33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Classification of Alzheimer's Disease from Structural MRI using Sparse Logistic Regression with Optional Spatial Regularization

Anil Rao, Ying Lee, Achim Gass, Andreas Monsch

Abstract— **In this paper, we apply Sparse Logistic Regression Classifiers to the classification of 69 Alzheimer's Disease and 60 normal control subjects based on voxel-wise grey matter volumes derived from structural MRI. Methods such as standard logistic regression cannot be used in such problems because of the large number of voxels in comparison to the number of training subjects. Sparse Logistic Regression (SLR) addresses this issue by incorporating a sparsity penalty into the log-likelihood, which effects an automatic feature selection within the classification framework. We apply two different formulations of sparse logistic regression and compare their classification accuracy with that of Penalized Logistic Regression (PLR) and Maximum uncertainty Linear Discriminant Analysis (MLDA). In the first approach, we use the original formulation of SLR in which correlated voxels are forced to have similar weights. In the second approach we use a spatially regularized formulation, SRSLR, to force the discriminating vector to be spatially smooth when viewed as an image. Evaluation of the methods using cross-validation shows similar classification accuracies for SLR and SRSLR, with both performing better than PLR and MLDA. In addition, SRSLR produced classifiers that were spatially smoother than those produced by SLR, which may better reflect the regional effects of Alzheimer's Disease.**

I. INTRODUCTION

Alzheimer's Disease (AD) is the leading form of dementia worldwide. It has been shown to be associated with reduced grey matter as measured by MRI over the whole brain [1], and within specific anatomical regions such as the hippocampus [2], compared to normal controls (NC). Recent interest in the computational neuroanatomy community has focused on developing tools for diagnosis of AD using multivariate pattern classification techniques applied to voxel-wise, rather than regional or whole brain, measures [3] [4]. Such an approach can potentially be used to develop novel biomarkers of AD and improve understanding of the disease process.

Multivariate voxel-wise classification from MRI is challenging because the number of voxels, ie. features, is typically many orders of magnitude greater than the number of available MRI scans, ie. training set size. In such situations, classifiers are prone to 'overfit' to the training data, and therefore perform poorly on unseen test examples, as a result of the so-called 'curse of dimensionality'. Feature reduction methods such as principal components analysis [5] and feature clustering [6] have been used to overcome this problem, but since these techniques are not fully embedded into the classification, it is highly likely that relevant features will be discarded or the new feature space will not be an appropriate one for classification.

Recent advances in machine learning have addressed these limitations by embedding feature selection into the classification framework by incorporating a sparsity penalty into the objective function that is optimized by the classifier. This results in a classifier where most of the coefficients, or 'weights', are zero which means that corresponding features are discarded as irrelevant to the classification problem. This should not only improve classification accuracy, but also produce classifiers that are more interpretable. Such methods have been successfully applied to functional MRI [7][8][9].

In this paper we propose the novel use of sparse logistic regression classifiers to classify AD from NC using voxel-wise grey matter volumes. The next section describes the mathematical framework for sparse logistic regression classification, while sections III and IV describe the application of the technique to structural MRI data.

II. LOGISTIC REGRESSION

In binary classification, logistic regression models the probability that a subject with row feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ belongs to class $y \in \{0,1\}$ as

$$\begin{aligned} P(y=0|\mathbf{x}) &= \frac{1}{1+e^{\mathbf{x}\beta+\beta_0}} \\ P(y=1|\mathbf{x}) &= \frac{e^{\mathbf{x}\beta+\beta_0}}{1+e^{\mathbf{x}\beta+\beta_0}} \end{aligned} \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)$ and β_0 are the 'weight' vector and intercept respectively. Given a training set $\mathbf{X} = (\mathbf{x}_1; \ldots; \mathbf{x}_n)$ of size n and class memberships $\mathbf{y} = y_i$, β and β_0 are estimated by maximizing the log-likelihood

$$L(\beta, \beta_0) = \sum_{i=1}^{n} y_i(\mathbf{x}_i\beta + \beta_0) - \log(1 + e^{\mathbf{x_i}\beta+\beta_0}) \tag{2}$$

The estimation of β and β_0 is performed by differentiating (2) to give the gradient of the log-likelihood:

$$g(\beta, \beta_0) = \mathbf{X}^T(\mathbf{y} - \mathbf{p}) \tag{3}$$

where $\mathbf{p} = (p_1, \ldots p_n)$, $p_i = P(y_i = 1|\mathbf{x})$, and in which we have added a column of ones to $\mathbf{X}$ to account for the intercept β_0. The 'score equations' are obtained by setting $g = 0$, and are solved using iterative reweighted least squares (IRLS) to

