Aboozar HOSSEINZADEH, Majid NOORHOSSEINI

Amirkabir Universisty of Technology, Iran

Optimal complexity for discriminant analysis

Abstract *A mechanism has been proposed to achieve optimal complexity for Discriminant Analysis (DA) based on Principal Component Analysis (PCA). We use PCA to filter non-informative features before applying to DA algorithm. In addition to significant accuracy improvement, the mechanism decreases the computational and storage costs of Linear Discriminant Analysis methods and makes the overall method more efficient. The mechanism helps classical linear DA methods outperform the state of the art and most superior linear and non-linear DA methods.*

Streszczenie. *Zaproponowano mechanizm analizy złożoności w analizie dyskryminacyjnej bazującej na Analizie Składowej Głównej PCA. Dodatkowo mechanizm umożliwia poprawę dokładności klasyfikacji danych. (**Analiza optymalnej złożoności w analizie dyskryminacyjnej**)*

Keywords: Face recognition, Object recognition, Discrimnant analysis, Computer vision, Pattern recognition
Słowa kluczowe: analiza dyskryminacyjna, PCA, rozpoznawanie obrazu

Introduction

Bayes classifier is known to be optimal since it has a minimum classification error. In most classification problems due to mathematical simplicity and possibility of reasonable approximation, patterns are assumed to have normal distribution. The most investigated classification problems are two class problems due to the fact that their results and theories usually can be expand or at least be a base of an induction to a multiclass case. In the classification of a vector x into one of the two normal distributions, Linear Discriminant Function (LDF) will be the most successful method. One of the main concerns of a designer of a plug-in Bayes classifier, including an LDF, is the Hughes phenomenon [1], that is, the existence of a peak in the classifier accuracy as the number of measurements, p, increases.

In last decades several researchers have involved in proposing new Discriminant Analysis (DA) algorithms. Some examples of the state of the art and most successful linear and also nonlinear (kernel based) DAs are Subclass Discriminant Analysis (SDA) [1], approximate Pairwise Accuracy Criterion (aPAC) [3], and Complete Kernel Fisher Discriminant (CKFD)[7]. Linear Discriminant Analysis (LDA) [5], [6], as the multiclass case of LDF, is the most popular DA algorithm. LDA assumes the data in each class can be grouped in a single cluster, i.e. having unimodal classes. In practice, this assumption may be vastly deviated. Based on a Nearest Neighbor (NN) clustering of each class to an optimal number of subclasses, Zhu and Martinez [1] presented Subclass Discriminant Analysis (SDA) to address multimodality in data. Approximate Pairwise Accuracy Criterion (aPAC) [3], decomposed a c class (multiclass) Fisher criterion into the $c(c-1)/2$ two class criterion. It introduced weighting functions of the class pairwise Mahalanobis distance to the overall criterion. Heteroscedastic linear dimension reduction (HLDR) [2], also known as HLDA, utilizes the so-called Chernoff criterion in LDA in a way that it can engage the heteroscedasticity of the data. Heteroscedastic data have differences in within class covariance matrices and HLDR uses the discriminatory information therein. Both aPAC and HLDR are based on the eigen-decompositions of some functions of positive semidefinite matrices. Therefore, they need to be calculated by reliable numerical methods such as SVD. Also both aPAC and HLDR need a c class problem to be decomposed into $c(c-1)/2$ two class problems. This action will become inefficient when c is large.

In this paper, we propose a mechanism to eliminate 'Hughes phenomenon' from DA algorithms. Principal Component Analysis (PCA) [6], [9], [10] usually is used to fade away singularity challenge of within class scatter matrices [6], [9], in DA methods when one involves in small sample size problems. Our approach is based on using PCA as a preprocessing subspace to filter non-informative features before applying to DA algorithms. In addition to accuracy improvements, our mechanism decreases the computational and storage cost and makes the overall DA methods more efficient. Also, we show that the usual vulnerability of LDA based algorithms versus PCA is removed by the mechanism. We conduct intensive experimental experiences to demonstrate the effectiveness of the proposed approach. A large number of most popular face databases as well as one object recognition database are used to evaluate the superiority of the proposed approach over a large number of the strongest and state of the art DA methods. Finally, we propose two suggestions for researchers in the area of statistical pattern classification which may improve the accuracy of their designed classifiers.

Hughes phenomenon and its relationship to Linear Discriminant Analysis

Let $\Pi_i \sim \mathcal{N}_p(\mu_i, \Sigma_i)$, $i = 1,2$ be two p-variate normal populations and $\Sigma_1 = \Sigma_2$. When all parameters are known the optimum classification rule, namely Bayes, is

$$U_0 = \{x - 1/2\,(\mu_1 + \mu_2)\}^T \Sigma^{-1}(\mu_1 - \mu_2) \tag{1}$$

where the term

$$F_0 = (\mu_1 - \mu_2)^T \Sigma^{-1} x \tag{2}$$

is Fisher's LDF of the populations. Usually, x is assigned to class 1 if $U_0 > 0$ and to class 2, otherwise.

