Using a Maximum Uncertainty LDA-Based Approach to Classify and Analyse MR Brain Images

Carlos E. Thomaz[1], James P. Boardman[2], Derek L.G. Hill[3], Jo V. Hajnal[2], David D. Edwards[4], Mary A. Rutherford[2], Duncan F. Gillies[1], and Daniel Rueckert[1]

[1] Department of Computing, Imperial College, London, UK
cet@doc.ic.ac.uk
[2] Imaging Sciences Department, Imperial College, London, UK
[3] Division of Imaging Sciences, King's College, London, UK
[4] Division of Paediatrics, Obstetrics and Gynaecology, Imperial College, London, UK

Abstract. Multivariate statistical learning techniques that analyse all voxels simultaneously have been used to classify and describe MR brain images. Most of these techniques have overcome the difficulty of dealing with the inherent high dimensionality of 3D brain image data by using pre-processed segmented images or a number of specific features. However, an intuitive way of mapping the classification results back into the original image domain for further interpretation remains challenging. In this paper, we introduce the idea of using Principal Components Analysis (PCA) plus the maximum uncertainty Linear Discriminant Analysis (LDA) based approach to classify and analyse magnetic resonance (MR) images of the brain. It avoids the computation costs inherent in commonly used optimisation processes, resulting in a simple and efficient implementation for the maximisation and interpretation of the Fisher's classification results. In order to demonstrate the effectiveness of the approach, we have used two MR brain data sets. The first contains images of 17 schizophrenic patients and 5 controls, and the second is composed of brain images of 12 preterm infants at term equivalent age and 12 term controls. The results indicate that the two-stage linear classifier not only makes clear the statistical differences between the control and patient samples, but also provides a simple method of analysing the results for further medical research.

1 Introduction

Multivariate pattern recognition methods have been used to classify and describe morphological and anatomical structures of MR brain images [5, 8, 17]. Most of these approaches analyse all voxels simultaneously, and are based on statistical learning techniques applied to either segmented images or a number of features preselected from specific image decomposition approaches. Although such preprocessing strategies have overcome the difficulty of dealing with the inherent high dimensionality of 3D brain image data, an intuitive way of mapping the classification results back into the original image domain for further interpretation has remained an issue.

C. Barillot, D.R. Haynor, and P. Hellier (Eds.): MICCAI 2004, LNCS 3216, pp. 291–300, 2004.

In this paper, we describe a new framework for classifying and analysing MR brain images. It is essentially a linear two-stage dimensionality reduction classifier. First the MR brain images from the original vector space are projected to a lower dimensional space using the well-known PCA and then a maximum uncertainty LDA-based approach is applied next to find the best linear discriminant features on that PCA subspace. The proposed LDA method is based on the maximum entropy covariance selection method developed to improve quadratic classification performance on limited sample size problems [15].

In order to demonstrate the effectiveness of the approach, we have used two MR brain data sets. The first contains images of 17 schizophrenic patients and 5 controls; and the second is composed of brain images of 12 preterm infants at term equivalent age (mean post-menstrual age [PMA] at birth 29 weeks, mean PMA at MR image acquisition 41 weeks), and 12 term born controls (mean PMA at birth 40.57 weeks, mean time of image acquisition day 4 of postnatal life). The results indicate that the two-stage linear classifier not only makes clear the statistical differences between the control and patient samples, but also provides a simple method of analysing the results for further medical research.

2 Principal Component Analysis (PCA)

Principal Component Analysis has been used successfully as an intermediate dimensionality reduction step in several image recognition problems. It is a feature extraction procedure concerned with explaining the covariance structure of a set of variables through a small number of linear combinations of these variables. In other words, PCA generates a set of orthonomal basis vectors, known as principal components (or most expressive features [12]), that minimizes the mean square reconstruction error and describes major variations in the whole training set considered [4]. For this representation to have good generalisation ability and make sense in classification problems, we assume implicitly that the distributions of each class are separated by their corresponding mean differences.

However, there is always the question of how many principal components to retain in order to reduce the dimensionality of the original training sample. Although there is no definitive answer to this question for general classifiers, Yang and Yang [16] have proved recently that the number of principal components to retain for a best LDA performance should be equal to the rank of the total covariance matrix composed of all the training patterns.

3 Linear Discriminant Analysis (LDA)

The primary purpose of the Linear Discriminant Analysis is to separate samples of distinct groups by maximising their between-class separability while minimising their within-class variability.

