CLASSIFICATION OF ALZHEIMER'S DISEASE PATIENTS AND CONTROLS WITH GAUSSIAN PROCESSES

Jonathan Young[1], Marc Modat[1], Manuel J Cardoso[1], John Ashburner[2] and Sebastien Ourselin[1,3]

[1]Centre for Medical Image Computing, University College London.
[2]Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London.
[3]Dementia Research Centre, Institute of Neurology, University College London.

ABSTRACT

There has been a great deal of recent work making use of multivariate classification techniques such as support vector machines to classify brain images obtained from MRI or PET as healthy or suffering from neurodegenerative disease. In the case of Alzheimer's disease, the results are as accurate as standard clinical tests and could potentially be used in a diagnostic setting. However these techniques give categorical class decisions. Here we show for the first time that Gaussian processes can be applied to structural neuroimaging data to perform classification of Alzheimer's disease subjects in a fully Bayesian framework. This offers advantages such as automatic setting of parameters via type II maximum likelihood and probabilistic predictions that may be useful in a clinical context, while maintaining the same accuracy as a state-of-the-art discriminative classifier applied to the same data.

Index Terms— Gaussian processes, Bayesian methods, Alzheimer's disease, Magnetic resonance imaging

1. INTRODUCTION

Alzheimer's disease (AD) is the most common form of dementia in the elderly, with prevalence expected to increase greatly in coming years largely due to aging population and possibly improvements in care of the elderly. A certain diagnosis can only be made by observing the presence of neurofibrillary plaques and tangles at autopsy. Hence in practice diagnosis is usually based on clinical history and batteries of cognitive tests. However, due to the unreliability, variability and somewhat subjective nature of these tests, biomarkers based on neuroimaging have been found to offer levels of accuracy at least as good.

A number of different imaging modalities have been used in these, including PET [1], and more recently both diffusion weighted and functional MRI [2], but the majority have used structural MRI [3-6]. This is because it is known that the early stages of AD are characterized by a pattern of atrophy in grey matter that is readily assessed and quantified on such images.

For example, the hippocampus is known to be one of the structures most affected by the disease process so the volume of the segmented hippocampus, normalized by intracranial volume, can distinguish between AD patients and controls with high accuracy [3]. Measurements of cortical thickness across the entire brain offer similar accuracy [4]. This study follows [5] in using maps of grey matter density across the entire brain as this approach includes all known areas of atrophy associated with AD.

Use of such images implies very high dimensionality in the data, as the dimensionality is equal to the number of voxels in the images being employed. Various methods have been used to cope with this, such as using complex feature extraction procedures [6] or well known methods such as principal components analysis to reduce the dimensionality of the images while attempting to preserve discriminative information, or selecting a subset consisting of the most discriminative features according to some statistical criterion. The most widely used classifier in these types of studies however is the support vector machine (SVM), which treats all the training images as points in a (potentially very) high dimensional space and attempts to find a hyperplane separating two labeled groups in the training data. It selects the hyperplane such that the distance to the closest training data on either side is maximized. Such classifiers can deal directly with images of the entire brain without having to reduce dimensionality.

However as discriminative classifiers, SVMs produce a simple decision value. Probabilistic predictions have a number of advantages. Firstly, they enable some clinically useful options such as a reject option for uncertain cases and use of decision theory to optimize classification rules. Furthermore, it is increasingly seen that the paradigm of evidence based medicine is best realized by Bayesian reasoning [7]. As a more practical consideration, in a probabilistic setting, classification parameters can be tuned via type-II maximum likelihood rather than computationally expensive cross-validation. Here we describe the Gaussian

process (GP) regression and classification method, which is fully Bayesian, and implement its first application to classification of AD from structural MRI images. We compare the performance with that of an SVM on the same data, and show that the former offers the advantages of probabilistic decisions while suffering no loss in classification accuracy.

2. MATERIALS AND METHODS

2.1 Images

The data used in this study consisted of 60 T1 weighted structural MRI images of healthy controls, and 60 T1 weighted scans of subjects diagnosed clinically with probable AD. All images were obtained from the ADNI database. The two groups were matched for age and sex. Each disease category was further split into 40 subjects for training the classifiers and 20 images to test them. The age and sex matching was preserved during this split.

