Classification of SPECT Images of Normal Subjects versus Images of Alzheimer's Disease Patients

Jonathan Stoeckel[1], Grégoire Malandain[1], Octave Migneco[2,3], Pierre Malick Koulibaly[3], Philippe Robert[4], Nicholas Ayache[1], and Jacques Darcourt[2,3]

[1] EPIDAURE - Project, INRIA,
2004 route des Lucioles - 06 902 Sophia Antipolis, France
{Jonathan.Stoeckel,Gregoire.Malandain,Nicholas.Ayache}@sophia.inria.fr
http://www-sop.inria.fr/epidaure/
[2] Service de Médecine Nucléaire - Centre Antoine Lacassagne
33 avenue de Valombrose - 06 189 NICE cedex 2, France
[3] Laboratoire de Biophysique - Faculté de Médecine
28 avenue de Valombrose - 06 107 NICE cedex 2, France
[4] Service de Psychiatrie - CHU Pasteur
30, voie Romaine, B.P 69 - 06 000 NICE cedex, France

Abstract. This work aims at providing a tool to assist the interpretation of SPECT images for the diagnosis of Alzheimer's Disease (AD). Our approach is to test classifiers, which uses the intensity values of the images, without any prior information. Such a classifier is built upon a training set, containing images with two different labels (AD patients and normal subjects). It will then provide a classification for any new unknown image. The main problem to be handled is the small number of available images compared to the large number of features (here the image's voxels): the so-called *small sample size* problem. We evaluate here the ability of two linear classifiers to correctly label a set of 79 images. Our experiments show promising results. They also show that image classification based on intensity values only is possible and might be used for other applications as well.

Clinical Context: Alzheimer's Disease

Alzheimer's disease (AD) is a neuro-degenerative disease that produces among others memory loss, behavioural changes and cognitive impairment. Mainly elderly are affected by this disease. Because of the aging populations the number of patients will probably rise in the coming years.

Single photon emission computed tomography (SPECT) is being largely used for the study of cerebral blood flow (CBF). These studies provide unique information for the identification of functional abnormalities relevant to Alzheimer's disease.

The process of diagnosing AD is based on a qualitative evaluation of neuropsychologic tests, combined with the analysis of SPECT images in case of serious

W. Niessen and M. Viergever (Eds.): MICCAI 2001, LNCS 2208, pp. 666–674, 2001.

suspicion. The aim of this work is to provide a tool which will assist the interpretation of these images and, we hope, clarify the ambiguous cases. This is especially important in the early stages of the disease, when the patient can benefit most from drugs that may have an impact on the progression of the disease, and to be able to support the development of such drugs.

1 Introduction

In the recent past a lot of research has been done on comparing *groups* to find the areas where differences exist in the regional cerebral blood flow between groups of AD and normal subjects. These results show on average significant abnormalities in the parietotemporal regions between AD an normal subjects [1]. But these typical patterns do not occur within all patients with probable AD [2]. The majority of these results were obtained using techniques based on statistical parametric mapping (SPM) [3] or on principal component analysis (PCA)/singular value decomposition (SVD) [4,5,6]. However these methods only give us a tool to find significant differences between groups of images. They were not developed to give us information about individual subjects.

However, in this article we will explore the possibility to develop a classifier that can classify a *single* SPECT image as being an image of a normal subject or of a probable Alzheimer patient. As described in the previous section, defining clear features on which a classification can be based is extremely hard. Therefore we choose not to use any prior knowledge about AD in SPECT images. We directly use the intensity values of the image as input to our classifier.

Yet, the number of available images to train such a classifier compared to the number of voxels in the image is very small. This most often leads to very poor classification performance. It is called the *small sample size* problem (see section 2.2) in the pattern recognition literature [7].

In this article we will present and test two classifiers that are able to circumvent the small sample size problem. Herewith we will show that it is possible to classify 3D images only using the intensity values. In our particular case this provides us with a valuable tool for assisting the interpretation of SPECT images for Alzheimer's Disease diagnosis.