Let the between-class scatter matrix S_b and within-class scatter matrix S_w be defined as

$$S_b = \sum_{i=1}^{g} N_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \quad \text{and} \quad S_w = \sum_{i=1}^{g}\sum_{j=1}^{N_i}(x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T , \tag{1}$$

where $x_{i,j}$ is the n-dimensional pattern j from class π_i, N_i is the number of training patterns from class π_i, and g is the total number of classes or groups. The vector $\bar{x}_i$ and matrix S_i are respectively the unbiased sample mean and sample covariance matrix of class π_i [4]. The grand mean vector $\bar{x}$ is given by

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{g} N_i\bar{x}_i = \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{N_i} x_{i,j} , \tag{2}$$

where N is the total number of samples, that is, $N = N_1 + N_2 + \cdots + N_g$. It is important to note that the within-class scatter matrix S_w defined in (1) is essentially the standard pooled covariance matrix multiplied by the scalar $(N-g)$, that is

$$S_w = \sum_{i=1}^{g}(N_i - 1)S_i = (N-g)S_p . \tag{3}$$

The main objective of LDA is to find a projection matrix P_{lda} that maximizes the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix (Fisher's criterion), that is

$$P_{lda} = \arg\max_{P} \frac{\left| P^T S_b P \right|}{\left| P^T S_w P \right|} . \tag{4}$$

It has been proved [4] that if S_w is a non-singular matrix then the Fisher's criterion is maximised when the column vectors of the projection matrix P_{lda} are the eigenvectors of $S_w^{-1}S_b$ with at most $g-1$ nonzero corresponding eigenvalues. This is the standard LDA procedure.

However, the performance of the standard LDA can be seriously degraded if there are only a limited number of total training observations N compared to the dimension of the feature space n. Since the within-class scatter matrix S_w is a function of $(N-g)$ or less linearly independent vectors, its rank is $(N-g)$ or less. Therefore, S_w is a singular matrix if N is less than $(n+g)$, or, analogously, might be unstable if N is not at least five to ten times $(n+g)$ [7].

4 The Maximum Uncertainty LDA-Based Approach

In order to avoid the singularity and instability critical issues of the within-class scatter matrix S_w when LDA is applied in limited sample and high dimensional problems, we have proposed a new LDA approach based on a straightforward stabilisation method for the S_w matrix [14].

4.1 Related Methods

In the past, a number of researchers [1, 2, 9, 10] have proposed a modification in LDA that makes the problem mathematically feasible and increases the LDA stability when the within-class scatter matrix S_w has small or zero eigenvalues.

The idea is to replace the pooled covariance matrix S_p of the scatter matrix S_w (equation (3)) with a ridge-like covariance estimate of the form

$$\widehat{S}_p(k) = S_p + kI\,, \tag{5}$$

where I is the n by n identity matrix and $k \geq 0$. However, a combination of S_p and a multiple of the identity matrix I as described in equation (5) expands all the S_p eigenvalues, independently of whether these eigenvalues are either null, small, or even large [14].

Other researchers have imposed regularisation methods to overcome the singularity and instability in sample based covariance estimation, especially to improve the Bayes Plug-in classification performance [3, 6, 13]. According to these regularisation methods, the ill posed or poorly estimated S_p could be replaced with a convex combination matrix $\widehat{S}_p(\gamma)$ of the form

$$\widehat{S}_p(\gamma) = (1-\gamma)S_p + (\gamma)\overline{\lambda}I\,, \tag{6}$$

where the shrinkage parameter γ takes on values $0 \leq \gamma \leq 1$ and could be selected to maximise the leave-one-out classification accuracy. The identity matrix multiplier would be given by the average eigenvalue $\overline{\lambda}$ of S_p calculated as

$$\overline{\lambda} = \frac{1}{n}\sum_{j=1}^{n}\lambda_j = \frac{tr(S_p)}{n}\,, \tag{7}$$

where the notation "tr" denotes the trace of a matrix.

The regularisation idea described in equation (6) has the effect of decreasing the larger eigenvalues and increasing the smaller ones, thereby counteracting the biasing inherent in sample-based estimation of eigenvalues [3]. However, such approach would be computationally expensive to be used in practice because it requires the

calculation of the eigenvalues and eigenvectors of an n by n matrix for each training observation of all the classes in order to find the best mixing parameter γ.

4.2 The Proposed Method

The proposed method considers the issue of stabilising the S_p estimate with a multiple of the identity matrix by selecting the largest dispersions regarding the S_p average eigenvalue.

