Ensemble Learning Incorporating Uncertain Registration

Ivor J.A. Simpson[1,2]
ivor.simpson@eng.ox.ac.uk
Julia A. Schnabel[1]
julia.schnabel@eng.ox.ac.uk
Jesper L.R. Andersson[2]
jesper@fmrib.ox.ac.uk
Adrian R. Groves[2]
adriang@fmrib.ox.ac.uk
Mark W. Woolrich[2,3]
mark.woolrich@ohba.ox.ac.uk

[1] Institute of Biomedical Engineering
Old Road Campus Research Building
Oxford OX3 7DQ

[2] FMRIB Centre
John Radcliffe Hospital
Oxford, OX39DU

[3] Oxford Centre for Human Brain Activity
Warneford Hospital
Oxford OX3 7JX

Abstract

In this paper we propose a novel approach for improving the robustness of quantitative prediction methods in spatially normalised statistical analysis of magnetic resonance (MR) images of the human brain. This is achieved by estimating the distribution of image registration mappings from subject to atlas space, rather than just using the *maximum-a-posteriori estimate*. As any derived predictions are highly dependent on the registration, the distribution of spatially normalised feature data can be derived from the set of probable mappings. This distribution of feature data can then be used to generate alternative training examples to create multiple predictors, which can then be combined in an ensemble learning approach. This allows for a more generalised prediction, which compensates for the inherent uncertainty in registration. Furthermore, extra testing examples can be generated to provide a measure of the prediction uncertainty. We demonstrate that using an ensemble learning approach with random feature data samples always leads to improvement in classification rate when separating subjects with Alzheimer's Disease from normal controls using a linear support vector machine.

1 Introduction

Medical imaging data is often used to make quantitative predictions about the current or future state of a subject. Machine learning techniques such as statistical classifiers and regressors are often used to facilitate this objective. Most of these machine learning techniques are *supervised*, which means they require a training set of data from which the methods learn about the relationship between the feature data $\mathbf{d}$, e.g. voxel intensities, and outcome variable o, e.g. disease score/group. When presented with unseen feature data, these methods should be able to predict the true value with a high degree of accuracy.

c 2012. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

Most standard machine learning algorithms require feature data of all subjects to be transformed into a common frame of reference to allow comparison. To map to this reference space, a registration tool is used to estimate the mapping between the subject image and atlas space (usually an average or representative subject). From this mapping, subject feature data can be transformed to the atlas space, a process known as spatial normalisation. This spatially normalised data can be then be examined in a voxel based morphometry (VBM) framework. Additionally, features of interest can be derived from the estimated transformation, a process known as deformation tensor based morphometry (TBM).

As has been previously shown, inter-subject brain registration is not an exact process [4]. Therefore the spatially normalised features, which are used as a basis for making predictions, are highly dependent on the registration procedure and are unlikely to be perfectly aligned. This misalignment of data will be a contributing factor to any incorrect predictions. The majority of registration methods simply estimate the *maximum-a-posteriori* (MAP), which is the most likely mapping subject to some constraints. However, recent registration methods have emerged which allow estimates of the registration uncertainty [6][8]. This facilitates the consideration of the set of probable mappings, as opposed to just the MAP.

We take an ensemble learning approach [2], where multiple statistical predictors are averaged. To create multiple classifiers we use a parametric variant of bootstrapping [3] to produce new examples of all the training subjects according to the distribution of the feature data. For spatially normalised feature data, the distribution of the feature data can be estimated based on the set of probable mapping inferred from the registration algorithm. Therefore, in this work, we randomly draw samples of feature data to train statistical predictors. This is repeated to generate an ensemble of predictive models. This ensemble accounts for the inherent registration uncertainty. Samples of the data distribution for test subjects also allows the estimation of the variability in prediction under uncertain registration.