2.2 Image Processing

To enable a classifier to be constructed using the training images, they had to be transformed into the same space. The images were masked to remove non brain material, and then used to perform groupwise registration. All images were repeatedly registered to a target image in an iterative procedure. At the end of each iteration, all registered images were averaged together to create an updated target image, with a randomly chosen image serving as the target in the first iteration. Initially, all images were rigidly registered to avoid bias to the target image. This was followed by a round of affine registrations, and then by 10 rounds of nonrigid registrations. All registrations were performed by NiftyReg [8], a registration toolkit that performs fast diffeomorphic nonrigid registrations.

The native space images were probabilistically segmented using the NiftySeg [9] tool into five tissue types: white matter, cortical grey matter, external cerebrospinal fluid, deep grey matter and internal cerebrospinal fluid. The transformations from each image's native space to the space of the final iteration groupwise template were then applied to the segmented native space images to warp the segmentations to the template space.

Finally, the segmentations were modulated by multiplying each voxel by the Jacobian determinant of the deformation field transforming it from its native space to the template space. This step ensures the total volume of tissue remains constant. Spatial smoothing of the image was not performed as part of this study.

2.3 Gaussian Process Regression

GP regression may be viewed in a number of different ways, but perhaps the simplest to introduce them is as, in the simplest case, a fully Bayesian extension of linear regression. Consider a set of target data $\mathbf{x}$, with observed scalar target values for regressing on $\mathbf{y}$. Assume there is zero mean Gaussian noise on the measurements. So if, as in linear regression, the model is parameterized by a set of weights $\mathbf{w}$, the target values $\mathbf{y}$ are normally distributed with mean given by the scalar product of the weights and data with observation noise variance σ^2.

$$y_i \sim N(\mathbf{w}^T x_i, \sigma^2)$$

Furthermore, assume a zero mean Gaussian prior on to weights $\mathbf{w}$ with diagonal covariance τ^2.

$$w \sim N(0, \tau^2 I)$$

Given observed data $D = \{X = [\mathbf{x}_1 ... \mathbf{x}_N], Y = [y_1 ... y_N]\}$ it can be shown that the posterior distribution on $\mathbf{w}$ also has Gaussian form.

$$\mathbf{w} \mid D \sim N(\mu_{\mathbf{w}}, \Sigma_{\mathbf{w}})$$

where $\Sigma_{\mathbf{w}} = \left(\sigma^{-2} XX^T + \tau^{-2} I\right)^{-1}$ and $\mu_{\mathbf{w}} = \sigma^{-2} \Sigma_{\mathbf{w}} XY^{\mathbf{T}}$

2.4 Making Predictions

To predict the target value y_{N+1} for a new data point $\mathbf{x}_{N+1}$, integrate over the posterior for the weights

$$\begin{aligned} p(y_{N+1} \mid \mathbf{x}_{N+1}, D) &= \int p(y_{N+1} \mid \mathbf{w}, \mathbf{x}_{N+1}) p(\mathbf{w} \mid D) d\mathbf{w} \\ &= \int N(y_{N+1} \mid \mathbf{w}^{\mathbf{T}} \mathbf{x}_{N+1}, \sigma^2) N(\mathbf{w} \mid \mu_{\mathbf{w}}, \Sigma_{\mathbf{w}}) d\mathbf{w} \\ &= N\left(\mu_{\mathbf{w}}^{\mathbf{T}} \mathbf{x}_{N+1}, \mathbf{x}_{N+1}^{\mathbf{T}} \Sigma_{\mathbf{w}} \mathbf{x}_{N+1} + \sigma^2\right) \end{aligned}$$

or alternatively we can take the mean of the posterior as a MAP estimate of $\mathbf{w}$.

2.5 Classification with Gaussian Processes

The treatment of GPs up to this point has assumed that our targets are real valued and continuous. In the classification framework, the target is either 0 or 1 to label each datapoint as belonging to one of two classes. To do this, a probit or logit function *f* is used to map the output of the Gaussian Process to the (0,1) interval. The former was used in this study. Define a GP over a function $g(\mathbf{x})$. If we then map the output to the desired interval using $y = f(g)$ the likelihood function becomes

$$p(\mathbf{y} \mid g) = \prod_{n=1}^{N} f(g_n)^{y_n}(1 - f(g_n))^{1-y_n}$$

where the GP prior on g is zero mean with covariance $\mathbf{C} = \mathbf{X}\mathbf{X}^T$. In two class classification, we need only calculate $p(y_{N+1}|\mathbf{y})$. Predictions can be made using

$$p(y_{N+1} = 1 \mid \mathbf{y}) = \int p(y_{N+1} = 1 \mid g_{N+1}) p(g_{N+1} \mid \mathbf{y}) dg_{N+1}$$

The above integral is not Gaussian in form and hence cannot be solved analytically, so approximations must be used. This study made use of the Expectation Propagation algorithm, for further details see [10].