Since the estimation errors of the non-dominant or small eigenvalues are much greater than those of the dominant or large eigenvalues [4], we have used the following selection algorithm [14] to expand only the smaller and consequently less reliable eigenvalues of S_p, and keep most of its larger eigenvalues unchanged:

i. Find the Φ eigenvectors and Λ eigenvalues of S_p, where $S_p = S_w/[N-g]$;

ii. Calculate the S_p average eigenvalue $\overline{\lambda}$ using equation (7);

iii. Form a new matrix of eigenvalues based on the following largest dispersion values

$$\Lambda^* = diag[\max(\lambda_1, \overline{\lambda}), \max(\lambda_2, \overline{\lambda}), ..., \max(\lambda_n, \overline{\lambda})]; \tag{8a}$$

iv. Form the modified within-class scatter matrix

$$S_w^* = S_p^*(N-g) = (\Phi \Lambda^* \Phi^T)(N-g). \tag{8b}$$

The proposed LDA is constructed by replacing S_w with S_w^* in the Fisher's criterion formula described in equation (4). It is a straightforward method that overcomes both the singularity and instability of the within-class scatter matrix S_w when LDA is applied in small sample and high dimensional problems. It also avoids the computational costs inherent to the aforementioned shrinkage processes.

5 Experimental Results

In order to demonstrate the effectiveness of the approach, we have used two MR brain data sets. The first contains images of 17 schizophrenic patients and 5 controls and the second is composed of brain images of 12 preterm infants at term equivalent age and 12 term controls.

Before registration, the MR brain images of each subject in the schizophrenia dataset have been stripped of extra-cranial tissue [18]. All MR images are then registered to the MNI brainweb atlas using the non-rigid registration described in [11]. Finally, the registered images have been resampled with a voxel size of 1x1x1 mm^3. After registration and resampling the schizophrenia dataset consists of 181x217x181 = 7,109,137 voxels. An analogous procedure has been adopted for the infant images consisting of 256x256x114 = 7,471,104 voxels.

Throughout all the experiments, we have used for the schizophrenia analysis a total of 13 training examples, i.e. all the 5 controls and 8 schizophrenia patients randomly selected. The remaining 9 schizophrenia subjects have been used for classification testing. The infant training and test sets have been composed of 8 and 4 images respectively. For instance, the PCA schizophrenia transformation matrix is composed of 7,109,137 (= number of voxels) rows and 12 (= total of training samples – 1) columns, where each column corresponds to a principal component. The LDA transformation matrix is composed of 12 rows and 1 (= number of groups – 1) column. Using these PCA plus LDA transformation matrices, every original image composed of 7,109,137 voxels has been reduced to a one-dimension vector on the final (or most discriminant [12]) feature space.

Figures 1 and 2 show the projected schizophrenia and infant data on the most expressive and discriminant features. White circles and squares represent the training sample of the controls and schizophrenia (or infant) examples used. The black circles and squares describe the corresponding subjects selected for testing. As can be seen, although the two and three most expressive features explain more than 50% of the total sample variance, the classification superiority of the two-stage dimensionality reduction technique based on a maximum uncertainty LDA approach is clear in both applications.

Another result revealed by these experiments is related to the linear discriminant feature found by the maximum uncertainty approach. In fact, this one-dimensional vector corresponds to a hyper-plane on the original space which direction describes statistically the most discriminant differences between the controls and the patients images used for training. A procedure of moving along this most discriminant feature and mapping back into the image domain might provide an intuitive interpretation of the results.

Figures 3 and 4 show respectively the schizophrenia and infant five points chosen from left to right on the most discriminant feature space and projected back into the image domain using the corresponding transpose of the LDA and PCA linear transformations previously computed. In Figure 3, although the differences are very subtle, the visual analysis of the example sagittal, axial, and coronal slices suggests that the regions of the normal brains are slightly better represented than the ones observed on the schizophrenia images. In looking at the preterm analysis illustrated in Figure 4, there is enlargement of the lateral ventricular system in the preterm infants at term equivalent age compared to the term control group. This is a common finding at term equivalent age among infants who have been born prematurely [19,20].

6 Conclusion

In this paper, we introduced the idea of using PCA plus the maximum uncertainty LDA-based approach to classify and analyse MR brain images. It avoids the computation costs inherent in the commonly used optimisation processes, resulting in a simple and efficient implementation for the maximisation and interpretation of the Fisher's criterion.