In this work we propose an ensemble learning scheme using a parametric bootstrap approach to provide robust classification which accounts for the registration uncertainty. We can apply this approach to voxel, or deformation tensor based morphometry approaches to estimate differences between subject groups using any standard black box prediction tool. To demonstrate the effect of this method, we apply it to discriminating between subjects with Alzheimer's disease and age matched healthy controls using a standard black box classifier. In our experiments we find that using ensemble learning with data features sampled according to registration uncertainty leads to a consistent improvement in classification accuracy.

2 Methods

2.1 Probabilistic Registration Methods

A probabilistic registration method that can estimate posterior transformation distributions is required to estimate the distribution of the feature data, which accounts for the uncertainty in registration. Standard registration procedures use a MAP approach to infer the mapping between images. These approaches do not provide any estimates on the confidence of the inferred mapping, and consequently do not lend themselves to this work.

There are two published methods which we are aware of which allow the inference of a distribution of probable mappings: Risholm et al. [6] use MCMC to numerically estimate the full posterior distribution of transformation parameters whilst marginalising over the regularisation parameters. This method is computationally very expensive even for low degrees

of freedom. The alternative approach proposed by Simpson et al. [8] uses Variational Bayes (VB) to infer an approximate posterior distribution of the set of transformation, regularisation, and noise parameters. Although this method is more efficient, it uses the assumption that the true distribution of transformation parameters follows a multi-variate normal distribution, which so far has not been justified as necessarily being appropriate. Nevertheless, we choose to use the method of Simpson et al. due to the large computational benefits. We include a brief summary of their method here:

Image registration can be described using a generative model as given in equation

$$\mathbf{y} \quad \mathbf{t}\,\mathbf{x}\,\mathbf{w} \quad \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is the target image, $\mathbf{t}\,\mathbf{x}\,\mathbf{w}$ is the transformed source image $\mathbf{x}$ and $\mathbf{w}$ parameterises the transformation. $\mathbf{e}$ models the image mismatch where $\mathbf{e} \quad 0 \quad {}^{1}\mathbf{I}$, $\mathbf{I}$ is the matrix identity and $\quad$ is the global inverse variance. We use a Free-Form Deformation (FFD) transformation model [7], where $\mathbf{w}$ is the set of b-spline knot displacements. Priors are included on all our unknown model parameters. Most importantly, a prior on $\mathbf{w}$ is required to provide regularisation of the mapping, $P\ \mathbf{w} \quad 0; \quad {}^{1}$ where $\quad$ encodes the bending energy regularisation, and $\quad$ is an inferred parameter which controls the strength of the prior.

Posterior distributions are inferred using a set of iterative update equations applied until parameter convergence. Of particular interest is the posterior distribution of $\mathbf{w}$, $P\ \mathbf{w}\ \mathbf{y} \quad q\ \mathbf{w} \quad {}^{1}$. The update equations for the hyper-parameters are given as:

$$\bar{}\mathbf{J}^{T}\mathbf{J} \quad \bar{} \tag{2}$$

$$_{new} \quad \bar{}\mathbf{J}^{T}\ \mathbf{J}\ _{old} \quad \mathbf{y} \quad \mathbf{t}\,\mathbf{x}\ _{old} \tag{3}$$

where $_{old}$ is the previous mean estimate of the transformation parameters $\mathbf{w}$ and $\mathbf{J}$ is the matrix of first order partial derivatives of the transformation parameters with respect to $\mathbf{t}\,\mathbf{x}\ _{old}$. $\bar{}$ and $\bar{}$ are the expectation of the regularisation and image noise distributions, respectively. $\quad$ is a factor which models the correlation in the image mismatch.

$q\ \mathbf{w}$ is the approximate posterior distribution of the inferred transformation parameters. The shape and scale of this distribution is dependent on the structure of the **image information**, weighted by the **noise precision**, which indicates the level of image mismatch. It also depends on the form of the **spatial prior**, e.g. the bending energy, weighted by the **spatial precision** which is related to the similarity of the transformation to the spatial prior.

2.2 Using sampled mappings

Once the registration model parameters have been inferred, we can consider taking a voxel- or deformation tensor-based morphometry [1] approach to provide a framework for the classification of subjects into their respective groups.