Equation (1) is obtained from the following minimum distance rule:

$$2U_0 = (x - \mu_2)^T \Sigma^{-1}(x - \mu_2) - (x - \mu_1)^T \Sigma^{-1}(x - \mu_1) \tag{3}$$

For a case of $\Sigma_1 \neq \Sigma_2$ the optimum rule is based on a quadratic form

$$U_1 = (x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1) - (x - \mu_2)^T \Sigma_2^{-1}(x - \mu_2) + log(|\Sigma_1|/|\Sigma_2|) \tag{4}$$

In contrast to (1), the evaluation of PMC (Probability of Misclassification) of (4), except for some special cases, is very complicated. When all parameters are not known the plug-in versions of (1) and (4) have been presented, by substituting $\bar{x}_i$ and S instead of μ_i and Σ_i, where $\bar{x}_i$ and S are the Maximum Likelihood estimation of μ_i and Σ_i respectively, based on N_i random samples from Π_i. It has been shown that when $N_i \longrightarrow \infty$, the distribution of the plug-in version of (1) tends to the true distribution. However,

when the sample size, N_i, is finite (limited training set), the estimations of $\bar{x}_i$ and S differ from their true values, μ_i and Σ_i. This is the reason for the degradation of plug-in versions and their suboptimal performances. An important question in this regard is: 'Is the Probability of Correct Classification (PCC) of the classifier monotone with respect to the number of measurements?' In other words, 'If $x \in U_p$, is it possible for a classifier to have a smaller PMC on a $U_{p'} \subset U_p$?' We refer interested readers to the fundamental work in [5] (especially chapters 35, 36, 39, and 40), and discuss only a brief and concise overview of them to answer questions regarding Bayes classifiers and their plug-in versions. It has been shown that increasing the number of measurements in the training set may decrease the accuracy of the classifier designed on this set. It has also been demonstrated by an experiment that while the Mahalanobis distance [5], [6], [9], [10], Δ_N^2, of the populations increases linearly to p, adding a new measurement, i.e. $U_p \mapsto U_{p+1}$, has no degradation effect in the classifier accuracy, except for cases with a very small N, where $N = \sum_{i=1}^{c} N_i$. Hughes [11], is the first work which shed more light on the above question by using mathematics. He built a mathematical model for establishing a peaking phenomenon in the mean accuracy of the Bayes classifier, trained on a training set with the size N_i, when p increases continuously. His model is commented on the selection of the Bayes rule estimation. Despite previous reports of the peaking phenomenon before his work, the phenomenon was called 'Hughes phenomenon' after his fundamental work. Since the phenomenon relates to p, it is also known as 'optimal measurement complexity' in literatures. In spite of existence of some examples which show peaking in the accuracy of Bayes classifiers, it has been shown that under some minor restrictions, the average accuracy of Bayes classifiers, designed on all possible generations, based on a given densities, namely Π_is, will never degrades, as p increases. The interrelationship between p and N_i and PCC has been studied based on Monte Carlo simulations by several researchers. It has been pointed out that an LDF can have better accuracy on a $U_{p'} \subset U_p$. For a constant p, to achieve a given accuracy, N_i should increase linearly for an LDF and Quadratic for a QDF. Numerical results, have a general agreement with an acceptable recommendation about a good practice in pattern recognition design which considers a linear relationship between p and the required N_i. That is, the number of training samples must be five to ten times the number of measurements [4], [5], i.e. $N/p > 5 \sim 10$. To use numerical results, e.g. results obtained by Monte Carlo simulations, two necessary conditions are the independency of the measurements and the knowledge of the true underlying distributions. Generally, the less the knowledge of the underlying distributions is, the more the required training samples for a given dimensionality would be. The amount of correlation within the measurements will change the optimal measurement complexity.

The formulation of Linear Discriminant Analysis and its relationship to Bayes Classifier

There are two main families of criteria to measure overlap of populations and separability of classes:

Chernoff distance, Bhattacharayya distance, etc, [5], [6], [9], [10], belong to a family of criteria which relates to upper bound of Bayes error for a given problem. Chernoff or Bhattacharayya distances as a criterion have two major drawbacks; they are applicable just for two class and normally distributed problems. However, optimizing of this family of criteria does not give Discriminant Functions as solutions.