In the following section we first introduce the notations used throughout this article and introduce briefly the concept of classification. This is followed by a description and analysis of the small sample size problem (section 2.2). In section 2.3 two classifiers well adapted to our problem are presented.

Results based on the data described in section 3.1 are shown in section 3.3. Finally we discuss our approach and give some directions for further research.

2 Methods

2.1 Classifiers

As pointed out in the introduction, we are looking for a classifier that will assign one of the two possible class labels (AD or normal) to an image given as an

input. This can be written as a function $g(x)$, which returns a positive value for one class and a negative value for the other class. x is a feature vector of length n describing the object we want to classify. Our images (objects) are described by simply putting the values of the n voxels in the feature vector. The objects can thus be seen as points in $\mathbb{R}^n$ (the feature space), and $g(x)$ as implicitly defining a surface in $\mathbb{R}^n$ discriminating the two classes.

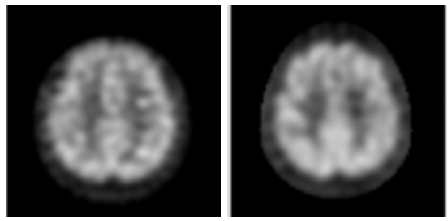

Fig. 1. *Example of a normal (left) and an AD (right) image shows the difficulty of visual interpretation.*

To define this surface a classifier needs to be trained. The training method is provided with a training set consisting of m example objects and their class labels. Once the classifier has been trained, it can be tested by comparing its results on a test set with the known labels of each object of this set. To make an unbiased estimation of the error rate of the classifier, the test set and the training set should be totally independent. In the ideal case, both sets are created by taking randomly objects out of the N objects with known class labels which form the total available object population. As might be expected the accuracy of the classification error depends not only on the independence of the test set but also on its size.

A classifier is said to have a good generalisation capability if it performs well on previously unknown objects. In this case, it does not model the noise of the learning set but the real structure of the data with respect to the class differences.

In practice the number of available objects is often too small to select large enough independent training *and* test sets. A number of alternative methods have been proposed for estimating the classification error [8]. The most well known method is the *re-substitution* method: both the training set and the test set contain exactly the same objects and this method is therefore optimistically biased. This measure does not say anything at all about the generalisation capabilities of a classifier. In this work, we will use the *leave one out* error estimate as described in section 3.2.

2.2 Small Sample Size Problem

Until recently it was believed that for classifiers trained with a number of training objects m smaller or around the dimensionality n of the feature space, no generalisation capability could be expected. This was based on the idea that the feature space needs to be *well filled* (curse of dimensionality). It principally came from the statistical point of view of needing many samples for being able to get good distribution estimates (see preface of [9]).

Logically it would seem that adding features to objects, and thus having relatively more information for classification of the objects, should lead to better results. The fact that it is often not the case, therefore leading to the small sample

size problem, might be explained by the idea that a higher dimensional feature space provides a classifier with more degrees of freedom, which can then exactly model the training set. In this case, it does mainly model the objects of the learning set and its associated noise but not the real structure of the data.

One way to solve the small sample size problem is to reduce the number of features. On the contrary, modern approaches, that are aimed at finding the inherent structure of the data instead of the parameters of an a-priori distribution, seem to be able to overcome the small sample size problem. An excellent overview of this problem and several related classifiers are found in [7]. Some classifiers which are proposed in the literature are the nearest mean classifier (NMC) [10], several variants of the Fisher Linear Discriminant (FLD) [11,12], and even highly nonlinear classifiers like the Parzen classifier [13] or the Support Vector Classifier (SVC) [9].