This work was supported by GlaxoSmithKline

A. Rao is with GlaxoSmithKline Clinical Imaging Centre, London, W12 0NN, UK anil.w.rao@gsk.com

Y. Lee is with GlaxoSmithKline Clinical Imaging Centre, London W12 0NN, UK ying.lee@gsk.com

A. Gass is with the Departments of Neurology and Neuroradiology, University Hospital, Basel, Switzerland AGass@uhbs.ch

A. Monsch is with the Memory Clinic, Department of Geriatrics, University Hospital, Basel, Switzerland andreas.monsch@unibas.ch

give the estimates for β and β_0 [10]. A test example is then classified by evaluating the probabilities of class membership in (1) and assigning to the more probable class. Logistic regression cannot be used when there is a large number of features in relation to the number of training samples, as the coefficients of β diverge to $\pm\infty$ to give a perfect fit to the training data but poor predictive performance with unseen test data. In our application, the number of features greatly outnumbers the number of training samples, and so standard logistic regression cannot be used.

A. Sparse Logistic Regression

In sparse logistic regression (SLR), a prior on the weight vector is included which penalizes the log-likelihood and regularizes the estimation of β. The penalized log-likelihood takes the form

$$L_P(\beta,\beta_0) = L(\beta,\beta_0) - \lambda_1 \|\beta\|_1 - \lambda_2 \|\beta\|_2^2 \tag{4}$$

where $L(\beta,\beta_0)$ is as defined in (2). The penalty term, which incorporates both an L_1 and L_2 penalty on the weight vector β, is the 'elastic net' penalty which has recently been applied to both regression and classification problems [11]. The effect of the L_1 penalty is to impose sparseness on β by shrinking its coefficients towards zero, with some weights exactly equal to zero depending on the value of the tuning parameter λ_1. Including this penalty therefore means that a kind of simultaneous continuous feature selection is performed within the classification framework. The L_2 penalty has the effect of regularizing the estimation of β, and tends to make correlated features have similar weights. Letting $\lambda_1 = 0$, ie. removing the sparsity penalty but keeping the L_2 penalty, gives the Penalized Logistic Regression (PLR) criterion which has been used in genetics classification tasks [12].

Classical methods for optimization such as IRLS cannot be used to maximize (4) because the L_1 term causes the penalized log-likelihood to be non-differentiable when any of the coefficients of β are equal to zero. However, the penalized log-likelihood is still a concave function and can be solved using a number of different algorithms [13] [14]. We adopt the bound optimization approach of [14] [7], in which (4) is optimized by iteratively maximizing a surrogate function Q at each iteration t

$$\hat{w}^{(t+1)} = \arg\max_w Q(w|\hat{w}^{(t)}) \tag{5}$$

where $w = (\beta,\beta_0)$. For SLR, the surrogate function is [7]

$$\begin{aligned} Q(w|\hat{w}^{(t)}) =& w^T(g(\hat{w}^{(t)}) - B\hat{w}^{(t)}) + \frac{1}{2}w^T B w \\ & - \lambda_1 \|\beta\|_1 - \lambda_2 \|\beta\|_2^2 \end{aligned} \tag{6}$$

where $g(w)$ is the gradient of the log-likelihood in (3), and B is the matrix

$$B = -0.25 \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \tag{7}$$

where each $\mathbf{x}_i$ is a row in the augmented data matrix $\mathbf{X}$ in (3). Instead of optimizing (6), it is maximised with respect to just one of its components at each iteration t using the update equations

$$\hat{w}_k^{(t+1)} = \begin{cases} \frac{s(-B_{k,k}\hat{w}_k^{(t)} + g_k(\hat{w}^{(t)}),\lambda_1)}{2\lambda_2 - B_{k,k}} & \text{if } k < p+1 \\ \frac{-g_{p+1}(\hat{w}^{(t)}) + B_{p+1,p+1}\hat{w}_{p+1}^{(t)}}{B_{p+1,p+1}} & \text{if } k = p+1 \end{cases} \tag{8}$$

where the updates of the weight vector components $w_1,\ldots,w_p$ and intercept component w_{p+1} are different since the intercept is not penalized in (4). The function s in (8) is the soft-thresholding operator:

$$s(x,y) = sign(x)\max\{0,|x|-y\} \tag{9}$$

B. Spatially Regularized Sparse Logistic Regression

Sparse logistic regression tends to give correlated features similar weights in the estimated weight vector β due to the L_2 penalty in (4). However, since in our application the training data, and hence β are images, we can use a penalty that enforces spatial smoothness on β to regularize the solution. This is achieved by maximizing the following penalized log-likelihood function:

$$L_{P_\Omega}(\beta,\beta_0) = L(\beta,\beta_0) - \lambda_1 \|\beta\|_1 - \lambda_2 \beta^T \Omega \beta \tag{10}$$

where Ω is chosen so that $\beta^T\Omega\beta$ is a discrete approximation to the integral of the 3D Lapacian of β, when viewed as an image, over a region of interest R.