The Bayes error is an optimum criterion for feature evaluation; however, it is not tractable in most practical cases. Therefore, in practice, to find Discriminant Functions, a simpler and more analytical criterion is desired. We can use a family of more tractable criterion functions, see next section, that depend on μ_i, Σ_i, and provide some measure of class seperability but have no direct relationship to Bayes error. However, it has been proven that [8] solving this family of criterion functions is equivalent to the mean-square-error of the general Bayes error estimate.

The most successful solutions to the FR or object recognition problems seem to be appearance-based approaches. They normally operate directly on images or 2-D appearances of the 3-D face objects. In appearance-based methods, one usually represents a 2-D $r \times c$ array of pixels by a vector in an n-dimensional space where $n = r.c$. These n-dimensional vectors are too large to be directly utilized in any practical and efficient face recognition system. Let the vector $z_i^j \in R^n$ denote the i-th sample (image) of class j. Give c, N_j, and N as the number of classes, samples in class j, and total number of training samples, respectively. Suppose μ and μ_j stand for the mean of all classes and class j respectively. Let us define:

$$S_W = \sum_{j=1}^{c} \sum_{i=1}^{N_j} (z_i^j - \mu_j)(z_i^j - \mu_j)^T \quad (5)$$

$$S_B = \sum_{j=1}^{c} N_j (\mu_j - \mu)(\mu_j - \mu)^T \quad (6)$$

$$S_T = \sum_{j=1}^{c} \sum_{i=1}^{N_j} (z_i^j - \mu)(z_i^j - \mu)^T = S_W + S_B \quad (7)$$

as the within class, between class, and total scatter matrices, respectively. PCA seeks to find:

$$W_{PCA} = \underset{W}{argmax} |W^T S_T W| \quad (8)$$

The most common case in DA applications is the multiclass problem, which usually utilizes LDA as in [5], [6], [9], [10], and is based on maximization of the following form of Fisher criterion [5], [6], [9], [10]:

$$J(W) = trace(W^T S_W^{-1} W W^T S_B W) \quad (9)$$

or find:

$$W = \underset{W}{argmax} \frac{|W^T S_B W|}{|W^T S_W W|} = [w_1\ w_2\ w_3\ \dots\ w_{c-1}] \quad (10)$$

where $\{w_i \mid i = 1, 2, 3, \dots, c-1\}$ is the set of $c-1$ generalized eigenvectors corresponding to the $c-1$ nonzero generalized eigenvalues $\{\lambda_i \mid i = 1, 2, 3, \dots, c-1\}$ of the generalized eigenvalue problems:

$$S_B w_i = \lambda_i S_W w_i \quad (11)$$

To solve (11), if S_W is a full rank matrix, then (11) can be rewritten as the following conventional eigenvalue problem:

$$S_W^{-1} S_B w_i = \lambda_i w_i \quad (12)$$

Based on aforementioned paragraphs, LDA is a plug-in Bayes classifier [4], [5]. LDA also could be seen as an approximation of aPAC. They usually yield similar results [3]. Thus, aPAC could be considered as a plug-in Bayes classifier, too. However, our proposed mechanism in Section 3 can be generalized to other DA algorithms, which use maximum likelihood estimations of their parameters, to achieve optimal complexity.

Hughes Phenomenon and Linear Discriminant Analysis

Usually, in practical classification problems (like FR or object recognition problems), designers don't have knowledge of underlying distributions and the parameters of classifiers are not known. They are forced to estimate classifiers parameters, e.g. μ, μ_j, S_T, S_B, S_W, etc, from available finite size training sets and utilize them in classifiers. The substitution of the maximum likelihood estimation of the parameters instead of the true values of them is the reason for the degradation of the plug-in versions [4], [5] of Bayes classifiers and their suboptimal performances. When someone involves in the design of a plug-in Bayes classifier one should follow an acceptable recommendation about a good practice in pattern recognition design to avoid Hughes phenomenon. That is, the number of training samples must be five to ten times the number of measurements [4], [5], i.e. $N/p > 5{\sim}10$. This is sometimes called the optimal complexity (or the optimal number of measurements) [5] for a classifier. Following the recommendation helps us to eliminate the effects of errors caused by the substitution of estimated parameters instead of true ones.

Unfortunately, sometimes concepts like 'the curse of finite sample size' or 'optimal measurements complexity' may be forgotten by researchers. We will show how the optimal complexity can be achieved by a simple mechanism based on PCA. *We use PCA as a filter for pruning non-informative features from original features spaces before applying to linear DA algorithms.* Finding optimal complexity can significantly improve the accuracies of DA algorithms. After achieving optimal complexity in Section 4, the classical LDA formulation and aPAC as two examples of linear DAs outperform the most superior DAs. This is predictable, since as we noted previously, a classifier based on (1) should be optimal [8].