2.6 Covariance Kernels

Note that here we defined the covariance $\mathbf{C}$ as $\mathbf{X}\mathbf{X}^\mathbf{T}$. Rather than defining the covariance as the pairwise inner products of the data in the input space, we can first map the data to some higher, possibly infinite dimensional space, although this makes interpretation more difficult. The covariance matrix is equivalent to the kernel in other methods such as SVMs, and any feature mapping may be used such that $\mathbf{C}$ is positive semidefinite and symmetric. This means that Gaussian Processes have a great deal of flexibility. In addition, the mapping may depend on some set of hyperparameters.

2.7 Learning Hyperparameters

Hyperparameters can be automatically learned from the data by maximizing the log likelihood with respect to them using optimization methods. It is possible to put a hyperparameter on the scale of variation of each input dimension to detect input dimensions that have little effect on the predictive distribution and eliminate them. This technique, known as Automatic Relevance Determination (ARD) [11] was used in the only previous work to use Gaussian Processes to classify structural MRI images [12]. However, this study used voxel level features, meaning the input dimensionality was too high for this to be tractable. Instead, the kernel function used was linear covariance $k(\mathbf{x},\mathbf{x}') = \mathbf{x}^\mathbf{T}\mathbf{x}'$, with a bias term. Only two hyperparameters had to be learned from the data: the mean, and the scale on the covariance. All others were automatically integrated out.

All GP calculations were done using the gpml toolbox (http://www.gaussianprocess.org/gpmlcode /matlab/doc). As a first step, hyperparameters were learned from the data, and then the log probabilities of the test data having labels equal to 1 (that is, having AD) were calculated using GPs. These log probabilities were then exponentiated and thresholded at 0.5 to provide a hard classification.

2.8 SVM Calculations

To provide a baseline classification accuracy for these data, classification was also performed with the widely used LIBSVM support vector machine library (http://www.csie.ntu.edu.tw/~cjlin/libsvm).

All analysis was conducted in MATLAB (The MathWorks Inc., Natick, MA, 2011).

3. RESULTS

Two sets of results are presented, those obtained from GP classification and those from SVM classification. Both sets of results were generated by training a classifier on the 80 designated training subjects, and then evaluating the accuracy of the resulting model in classifying the 40 subjects set aside for training.

The SVM correctly classified 31 out of the 40 test subjects, giving an accuracy of 77.5%. The GP model correctly classified 33 out of the 40 test subjects, equal to an accuracy of 82.5%. However, the difference in absolute classification accuracy is not statistically significant. Both classifiers appeared to be well balanced, with the SVM having a sensitivity (assuming AD = positive) of 75% and specificity of 80%, and the GP having a sensitivity of 80% and specificity of 85%. The results using SVM classification are also in line with those given in [13], a large study comparing various methods of classifying Alzheimer's disease that included a procedure very similar to this one. ROC curves and the associated areas under the curve are given below.

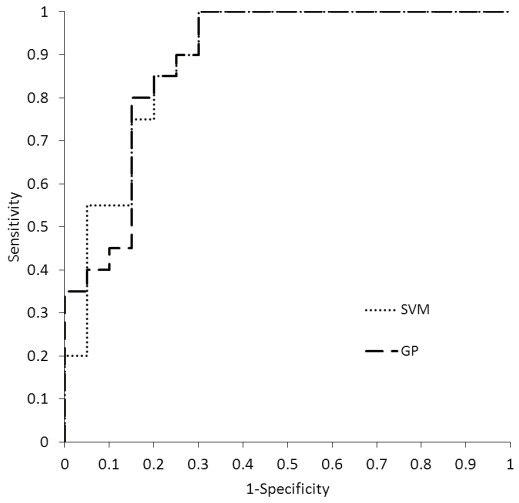


Figure 1: ROC curves and areas for both GP and SVM classification.