In the standard setting, VBM is performed by transforming the subject data to the atlas space based on the expectation of the transformation distribution, $\mathbf{d} \quad \mathbf{t}\,\mathbf{x}$. In TBM, instead of examining the spatially normalised image information, the assumption is made that the discriminative differences between subjects are contained in the deformation field used to map each subject to the template image. Feature data is often derived from the voxelwise $3 \quad 3$ Jacobian matrix of the transformation of the mapping $\mathbf{J}_m$. Often a scalar measure of the Jacobian matrix is used for comparison, most commonly $\mathbf{d}_i \quad \log\ \mathbf{J}_{m\ i}$, where i is a voxel index. $\log\ \mathbf{J}_{m\ i}$ provides a measure of the expansion/contraction of a particular voxel as a

result of the mapping. The spatially normalised images can be compared between images in either a voxelwise or a multivariate fashion as features for classification. By sampling from the approximate posterior distribution of mappings $q(\mathbf{w})$ an estimation of the distribution $P(\mathbf{d})$, where $\mathbf{w}$, can be built up. The novel contribution of this work is to use this distribution of features to provide robust classification which properly accounts for the uncertainty in the estimated registration.

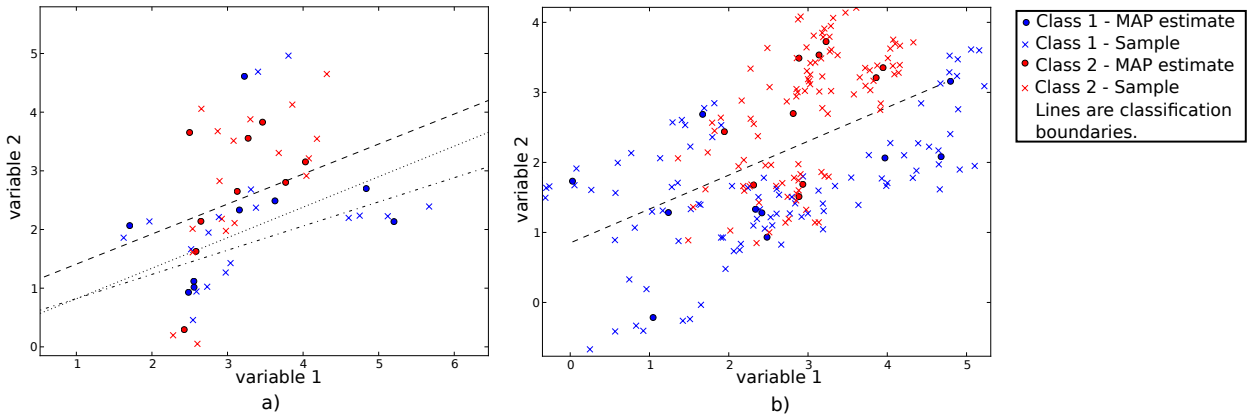

Figure 1: Graphical examples of using sampled data features in classification. a) illustrates ensemble learning with parametric bootstrapping (train+), each classification boundary is estimated using sampled data features from a fixed set of subjects. b) shows estimating variability in classification label using 10 random samples for each subject (test+).

Ensemble learning methods [2], are approaches which construct a set of multiple predictive models, and take an average of their predictions. In this case, we create several predictive models by drawing feature data samples for the set of training subjects in place of observations, in a parametric bootstrapping approach [3]. A graphical illustration of how multiple predictive models can be generated is illustrated in Figure 1 a). In this work we use an un-weighted averaging of predictions, but several more advanced approaches exist.

The predicted class variability can be estimated using a set of probable feature data for each subject, rather than the most likely observation. This is illustrated in Figure 1 b).