The Optimal Complexity for Discriminant Analysis

PCA representation of data is the most compact representation of the data with no data structure distortions. Thus, PCA is used by researchers to obtain the lowest dimensional representation of data in small sample size problems. Those problems have $\mathrm{N} \ll \mathrm{n}$. Within class scatter matrices e.g. (5) are often utilized in DA algorithms in addition to LDA. A usual challenge in DA algorithms is the singularity challenge of within class scatter matrices, i.e. these matrices are noninvertible. In addition to finding the most compact representation of data, researchers usually use PCA as a tool for overcoming the singularity challenges in DA algorithms as well. Therefore, their criterion to determine how many principal components should be kept is the rank of within class scatter matrices. The singularity of S_W, causes a common computational challenge especially for a laboratory face database. Representing the data in a new p-dimensional subspace with $p < N - c$ may be a mathematical solution for this problem since the maximum rank of the within class scatter matrix is $N - c$. However, it is possible that the rank of S_W be smaller than $N - c$. Thus, to completely avoid rank deficiency problems, the rank of S_W should be calculated before projecting data to a p-dimensional subspace with $p < N - c$. Let W_{PCA}, be a matrix made by concatenating those eigenvectors of S_T that correspond to the $N - c$ largest eigenvalues. Researchers usually map data by means of W_{PCA} to a $N - c$-dimensional Principal Component Analysis (PCA) subspace, as the preprocessing subspace and then implement their DA algorithms. The maximum theoretical rank of these matrices is $N - c$, where c is the number of classes. The criterion of researchers for ordering the principal components is the amount of their corresponding eigenvalues. Thus, they sort the components with respect to their descending order of their eigenvalues and then keep those with the $N - c$ largest eigenvalues. Usually, S_W in the new coordinates is a full rank matrix and therefore their solution helps us to overcome the aforementioned computational problem caused by rank deficiency of S_W, however, it is not a good idea to cause LDA to become optimal. A perfectly reliable sorted subset of independent measurements are obtained from a random ordered set of correlated measurements, namely the face image pixels, by means of W_{PC}, before utilizing LDF. Therefore, the recommendation for a good practice in the pattern classification, i.e. holding $N/p > 5{\sim}10$, can be applied to the resultant subsets. We suggest to keep $p = p_{Opt} \approx N/10 \sim N/5$ instead of $p = N - c$. By using $\mathrm{p_{Opt}}$, the rank deficiency problem of the within class scatter matrix will be avoided more powerfully when compared with the case of using $p = N - c$. Since the computational and storage cost of linear DAs are of the order $O(N^2 p_{org})$ and $O(N p_{org})$, thus their decreases are of the order of $1/10 \sim 1/5$. These modifications results in a significant improvement in the accuracy of linear DAs, based on empirical results, and simultaneously decreasing the cost. Experiments show that the optimum p is almost among $N/10 \sim N/5$. Mathematical determination of the optimum p as a constant portion of N, seems to be so complicated and a tuning process is inevitable for a new unknown database to get an exceptional accuracy improvements. However, in the absence of such tuning, the accuracy improves when a designer follows a rule of thumb like limiting p to $N/10 \sim N/5$.

In this paper, we propose two optimum complexity versions of LDA and aPAC. They are based on features obtained by principal components corresponding to the $\lfloor N/10 \rfloor$ and $\lfloor N/5 \rfloor$ number of the largest eigenvalues and are shown by indices "opt1" and "opt2", respectively. This paper experiments could be seen as a proof to the optimality of LDA in minimization of Bayes error (see the 15th chapter of [5]), of coarse when the Hughes phenomenon is eliminated. A pseudo code for "opt1" version of LDA can be as follows. For "opt2" version LDA, one should only substitute p_2 instead of p_1, in the second step of the pseudo code. For aPAC, steps 4 and 5 should be substituted by the algorithm stated in [3]:

Input: an available sample collection and z_i^h, i-th sample belongs to h-th class; have to be classified in R^n

Output: g, classifier recognized class of z_i^h

Calculate $\mathrm{S_T}$ from (7)

Compute W_{PCA} from (4) and sort them according to the descending order of their corresponding eigenvalues

$p_1 = \lfloor N/10 \rfloor$ and $p_2 = \lfloor N/5 \rfloor$

$$V \longleftarrow W_{PCA}(1{:}p_1)$$

$$z'^h_i \longleftarrow V^T z_i^h$$

Calculate S_W and S_B for z'^h_is (from (5) and (6))