2.3 Nearest Mean Classifier (NMC) and Pseudo Fisher Linear Discriminant (PFLD)

Because of the extremely high number of features compared to the number of available objects we chose to use classifiers with a low complexity. This leads to linear discriminant classifiers, in other words classifiers defining one single hyperplane in feature space which separates the two classes. The most basic approach is the nearest mean classifier (NMC), it classifies the object to the nearest class mean:

$$g_{\mathrm{NMC}}(x) = \left(x - \bar{x}^{(2)}\right)^T \left(x - \bar{x}^{(2)}\right) - \left(x - \bar{x}^{(1)}\right)^T \left(x - \bar{x}^{(1)}\right) \tag{1}$$

where x is the object to be classified, $\bar{x}^{(1)}$ and $\bar{x}^{(2)}$ are the means of the feature vectors in the training set for the classes one and two respectively. It spans up an equidistant hyperplane between both class means.

The probably most used classification rule is the Fisher Linear Discriminant (FLD), it not only takes the class means into account but also the sample covariance S (assumed to be common to both classes):

$$g_{\mathrm{FLD}}(x) = \left[x - \frac{1}{2}\left(\bar{x}^{(1)} + \bar{x}^{(2)}\right)\right]^T S^{-1} \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right) \tag{2}$$

Please note that the NMC is a special case of the FLD, when S is the identity matrix the two classifiers are equivalent (Eq.(1)=Eq.(2) when $S = \lambda I$).

When the number of training samples is smaller than the number of features ($m < n$) the covariance matrix yields a singular matrix that cannot be inverted. The Pseudo Fisher Linear Discriminant (PFLD) [11] is formed by replacing the inverse of the covariance matrix by its pseudo inverse, $(S^T S)^{-1} S^T$. The pseudo inverse relies on the singular value decomposition of S and it becomes equal to the inverse of S when $m \geq n$. In this case, the PFLD is equivalent to the FLD.

Even though the FLD is seen in the classical way as designed for two multivariate Gaussian populations with equal prior probabilities differing in mean

vectors, but sharing the same covariance matrix, the PFLD has also a more general geometric interpretation [14]. It defines a hyper-plane that maximises the distances to all given training objects. If $m < n$ all the training objects are in a linear subspace, and the discriminant is perpendicular to that subspace. This makes sense as it corresponds with an indifference to directions for which no training objects are given.

The PFLD is known to have very bad generalisation capabilities when $m \approx n$ [11]. In this application, we assume this problem will not play a role because we have $m \ll n$.

2.4 Preprocessing

This section describes the necessary pre-processing steps before using the images as input to a classifier: the two first points consists in spatial and intensity normalisation, while the third one is only necessary from a computational point of view.

Spatial Normalisation: Registration Due to different positions of the subjects in the scanner as well as differences of brain size and morphology between subjects, the same voxel location in two images may not correspond to the same anatomical position. Therefore image registration is necessary. Because global intensity variations exist (see below) a robust approach is needed. We searched for the affine transformations which maximise the correlation coefficient [15] using Powell's optimisation method. We have also tested a non-rigid [16] approach, but this did not improve classifier performance.

Intensity Normalisation A basic problem in the use of SPECT HMPAO images is the lack of an absolute signal level. There are several sources of variation of the measures signal among SPECT images. *Global* variations: they may be caused by differences in the dose of the radioactive tracer being present at the time of the image acquisition, scanner sensitivity, or the positioning of the patient with respect to the detectors. *Local* variations: differences occur normally between different subjects as well as within the same subject over time. Of course pathologies as AD do cause differences as well. Before analysis of the images global variations have to be corrected whereas local variations, especially those due to a pathology, should be preserved. The standard approach is to assume the global CBF to be the same for all subjects. This leads to simply dividing the intensities by the sum of the intensities in the brain. This approach is not very robust in the case of strong local variations. Saxena [17] proposed to divide by the mean of the top one percent intensities. Here, the assumption is that the voxels with the highest intensities (denoting highest levels of perfusion) would be in those areas that are relatively unaffected by AD. In our application choosing either one of these two methods did not influence the classification results.

Subsampling by Mean Filtering All images were subsampled by a certain factor in each dimension by simply taking the mean of the implicated voxels. This was done for the following reasons: firstly to obtain acceptable memory and calculation time requirements; secondly to eliminate noise and compensate for eventual registration imprecision. Misregistration can lead to two voxels at the same position for different subjects not to represent exactly the same anatomical position; this effect may be suppressed by the averaging of the voxels. Of course only the image voxels being part of the brain were used for the classification.