3 Experiments

162 subject images were taken from the ADNI database[5], 81 suffered from Alzheimer's disease (AD), the remaining 81 were age matched normal control subjects (NC). All the images were initially affinely aligned to the MNI 152 atlas space using 9 degrees of freedom, and resampled to 1mm isotropic resolution. To apply VBM or TBM, a suitable atlas image is required. An atlas image was created by iteratively registering 40 NC subjects to their average. This resulted in a sharp atlas image which is representative of the normal control population. Figure 2 illustrates how regions of interest were selected for left and right hippocampi using an enlarged bounding box.

The subject images were non-rigidly registered to the atlas image using a 5mm FFD knot spacing for TBM, and a 10mm spacing for VBM. Samples of probable mappings were drawn from , these mappings were used to create samples of feature data. From these samples a subject feature mean, and covariance matrix was estimated. To allow storage of the covariance matrices, the feature data was sub-sampled by a factor of 4 to 1800 voxels.

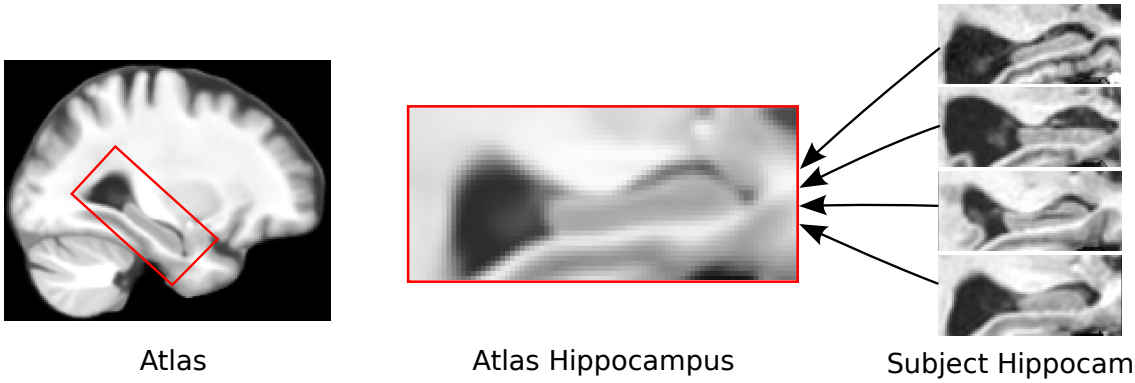

Figure 2: The experimental setup. A hippocampal bounding box of size (40x80x35 voxels), in red, selected the region of interest. Each subjects' hippocampi were registered to the atlas.

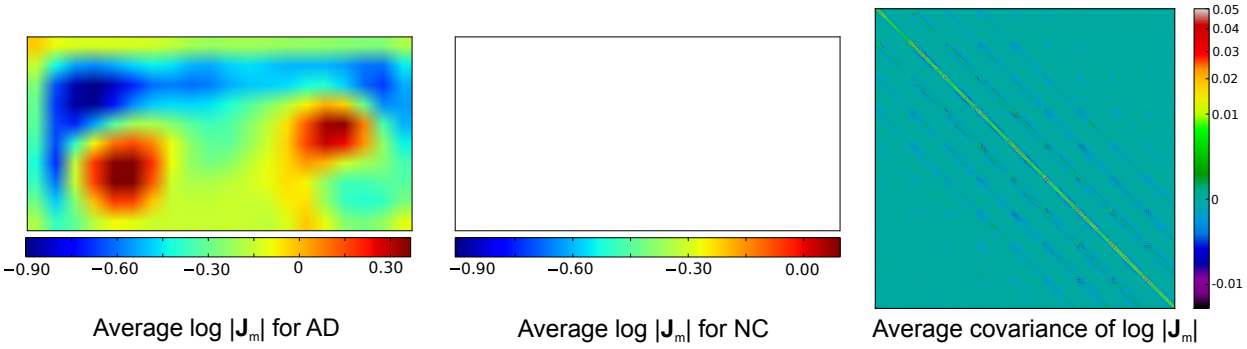

Figure 3: Left and middle images show the average log $\mathbf{J}_m$ inferred from registering the subject to the atlas images, as shown in Figure 2, for the AD and NC groups. The image on the right shows the average of the intra-subject log $\mathbf{J}_m$ covariance matrix as calculated based on the distribution of probable registration mappings.