Obtain $W = [w_1\, w_2\, w_3 \ldots w_{c-1}]$ (from equations (11) or (12) and (10))

$$z''^h_i \longleftarrow W^T z'^h_i$$

Experimental results

We use MATLAB platform for simulation. We show results in Tables 2, 3, 4, and 5 when 1-Nearest Neighbor (1-NN), 5-NN, Nearest Mean (NM), and Support Vector Machine (SVM) have been used for classification, respectively. We use "Hold out" method with cross

validation factor 0.5. To see the effectiveness of finding p_{Opt}, seven popular publicly accessible databases in the face recognition area: FERET [12], [13], ORL [15], UMIST [16], GEORGIA TECH [17], Essex 94 [18], and Essex 95 [19] are used. Also, ETH-80 [14] has been used as a benchmark database for object recognition. Fundamental parameters of these databases are summarized in Table 1. In addition to popularity of FR datasets, we used these databases with different challenging factors in the area of FR, e.g. Pose, Facial expression, Image orientation, etc, for a face classifier. For FERET, we use images of its *b-subset*. This subset consists of the images whose names are marked with two character strings: "ba", "bj", "bk", "be", "bf", "bd", and "bg". For computational consideration, the images of FERET, ETH-80, Essex 94, and Essex 95 are resized to 96×64, 64×64, 50×40, and 50×40 respectively.

Table 1. Important information of this paper used databases

DATABASE	*Class*	*Images per Class*	*N*	*n*
FERET	*200*	*7*	*1400*	*98304*
UMIST	*20*	*between 19 to 48*	*575*	*10304*
Essex 94	*152*	*20*	*3040*	*36000*
Essex 95	*72*	*20*	*1440*	*36000*
ORL	*40*	*10*	*400*	*10304*
ETH-80	*8*	*410*	*3280*	*16384*
GEORGIA TECH	*50*	*15*	*750*	*Various*

Georgia Tech database requires a special attention regarding to variable resolution of its images. Since we initially need to get equal size vectors from these images, we forced images to occupy a fixed size 2-D array of pixels. This task was done by symmetric cropping of images with respect to the centre of each image. Hence, the top, down, right, and left margins of images are eliminated as needed. Since images in Georgia Tech database have at least 111×111 pixels, we only keep a symmetrical 111×111 region around the centre of each image.

Table 2: recognition rate (%) for nearest neighbor

DATABASE	*LDA_{opt1}*	*LDA_{opt2}*	*$aPAC_{opt1}$*	*$aPAC_{opt2}$*	*SDA*
FERET	***90.7****	***86.7***	***86.43***	***86.93***	*79.73*
Essex 94	***99.16***	***98.82***	*97.65*	***98.65***	***98.44***
Essex 95	***89.14***	***88.56***	***88.14***	***90.72****	***88.06***
UMIST	***99.02***	***99.23****	*98.11*	***99.16***	*97.82*
ORL	*93*	*95.4*	*90.9*	*91.3*	*92.13*
GEORGIA TECH	***70.74****	***69.71***	***68.29***	***70.63***	***70.46***
ETH-80	***74.7***	***75.68***	***74.66***	***75.23***	***84.39***
DATABASE	***CKFDP***	***CKFDG***	***LDA***	***aPAC***	***PCA***
FERET	*77.73*	***87.57***	*51.4*	*51.53*	*77.87*
Essex 94	***98.27***	***99.47****	*95.31*	*97.6*	*96.78*
Essex 95	***90.03***	***90.53***	*71.75*	*72.5*	*84.22*
UMIST	*23.86*	*67.16*	*98.32*	*78.88*	*94.39*
ORL	***96****	*46.7*	*91.9*	*94.3*	*89.7*
GEORGIA TECH	*67.89*	*46.86*	*33.94*	*59.09*	***68.63***
ETH-80	*66.54*	***72.84***	*65.83*	*65.85*	***86.5****

We implement our mechanism on LDA and aPAC and propose LDA_{opt1}, LDA_{opt2}, $aPAC_{opt1}$, and $aPAC_{opt2}$. They are based on features obtained by principal components corresponding to the $\lfloor N/10 \rfloor$ and $\lfloor N/5 \rfloor$ number of the largest eigenvalues and are shown by indices "opt1" and "opt2", respectively. We compare the resultant "opt1" and "opt2" versions with the most superior DAs like CKFD [7] with both polynomial and Gaussian kernel type (optimized by using 5-fold cross validation and shown by CKFDP and CKFDG, respectively), SDA [1], aPAC [3], LDA, and PCA. In order to increase the confidence level and statistical reliability of our results and to eliminate the effects of exceptions, the accuracies of the methods in Fig. 1 of this work are evaluated by averaging its performance over a large number of iterations. Higher accuracies in each row are bold face. The best accuracy in each row is shown by asterisk.