3 Results

3.1 Materials

A set of 29 images of probable AD patients (mean Mini Mental Test score 23.5), with clinically confirmed diagnosis were acquired as well as a set of 50 images of normal subjects. This last set of images was made available by The Society of Nuclear Medicine Brain Imaging Council (`http://brainscans.med.yale.edu/`). The rCBF was assessed with technetium-99m-D L-hexamethyl-propylene amine oxime (Tc-99m HMPAO), a tracer that is trapped inside the brain in proportion to the regional blood flow. All images were acquired using triple-head camera systems (Picker Prism 3000 series). The cameras were equipped with fan beam collimators.

Raw projection images were acquired in a 128×128 matrix. They were reconstructed using a ramp filter and the resulting slices post-filtered using a low-pass filter (Butterworth, order=6, cutoff=0.26). Chang attenuation correction [18] (cutoff=0.07) was applied to the filtered data. After registration and reslicing all images had 1.8 mm cubic voxels and 93 axial slices.

3.2 Leave One Out Method

A widely used classification error estimate when the number of objects is very small, is the *leave one out* method [8]. In this method all N available objects are used. The classifier is trained N times on a training set of $m = N - 1$ objects leaving out a different object each time. This object is used to test the classifier. This provides us with an unbiased error estimate if the observations of the objects are statistically independent.

3.3 Experiments

Table 1 show results using the 79 images described in section 3.1. The success rates were calculated using the leave one out method (see previous section). Note, that due to roundoff effects when selecting the voxels being part of the brain, the number of resulting features 4819 and 618 are not exactly 8 times more between subsampling factors 4 and 8. The results show a little improvement when using more features, probably some information is lost by too much subsampling. The

Table 1. Percentages of successfully classified images obtained using the leave one out method. Results are shown for all 79 images and for the two classes individually (AD = Alzheimer's Disease, NO = normal). Both the NMC and PFLD were tested for subsampling factors 4 and 8, corresponding respectively to 4819 and 618 features (i.e. voxels).

subsampling factor	NMC			PFLD		
	Total	AD	NO	Total	AD	NO
4	84.8%	79.3%	88.0%	89.9%	82.8%	94.0%
8	81.0%	72.4%	86.0%	88.6%	82.8%	92.0%

PFLD outperforms the NMC, this can be explained by the fact that it also takes the covariance into account, and thus the shape of the classes in feature space. The differences in success rates between normal and AD images might be explained by the presence of nearly two times more normal images than AD images. We also investigated if using less training objects would increase the performance of the PFLD. This was not the case and shows the assumption made (number of features is sufficently different from the number of training images) at the end of section 2.3 to hold for our data. Please note that four highly experimented observers, who were presented the 29 AD images amidst other images, classified on average only 66.4% of the AD images correctly (PFLD and subsampling factor 4 resulted in 82.8% for AD images, see table 1).

4 Discussion and Conclusion

In this article, we have presented a general framework for using classifiers directly on 3D volumetric images without using prior knowledge based on recent findings in the pattern recognition literature for the small sample size problem. Promising results for use in AD were shown in the previous section, especially when considering the difficulty human observers have in classifying this type of images. Of course feature extraction (e.g. symmetry features, texture features) may improve the results. But we have shown it feasible to do without. Our further research will explore the possibilities for feature extraction as well as using other well behaving classifiers in the small sample size problem. Another important point will be to compare the performance of our classifiers to that of human observers who do normally analyse this type of images. A classifier that not only gives a binary result but also a rating which indicates a degree of class-membership might provide an even more useful tool. It is important to notice that this type of classification methods could be applied to a lot of different applications in medical imaging. They can help to solve the problems related to the often complex diagnosis making in 3D volumetric images.

5 Acknowledgements

We would like to thank P. Cachier and A. Roche for fruitful discussions, and R.P.W. Duin for making PrTools [19] available. This work was partially supported by the Conseil Régional Provences-Alpes-Côte d'Azur and by a grant from the European Community *SPECT in Dementia BIOMED2*.