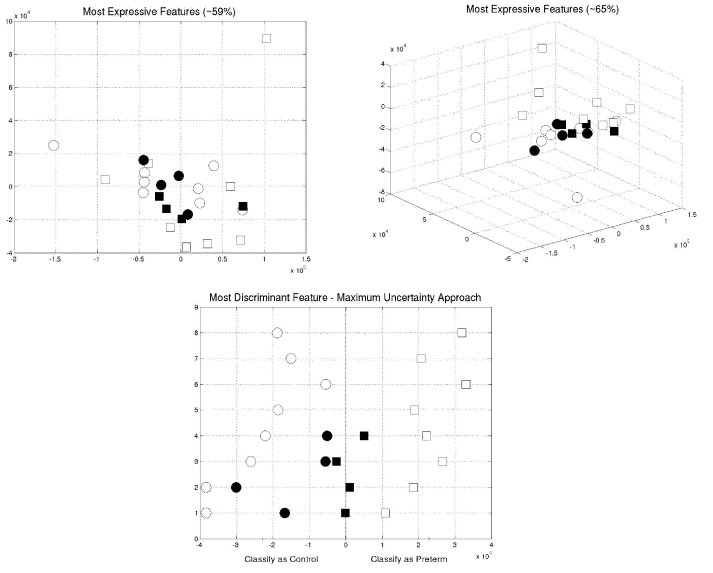

Fig. 1. Schizophrenia sample data projected on the most expressive and discriminant* features.

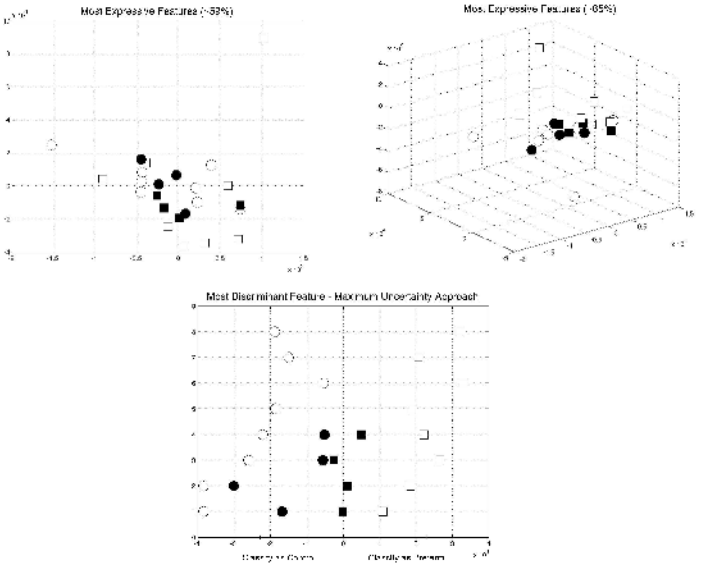

Fig. 2. Infant sample data projected on the most expressive and discriminant* features.

* The vertical value of each point is illustrative only and represents its corresponding index in the sample set.

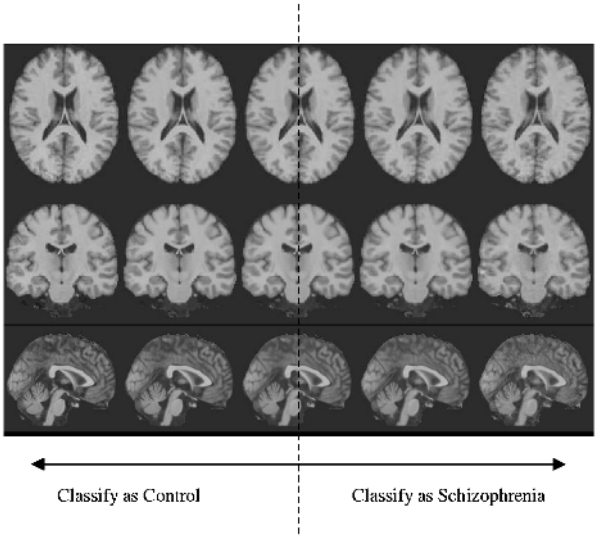

Fig. 3. Visual analysis of the schizophrenia most discriminant feature.

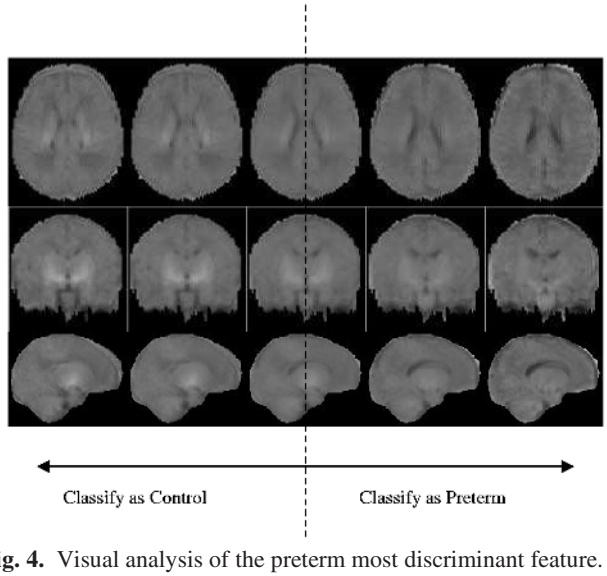

Fig. 4. Visual analysis of the preterm most discriminant feature.