Table 3: recognition rate (%) for 5-nearest neighbor

DATABASE	*LDA_{opt1}*	*LDA_{opt2}*	*$aPAC_{opt1}$*	*$aPAC_{opt2}$*	*SDA*
FERET	***89.83****	***85.83***	***83.77***	***84.43***	*75.2*
Essex 94	***98.71***	***98.5***	*95.49*	***97.83***	***97.23***
Essex 95	*80.64*	***83.61***	*79.69*	***86.22****	*78.89*
UMIST	*96.7*	***98.88****	***98.11***	***97.82***	*92.63*
ORL	*89.5*	***91.7***	*79.5*	*80.5*	*86.75*
GEORGIA TECH	***69.03****	***66.97***	*63.94*	***67.94***	***68.86***
ETH-80	***76.77***	***76.87***	***76.84***	***76.77***	***81.07***
DATABASE	***CKFDP***	***CKFDG***	***LDA***	***aPAC***	***PCA***
FERET	*72.87*	***86.8***	*53.8*	*44.93*	*72.77*
Essex 94	***97.56***	***99.37****	*94.72*	*95*	*93.32*
Essex 95	***82.28***	***83.56***	*71.86*	*64.11*	*73.64*
UMIST	*21.33*	*65.26*	***98.32***	*67.93*	*83.51*
ORL	***95.9****	*45.7*	***91.9***	***92.6***	*78.6*
GEORGIA TECH	***67.37***	*45.03*	*35.83*	*55.66*	*64.63*
ETH-80	*67.35*	***74.28***	*66.27*	*66.2*	***83.06****

Table 4: recognition rate (%) for nearest mean

DATABASE	*LDA_{opt1}*	*LDA_{opt2}*	*$aPAC_{opt1}$*	*$aPAC_{opt2}$*	*SDA*
FERET	***92.93****	***91.6***	*86.37*	*86.77*	*71*
Essex 94	***99***	***98.66***	***97.09***	***98.49***	***98.31***
Essex 95	***85.97****	***85.47***	*64.58*	***85.28***	*76.56*
UMIST	***97.4***	***98.46****	*91.3*	***97.47***	*92.63*
ORL	***91.5***	***93.7***	*76.6*	*77.8*	***89.88***
GEORGIA TECH	***72.29****	***71.89***	*52*	*67.49*	*63.43*
ETH-80	***75.09***	***76.48****	***75.09***	***76.35***	*72.18*
DATABASE	***CKFDP***	***CKFDG***	***LDA***	***aPAC***	***PCA***
FERET	*76.67*	***91.63***	*60.1*	*18.47*	*41.7*
Essex 94	*97.98*	***99.42****	*94.99*	*96.91*	*96.1*
Essex 95	***84.64***	***83.61***	*71.83*	*31.58*	*50.67*
UMIST	*9.19*	*61.68*	***98.32***	*33.4*	*94.88*
ORL	***95.8****	*46.4*	***92***	***91.9***	*77.1*
GEORGIA TECH	***70.06***	*44.34*	*32.8*	*41.77*	*48.91*
ETH-80	*62.9*	*69.61*	*66.61*	*66.29*	*61.23*

Table 5: recognition rate (%) for SVM

DATABASE	*LDA_{opt1}*	*LDA_{opt2}*	*$aPAC_{opt1}$*	*$aPAC_{opt2}$*	*SDA*
FERET	***85.57****	***85.5***	***82.77***	***84,53***	*30.9*
Essex 94	***96.76***	***97.67***	*71.09*	*85.5*	*87.05*
Essex 95	***82.42***	***81.78***	***80.97***	***83.06****	*58.67*
UMIST	*95.37*	***97.54****	*95.09*	***97.4***	*91.23*
ORL	*86.1*	***95.5***	*86.7*	***93.8***	*91.75*
GEORGIA TECH	*51.26*	***57.71***	*47.54*	***58.34****	*32.4*
ETH-80	***59.57***	***66.39***	*52.9*	***60.2***	*64.35*
DATABASE	***CKFDP***	***CKFDG***	***LDA***	***aPAC***	***PCA***
FERET	*76.87*	***84.6***	*67.87*	*23.6*	*44*
Essex 94	*91.74*	***98.01****	***96.43***	*83.08*	*77.41*
Essex 95	***82.42***	***82.36***	*71.06*	*46.69*	***81.25***
UMIST	*19.44*	*67.02*	***97.19***	*91.02*	***96.7***
ORL	***94.9***	*39.9*	*92.8*	***95.7****	*93.4*
GEORGIA TECH	***57.94***	*34.57*	*39.77*	*39.03*	*40.29*
ETH-80	*41.41*	*58.48*	*59.77*	*60.22*	***73.6****