The mean log $\mathbf{J}$ images for both groups and the covariance matrix are given in figure 3.

Based on the parametric model of feature data for each subject, new probable samples can be drawn, and utilised as described in section 2.2. In the experiments we use a linear support vector machine (SVM), using all 1800 feature voxels to classify between subject groups. Subject age is regressed out for each voxel. We compare 4 method variants: original: where simply the subject mean image data is used; train+: where ensemble learning with parametric bootstrap is used to create multiple classifiers; test+: where extra test samples are classified; and finally: train+test+ which combines train+ and test+. Experimental results are given in table 1. For computational efficiency, all samples are drawn using the subject average covariance matrix. We also test averaging predictions across different feature types.

As shown in table 1, using train+, or train+test+ appears to provide a marked increase in

Table 1: Classification correct rate using the different predictor training and testing variants in TBM and VBM. L and R indicate the left and right hippocampus data, respectively.

Feature data	Original	Train+	Test+	Train+ Test+
L log $\|\mathbf{J}\|$	0.704	**0.796**	0.697	0.790
R log $\|\mathbf{J}\|$	0.709	0.759	0.722	**0.772**
Average L & R log $\|\mathbf{J}\|$	0.747	0.809	0.772	**0.821**
L VBM	0.759	**0.790**	0.759	**0.790**
R VBM	0.734	0.753	0.734	**0.759**
Average L & R VBM	0.778	**0.802**	0.772	0.784
Average L & R log $\|\mathbf{J}\|$ & VBM	0.802	**0.827**	0.802	0.809

classification rate over using the estimated MAP in both VBM and TBM approaches. The improvement in TBM is more substantial as TBM features are entirely dependent on the mapping used, whereas in VBM the feature data is less dependent on the mapping. Finally, we can see that all variants of ensemble learning provide reasonable estimates of classification uncertainty as the results from averaging multiple feature data types are improved.

4 Discussion and Conclusions

In this work we have demonstrated that using random sampled observations of spatially normalised feature data in statistical prediction leads to more robust prediction. The random feature data samples were derived from the set of probable mappings between subject and atlas space as inferred by a probabilistic registration tool. These samples were incorporated into an ensemble learning framework. In our experiments we provide results on the problem of classification of subjects with Alzheimers Disease, from age matched healthy controls using a linear SVM. The results show a consistent improvement in classification rate when using random samples compared to MAP observations, with a maximum improvement of 9% for some image derived features.

Possible future work includes using feature selection instead of voxel sub-sampling and investigating alternative ensemble learning approaches which use weighted averaging.

References

[1] J. Ashburner and K.J. Friston. Voxel-based morphometry–the methods. *Neuroimage*, 11 (6):805–821, 2000.

[2] T. Dietterich. Ensemble methods in machine learning. *Multiple classifier systems*, pages 1–15, 2000.

[3] B. Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 7 (1):1–26, 1979.

[4] A. Klein and et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, 46(3):786–802, 2009.

[5] S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, C. Jack, W. Jagust, J.Q. Trojanowski, A.W. Toga, and L. Beckett. Alzheimer's Disease Neuroimaging Initiative. *Advances in Alzheimer's and Parkinson's Disease*, pages 183–189, 2008.

[6] P. Risholm, E. Samset, and W. Wells. Bayesian Estimation of Deformation and Elastic Parameters in Non-rigid Registration. *WBIR*, pages 104–115, 2010.

[7] D. Rueckert, LI Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach, and D.J. Hawkes. Nonrigid registration using Free-Form Deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[8] I.J.A. Simpson, J.A. Schnabel, A.R. Groves, J.L.R. Andersson, and M.W. Woolrich. Probabilistic inference of regularisation in non-rigid registration. *NeuroImage*, 59(3): 2438–2451, 2012.