The two-stage dimensionality reduction technique is a straightforward approach that considers the issue of stabilising the ill posed or poorly estimated within-class scatter matrix with a multiple of the identity matrix. Although the experiments carried out were based on small MR data sets, we believe that such recent multivariate statistical advances for targeting limited sample and high dimensional problems can

provide a new framework of characterising and analysing the high complexity of MR brain images.

Acknowledgments. This work is part of the UK EPSRC e-science project "Information eXtraction from Images" (IXI). Also, the first author was partially supported by the Brazilian Government Agency CAPES under Grant 1168/99-1.

References

1. N.A. Campbell, "Shrunken estimator in discriminant and canonical variate analysis", *Applied Statistics*, vol. 29, pp. 5-14, 1980.
2. P.J. Di Pillo, "Biased Discriminant Analysis: Evaluation of the optimum probability of misclassification", *Comm. in Statistics-Theory and Methods*, vol. A8, no. 14, pp. 1447-1457, 1979.
3. J.H. Friedman, "Reguralized Discriminant Analysis", *JASA*, vol. 84, no. 405, pp. 165-175, 1989.
4. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, 1990.
5. P. Golland, W. Grimson, M. Shenton, and R. Kikinis, "Small Sample Size Learning for Shape Analysis of Anatomical Structures", In Proc. *MICCAI*, pp. 72-82, 2000.
6. T. Greene and W.S. Rayens, "Covariance pooling and stabilization for classification", *Computational Statistics & Data Analysis*, vol. 11, pp. 17-42, 1991.
7. A. K. Jain and B. Chandrasekaran, "Dimensionality and Sample Size Considerations in Pattern Recognition Practice", *Handbook of Statistics*, vol. 2, pp. 835-855, 1982.
8. Z. Lao, D. Shen, Z. Xue, B. Karacali, S. Resnick, and C. Davatzikos, "Morphological classification of brains via high-dimensional shape transformations and machine learning methods", *NeuroImage*, vol. 21, pp. 46-57, 2004.
9. R. Peck and J. Van Ness, "The use of shrinkage estimators in linear discriminant analysis", *IEEE PAMI*, vol. 4, no. 5, pp. 531-537, September 1982.
10. W.S. Rayens, "A Role for Covariance Stabilization in the Construction of the Classical Mixture Surface", *Journal of Chemometrics*, vol. 4, pp. 159-169, 1990.
11. D. Rueckert, A. F. Frangi, and J. A. Schnabel, "Automatic Construction of 3-D Statistical Deformation Models of the Brain Using Nonrigid Registration", *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 1014-1025, 2003.
12. D. L. Swets and J. J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval", *IEEE PAMI*, vol. 18, no. 8, pp. 831-836, 1996.
13. S. Tadjudin, "Classification of High Dimensional Data With Limited Training Samples", PhD thesis, Purdue University, West Lafayette, Indiana, 1998.
14. C. E. Thomaz and D. F. Gillies, "A Maximum Uncertainty LDA-based approach for Limited Sample Size problems - with application to Face Recognition", *Technical Report TR-2004-01*, Department of Computing, Imperial College, London, UK, January 2004.
15. C. E. Thomaz, D. F. Gillies and R. Q. Feitosa. "A New Covariance Estimate for Bayesian Classifiers in Biometric Recognition", *IEEE CSVT*, vol. 14, no. 2, pp. 214-223, February 2004.
16. J. Yang and J. Yang, "Why can LDA be performed in PCA transformed space? ", *Pattern Recognition*, vol. 36, pp. 563-566, 2003.

17. P. Yushkevich, S. Joshi, S.M. Pizer, J.G. Csernansky and L.E. Wang. "Feature Selection for Shape-Based Classification of Biological Objects", *Information Processing in Medical Imaging,* 2003.
18. S. Smith, "Fast robust automated brain extraction",*Human Brain Mapping*,17(3):143-155, 2002.
19. L. R. Ment et al. "The etiology and outcome of cerebral ventriculomegaly at term in very low birth weight preterm infants", *Pediatrics* 104, 243-248, 1999.
20. J. P. Boardman et al. "An evaluation of deformation-based morphometry applied to the developing human brain and detection of volumetric changes associated with preterm birth", In Proc. *MICCAI*, Lecture Notes in Computer Science (2878), 697-704, 2003.