Discussion

LDA_{opt1} is the best method when 1-NN, 5-NN, and NM are used; an experimental demonstration for the claim of [8], i.e. a classifier based on (1) should be optimal. When SVM is used $aPAC_{opt2}$ is the best. As it can be seen, both versions of LDA and aPAC outperform some of the most

superior methods in the area of DA like SDA, CKFDG, CKFDP, and aPAC. The performance of the "opt1" and "opt2" versions of LDA and aPAC is incredibly superior than those methods especially in FERET and ORL. Also, in all of the used databases, the performance of the "opt1" and "opt2" versions is better than those methods. SDA has significantly larger training time than others in FERET and Essex 95. The reasons in those cases are the special definition of between class scatter matrix in SDA. SDA requires also a search for the optimal number of subclasses which requires significantly additional training time. However, in addition to achieving to much better accuracies for LDA and aPAC, tremendous computational and storage cost savings are achieved because of working with much smaller matrices and vectors. The optimum complexity versions of LDA and aPAC have much lower training times than SDA (as well as CKFDG and CKFDP). For FERET, Essex 94, Essex 95, and Georgia Tech, these versions are 1.5×10^3, 15, 5×10^3 and 25 times faster than SDA! There exists a reported vulnerability of LDA versus PCA due to insufficient or unsuitable training data per class in face recognition world. This vulnerability is faded away perfectly where PCA achieves to almost 70% and 78% for Georgia Tech and FERET, respectively. In contrast to LDA and aPAC, their optimized versions achieve to clearer peak point in the accuracy curves of 1-NN and NM when the number of features is between $\lfloor(c-1)/4\rfloor$ to $\lfloor(c-1)/2\rfloor$. Both of these benefits are worthwhile regarding to feature evaluation [8] concerns and required online classification times. However, we forsake these optimum values and report the accuracies of the ends of accuracy curves. Since for ORL and Georgia Tech $\lfloor N/10\rfloor < c-1$, thus the features numbers of LDA_{opt1} in these databases are limited to $\lfloor N/10\rfloor$. LDA_{opt1} and LDA_{opt2} have almost the same performances in UMIST and Georgia Tech.

LDA_{opt1} achieves to highest accuracy in FERET and Georgia Tech with only 60 and 35 features, respectively. Other methods have twice or triple number of features in these databases. This is much worthwhile regarding to required *online* classification times. The less the number of features is the less the required online classification times will be. Also, the modified versions have a clearer peak point in the accuracy curve than others. This is another worthwhile outcome of following the optimal complexity rule for a classifier. Having a clearer peak in the accuracy of a classifier for feature numbers less than $c-1$, is favorable regarding to *feature evaluation* concerns. The designer of a classifier always seeks to find the minimum number of features which yields to most accuracy simultaneously. The designer desires to find a way to sort the features in an order in which the most valuable features are in the first. One should note the reduction in the computational errors e.g. in necessary eigen-decomposition in the framework of LDA. The reduction is because of working by much smaller matrices during eigen-decompositions.

It should be noted that, as experiments show for the used databases, even for a very small p, i.e. $N/20\sim N/10$, the accuracy of the classifier is still better than that obtained from LDA and aPAC. Thus, $N/20\sim N/10$, is better suggestion than $N-c$ as the dimension of PCA representations before application of DA algorithms. In those cases, the storage cost will decrease in the order of $1/20\sim1/10$ and a significant efficiency improvement will be achieved.

Finally, our two suggestions for performance improvement of plug-in classifiers are as follows:

Considering the curse of finite sample size, we suggest following optimal measurement complexity recommendation with regard to the sample size, in the design of a plug-in classifier, as a plug-in Bayes classifier. Holding the number of measurements between $N/10\sim N/5$, in addition to significant accuracy improvement, usually decreases computational and storage costs. This is due to the fact that DA techniques often work on a much larger number of measurements. As stated above, W_{PCA} is one of the best means to utilize this idea. Instead of working on all available raw measurements of existing samples, we suggest working on $N/10\sim N/5$ number of good measurements. These are initially extracted from all available measurements, e.g. by means of W_{PCA}. Then apply DA techniques to the optimal number of extracted measurements.

Since mathematical determination of p_{Opt} is very complicated, instead of rough estimation of p_{Opt} like $N/10\sim N/5$, which is based on i.i.d. normal distribution assumption, to achieve the best accuracy, a tuning process seems to be inevitable for each individual database, since its distribution can only be reasonably approximated to normal distribution. Our experiments show that p_{Opt} and the overall shape of the accuracy curves for each of four tried databases is almost similar in different iterations. This is rational, since in determining p_{Opt} for a given database, the particular underlying distribution of that database plays the main role. Thus, we suggest offline tuning of the classifier just once and find p_{Opt} by stepwise increase of p. Empirical results show that for a usual FR or object recognition problem, a good suggestion for the step size, start, and stop points of the tuning process may be $N/10$, $N/10$, and $2N/5$, respectively. Therefore, the offline tuning computational costs due to our tuning mechanism is at most 0.3 times of that of the original method, i.e. LDA, aPAC, etc. To achieve the probable slight improvement in the classifier accuracy we can decrease the step size and increase the span of tuning. However, a reasonable theoretical stop point for the tuning process is $N-c$.

