1597**
(j)	0.2816	0.2816	0.4984	0.4135	0.4135	0.5363	0.4135	**0.0413**	**0.4075**	**0.0344**	**0.0344**	0.6583

For all other datasets, the evaluation consists of randomly divide the data set into K non overlapping folds of equal size and for k times , each time choose one fold to be designated as a test data and the others will be combined to compose the training data. The choice of the number of folds k is dictated by the bias-variance trade-off. 10 to 20-fold CV is widely accepted that it offers a good bias-variance compromise, and these values are often used as default. Stratification will give further improvements in terms of both bias and variance, where the relative class frequencies over folds roughly match those of the original data set. Therefore, for large data sets (N_i > 500), a stratified 20-fold CV was used (see Table 1). Other data sets were tested using a stratified 10-fold CV.

In the d-dimensional reduced feature space, the classification error is estimated empirically based on three different classifiers (Devijver and Kittler, 1982, Fukunaga, 1990):

-The linear classifier assuming all classes to be normally distributed with equal covariance matrix

-The quadratic classifier assuming the underlying distributions to be normal with covariance matrices that are not necessarily equal.

-Nearest mean classifier that is based on Euclidean distance to the nearest mean.

These classifiers are chosen because they stay close to the assumption that most of the relevant information is in the first and second order central moments, i.e., the means and the co-variances.

3.2 Analysis of Results

The per-data set-performances of these three reduction techniques are compared. Therefore, for each data set, per classifier, the mean estimated classification error over the multiple runs is determined for dimension d=1 and d=2 (see Table 2 and 3). This gives a final estimate of the classification error for the respective settings. The overall optimal error rate, per classifier, over all transforms is typeset in bold. To compare the results, we have used a signed rank test where the desired

level of significance is set to 0.01 (Rice, 1995). Tables 2 and 3 give the Mean Classification Error (MCE) obtained when not performing any dimension reduction.

We start with two general observations: First, the quadratic classifier, in general, gives better results for most of the data sets. This may indicate that in most data sets, there is indeed information separation present in the second order moments of the class distributions. Second, the average error rates after reduction to d=1 or d=2 remain, in general, smaller than those in the full space, thus confirming that a gain in performance can be achieved by reducing the dimensionality of the problem.

Also, note that the average error rates of the PF method compare favorably to those of other techniques for the 1d and 2d subspace dimensions. This advantage seems to correlate with the difficulty of the classification problem. In particular, for linear and quadratic classifier, PF is uniformly superior to other methods.

We begin with the analysis of the two-class problem (data sets a, b, c and d).In case of using the nearest mean classifier; we can see that the Patrick-Fisher criteria as well as the LDA ranked better result than ACC. For the quadratic and linear classifiers, the optimal results were provided by PF and ACC, with the best overall performance significantly different from the best performances of the LDA technique. Note that the performance of LDA is seriously limited by the constraint d < K (number of classes).

We now turn to the analysis of the multi-classes case were the K-fold CV was used (data sets e, f, i and j). Clearly, a similar analysis of the two class case is observed: where the advantage of PF persists and it is much better than LDA. Note that the PF and ACC error rates are in order of 10^{-2} whereas those of the LDA are in the order of 10^{-1}.

For data sets (g and h) where validation is based on a test set, the best error rates are those given by PF and LDA, these methods provide much better separability in data set than the ACC criteria or all classifiers results.

4 CONCLUSIONS

In this paper, 2D dimensionality reduction method is proposed. Its novelty lies on the study of a new L^2 probabilistic dependence measure estimate obtained by the orthogonal Fourier series expansion.

The real dataset experiments show that the suggested method increases the separability measure between the projected classes onto the reduced space consistently better than the well-known LDA method.

Since results given by the proposed method are promising and could be used as a step before a classification process. We will concentrate our future work on the evaluation of the effectiveness of this method by studying the classification accuracy in term of the probability error.

REFERENCES

Aladjem, M. E., (1996). Two class pattern discrimination via recursive optimization of Patrick-Fisher distance. *Proc. of the 13th International Conference on Pattern Recognition* (ICPR), vol. 2, pp. 60-64.

Devijver, P. A. and Kittler, J. (1982). Pattern Recognition: A Statistical Approach. London: Prentice-Hall.

Drira, W. and Ghorbel, F. (2011). Une 2D-réduction de dimension par un estimateur de la distance en probabilité de Patrick Fisher. *43èmes Journées de Statistique*, Tunis.

Drira, W. and Ghorbel F. (2010). Réduction de dimension par un nouvel estimateur de la distance de Patrick Fisher à l'aide des fonctions orthogonales. *42èmes Journées de Statistique*, Marseille.

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

Fukunaga, K. (1990) Introduction to Statistical Pattern Recognition. *New York: Academic Press*.

Hillion, A. (1988). Une méthode de classification de textures par extraction linéaire non paramétrique de caractéristiques. *Traitement du signal*, Volume 5,N° 4.

Loog, M. et al. (2001). Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria, *IEEE trans. on PAMI.*, Vol. 23 N°7.

Murphy, P. M. and Aha D. W., (2004).UCI Repository of Machine Learning Databases, http://archive.ics.uci.edu /ml/citation_policy.html.

Nenadic, Z., (2007) Information Discriminant Analysis: Feature Extraction with an Information-Theoric Objective. *IEEE Trans. on PAMI*, Vol. 29 N° 8.

Patrick, E A. and Fisher P F (1969). Non parametric feature selection. *IEEE Trans On In. Theory*, Vol. 15, 577-84.

Rice, J. A. (1995), *Mathematical Statistics and Data Analysis*, second ed. Belmont: Duxbury Press.