As the above figure shows, the ROC curves for the two classification methods are virtually identical. The areas under the two ROC curves are also very similar, at 0.890 for the SVM and 0.888 for the GP.

5. DISCUSSION

Here we have presented the first application of Gaussian process classifiers to distinguish between healthy elderly controls and subjects with AD. While many previous studies have achieved similar accuracy on this type of data [5, 13], they have all used classification algorithms that simply produce a binary decision between two classes, mostly using support vector machines. While the output of SVMs can be converted to probabilities, the methods for doing so are somewhat ad-hoc due to the SVM's non probabilistic objective function [14].

We have chosen to threshold the probabilistic predictions made by the GP classifier at $p(y = 1) = 0.5$ to produce a binary decision that can be directly compared to results from an SVM. The results from the two methods are extremely similar, showing that it is possible to switch to the probabilistic framework without a significant loss of accuracy. Moreover, all the subjects wrongly classified by the GP were also wrongly classified by the SVM, suggesting that the decision boundaries applied by the two classifiers are similar.

However, it would be a waste to limit ourselves to the use of probabilistic predictions in this way. This formulation allows automatic variable selection via maximum likelihood that avoids the effects of overfitting. The equivalent feature selection must be done via computationally expensive cross validation or inaccurate filtering for other methods. While it was not possible to fully apply such techniques in this method, using a labeled atlas to aggregate the grey matter within anatomical regions as in [12] would allow ARD to be used. Alternatively, the feature dimensionality could be reduced by focusing on a small region of interest such as the area around the hippocampus as was also done in [5], or by aggregating voxels into anatomical regions [12]. In the context of multikernel learning, a probabilistic formulation also potentially offers the automatic tuning of the kernel mixing weights.

Probabilistic predictions also open up possibilities that cannot be easily done otherwise, such as a reject option in which more uncertain classifications are passed to a different classifier or human expert, and tuning of the probability threshold to maximize positive predictive value of the test in a clinical context [7].

6. REFERENCES

[1] K. R. Gray et al., "Regional analysis of FDG-PET for use in the classification of Alzheimer's Disease", in 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2011, pp. 1082-1085.

[2] C.-Y. Wee et al., "Identification of individuals with MCI via multimodality connectivity networks", in 2011 Medical Image Computing and Computer-Assisted Intervention: MICCAI, vol. 14, no. 2, pp. 277-284, 2011.

[3] J. Barnes et al., "Differentiating AD from aging using semiautomated measurement of hippocampal atrophy rates", NeuroImage, vol. 23, no. 2, pp. 574-581, Oct. 2004.

[4] J. P. Lerch et al., "Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls", Neurobiology of Aging, vol. 29, no. 1, pp. 23-30, Jan. 2008.

[5] S. Klöppel et al., "Automatic classification of MR scans in Alzheimer's disease", Brain, vol. 131, no. 3, pp. 681 - 689, Mar. 2008.

[6] C. Davatzikos et al., "Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging," *Neurobiology of Aging*, vol. 29, no. 4, pp. 514-523, Apr. 2008.

[7] D. Ashby and A. F. Smith, "Evidence-based medicine as Bayesian decision-making," *Statistics in Medicine*, vol. 19, no. 23, pp. 3291-3305, Dec. 2000.

[8] M. Modat et al., "Fast free-form deformation using graphics processing units," *Computer Methods and Programs in Biomedicine*, vol. 98, no. 3, pp. 278-284, Jun. 2010.

[9] M. J. Cardoso et al., "LoAd: A locally adaptive cortical segmentation algorithm," *NeuroImage*, vol. 56, pp. 1386-1397, Jun. 2011.

[10] T. P. Minka, "Expectation Propagation for approximate Bayesian inference," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, San Francisco, CA, USA, 2001, pp. 362–369.

[11] R. M. Neal, Bayesian learning for neural networks. Springer, 1996.

[12] C. Chu et al., "Classification of Neurodegenerative Diseases Using Gaussian Process Classification with Automatic Feature Determination", in 2010 First Workshop on Brain Decoding: Pattern Recognition Challenges in Neuroimaging (WBD), 2010, pp. 17-20.

[13] R. Cuingnet et al., "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database", NeuroImage, Jun. 2010.

[14] J. Platt, "Probabilities for SV Machines," in Advances in Large-Margin Classifiers, Cambridge, MA: MIT Press, 2000, pp. 61-74.