Conclusion

A mechanism has been proposed to eliminate Hughes phenomenon from linear DA algorithms. It has been shown that using the mechanism yields a significant improvement in the accuracy of used DAs while their computational and storage costs are reduced. Also, the usual vulnerability of LDA based algorithms versus PCA has been removed by utilizing the mechanism. The mechanism results in achieving the optimal complexity for linear DAs. It has been shown by intensive experiments that following the recommendation yields significant improvement in DA algorithms accuracies. In addition to accuracy improvement, the modification decreases their computational and storage costs and makes the overall method more efficient. The mechanism causes LDA and aPAC to outperform the state of the art and most superior linear and nonlinear DAs. This could be a very good demonstration for the strength of the idea of finding the optimal complexity for a classifier.

REFERENCES

[1] Zhu, M., and Martinez, A. M.: 'Subclass discriminant analysis," IEEE Trans. on Pattern Analysis and Machine Intelligence', 2006, 28, (8), pp. 1274-1286.

[2] Loog, M., and Duin, R.P.W.: 'Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion', IEEE Trans. on Pattern Analysis and Machine Intelligence, 2004, 26, (6), pp. 732-739.

[3] Loog, M., Duin, R.P.W., and Haeb-Umbach, T.: 'Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria', IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001, 23, (7), pp. 762-766.

[4] Jain, A. K., Duin, R. P.W., and Mao, J.: 'Statistical Pattern Recognition: A Review', IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000, 22, (1), pp. 4-37.
[5] Chandrasekaran, B., and Jain, A. K.: 'Dimensionality and Sample Size Considerations in Pattern Recognition Practice', Krishnaiah, P.R., and Kanal, L.N., (Eds.): 'Handbook of Statistics, Vol. 2' (Amsterdam: North-Holland, 1982), pp. 835-855
[6] Fukunaga, K.: 'Introduction to Statistical Pattern Recognition' (Academic Press, 1990, 2nd edn.).
[7] Yang, J., Frangi, A. F., Yang, J., Zhang, D., and Jin, Z. : 'KPCA plus LDA: A complete kernel fisher discriminant framework for feature extraction and recognition', IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27, (2), pp. 230-244.
[8] Fukunaga, K., and Short, R. D.: 'A class of feature extraction criteria and its relation to the Bayes risk estimate', IEEE Trans. on Inform. Theory, 1978, IT-26, pp. 59-65.
[9] Alpaydin, E.: 'Introduction to Machine Learning' (MIT Press, 2004).
[10] Duda, R., Hart, P., and Stork, D.: 'Pattern Classification' (John Wiley & Sons, 2001).
[11] Hughes, G. F.: 'On the mean accuracy of statistical pattern recognizers', IEEE Trans. Inform. Theory, 1968, 14, pp. 55-63.
[12] Georghiades, A.S., Belhumeur, P.N., and Kriegman, D.J.: 'From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose', IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001, 23, (6), pp. 643-660.
[13] Lee, K.C., Ho, J., and Kriegman, D.: 'Acquiring Linear Subspaces for Face Recognition under Variable Lighting', IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27, (5), pp. 684-698.
[14] Phillips, P.J.: 'The Facial Recognition Technology (FERET) Database', http://www.itl.nist.gov/iad/humanid/feret/feret_master.html, 2004.
[15] Phillips, P.J., Moon, H., Rizvi, S.A., and Rauss, P.J.: 'The FERET Evaluation Methodology for Face-Recognition Algorithms', IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000, 22, (10), p. 1090-1104.
[16] Leibe, B., and Schiele, B.: 'Analyzing Appearance and Contour Based Methods for Object Categorization', In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2003.
[17] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html, accessed August 2010.
[18] http://www.shef.ac.uk/eee/research/vie/research/face.html, accessed August 2010.
[19] http://www.anefian.com/research/face_reco.htm, accessed August 2010.
[20] http://cswww.essex.ac.uk/mv/allfaces/faces94.html, accessed August 2010.
[21] http://cswww.essex.ac.uk/mv/allfaces/faces95.html, accessed August 2010.

***Authors**: Aboozar Hosseinzadeh, Amirkabir Universisty of Technology, Electrical Engineering Dept., Tehran, Iran;*
Address: 4th floor, no.31, Shabnam 2 street, Karaj, Iran.
Tel: 00982612515636, Fax: 00982612515636
Email: aboozar@aut.ac.ir;
Dr. Majid Noorhosseini, Amirkabir Universisty of Technology, Computer Engineering Dept., Tehran, Iran; Email: majidnh@aut.ac.ir.