References

1. Jagust W.J. Functional imaging patterns in Alzheimer's disease. *Annals of the New York Academy of Sciences*, 777:30-36, 1996.
2. Nitrini R., Buchpiguel C.A., Caramelli P., Bahia V.S., Mathias S.C., Nascimento C.M.R., Degenszajn J., and Caixeta L. SPECT in Alzheimer's disease: features associated with bilateral parietotemporal hypoperfusion. *Acta Neurologica Scandinavica*, 101:172-176, 2000.
3. Frackowiak R.S.J., Friston K.J., Frith C.D., and Dolan R. *Human Brain Function*. Academic Press, 1997.
4. Jones K., Johnson K.A., Becker J.A., Spiers P.A., Albert M.S., and Holman B.L. Use of singular value decomposition to characterize age and gender differences in SPECT cerebral perfusion. *Journal of Nuclear Medicine*, 39:965-973, 1998.
5. Houston A.S., Kemp P.M., and Macleod M.A. A method for asseing the significance of abnormalities in HMPAO brain spect images. *Journal of Nuclear Medicine*, 35:239-244, 1994.
6. Johnson K.A., Jones K., Holman B.L, Becker J.A., Spiers P.A., Satlin A., and Albert M.S. Preclinical prediction of Alzheimer's disease using SPECT. *Neurology*, 50:1563-1571, 1998.
7. Duin R.P.W. Classifiers in almost empty spaces. In *Proceedings of the 15th International Conference on Pattern Recognition (ICPR2000), Vol. 2*, pages 1-7, Las Alamitos (CA), 2000. IAPR, IEEE Computer Society.
8. Raudys S.J. and Jain A.K. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 13(3):252-264, March 1991.
9. Vapnik V.N. *Statistical learning theory*. John Wiley & Sons, 1998.
10. Skurichina M. and Duin R.P.W. Stabilizing classifiers for very small sample sizes. In *Proceedings of the 13th International Conference on Pattern Recognition (ICPR1996), Vol. 2*, pages 891-896. IAPR, IEEE Computer Society, 1996.
11. Raudys S.J. and Duin R.P.W. Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters*, 19:385-392, 1998.
12. Chen L.F., Liao H.Y.M., Ko M.T., Lin J.C., and Yu G.J. A new LDA-based face recognition which can solve the small sample size problem. *Pattern Recognition*, 33:1713-1726, 2000.
13. Hamamoto Y., Fujimoto Y., and Tomita S. On the estimation of a covariance matrix in designing Parzen classifiers. *Pattern Recognition*, 29(10):1751-1759, 1996.
14. Duin R.P.W. Small Sample Size Generalization. In *Proceedings of the 9th Scandinavian Conference on Image Analysis (SCIA95)*, pages 957-964, 1995.
15. Roche A., Malandain G., and Ayache N. Unifying Maximum Likelihood Approaches in Medical Image Registration. *International Journal of Imaging Systems and Technology*, 11:71-80, 2000.

16. P. Cachier and X. Pennec. 3D Non-Rigid Registration by Gradient Descent on a Gaussian-Windowed Similarity Measure using Convolutions. In *Proc. of MMBIA'00*, pages 182-189, Hilton Head Island, USA, June 2000.
17. Saxena P., Pavel D.G., Quintana J.C., and Horwitz B. An automatic threshold-based scaling method for enhancing the usefulness of Tc-HMPAO SPECT in the diagnosis of Alzheimer's disease. In *Proceedings of the 1st international conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI'98)*, volume 1496 of *Lecture Notes in Computer Science*, pages 623-630. Springer, 1996.
18. Chang L.T. A method for attenuation correction in radionuclide computed tomography. *IEEE Transactions on Nuclear Science*, 25:638643, 1978.
19. Duin R.P.W. *PRTools 3.1, A Matlab Toolbox for Pattern Recognition*. Delft University of Technology, Januray 2000.