$$\beta^T \Omega \beta \approx \iiint_R (\beta_{xx} + \beta_{yy} + \beta_{zz})^2 \tag{11}$$

Note that if we let $\Omega = I$, then we have the SLR penalized log-likelihood in (4). The surrogate function for maximising (10) now becomes

$$\begin{aligned} Q(w|\hat{w}^{(t)}) =& w^T(g(\hat{w}^{(t)}) - B\hat{w}^{(t)}) + \frac{1}{2}w^T B w \\ & - \lambda_1 \|\beta\|_1 - \lambda_2 \beta^T \Omega \beta \end{aligned} \tag{12}$$

Differentiating (12) with respect to a component of w gives the following update equation for the weights

$$\hat{w}_k^{(t+1)} = \frac{s(\alpha_k \hat{w}_k^{(t)} - 2\lambda_2 \Omega_{k,} \hat{w}_{1:p}^{(t)} + g_k(\hat{w}^{(t)}),\lambda_1)}{\alpha_k} \tag{13}$$

where $\alpha_k = 2\lambda_2\Omega_{k,k} - B_{k,k}$, $\Omega_{k;}$ is the kth row of Ω, and $\hat{w}_{1:p}^{(t)}$ refers to the weight-only components of $\hat{w}^{(t)}$. The update equation for the intercept w_{p+1} is the same as in (8). We refer to this formulation of SLR as Spatially Regularized Sparse Logistic Regression (SRSLR).

III. Application to Classification of Alzheimer's Disease

A. Materials

We applied SLR and SRSLR to the classification of AD using structural MRI of the brain. 75 AD and 65 NC subjects were recruited at the Memory Clinic, University Hospital (USB), Basel, Switzerland. Patients were diagnosed as probable/possible AD if they met the criteria defined by NINCDS and ADRDA [15]. MRI data acquisition was performed on a

3 T Siemens Allegra head-only MRI system (Siemens, Erlangen, Germany) at the USB. High-resolution structural images (FOV = 256x256mm, voxel size = 1.1x1.1x1.1mm, plane: sagittal) were acquired using a T1-weighted magnetization-prepared rapid gradient-echo (MPRAGE) sequence (TR = 2150 ms, TI = 1000 ms, TE = 3,49 ms, TA = 7 min, flip angle = 7). All subsequent image analysis was performed at the GlaxoSmithKline Clinical Imaging Centre, London, United Kingdom.

B. Preprocessing

The T1-weighted images were pre-processed using SPM8 software [16]. Firstly, all native images were segmented into grey matter (GM) and white matter (WM) [17]. Of these images, 6 AD and 5 NC were excluded from the rest of the analysis due to large amounts of mistakenly segmented dura in their native GM segmentations. The GM and WM segmentations were then non-rigidly aligned to an intensity average template using DARTEL [18], and the aligned GM segmentations were then affine transformed into MNI space. Volume preservation was used throughout so that the resulting images have the same volume of GM as the native segmentations. Finally, the GM images were corrected for volume differences due to head size using the affine part of the native-to-MNI space mapping determined during segmentation.

The images were then smoothed with an isotropic Gaussian 1mm FWHM kernel, giving the training and test data for the classifier. We also produced a corresponding set of images smoothed with an isotropic Gaussian 8mm FWHM kernel to which we applied a GM masking threshold of 0.1 to exclude voxels which were likely not to be GM. Finally, we took the corresponding voxels in the 1mm-smoothed images and standardized the GM at each of these voxels to have zero mean and unit variance over the set of subjects, as is commonly performed with penalized methods. The standardized GM images were then used as input features in the classifiers.

C. Evaluation of Classification Performance

We used 10-fold cross-validation to evaluate the performance of the classifiers on the 129 subjects. This requires that we randomly split the training set into 10 folds and predict each fold in turn after training on the remaining 9 folds. The sensitivity (proportion of AD subjects correctly predicted), specificity (proportion of NC subjects correctly predicted), and classification accuracy (proportion of all subjects correctly predicted) were then calculated for each fold, and the mean and standard errors of these measures across folds was determined.

Since SLR and SRSLR both require two tuning parameters, λ_1 and λ_2, we estimated the optimum values using nested cross validation, ie., for each test fold the remaining 9 training folds were further divided into 5 inner folds, over which cross validation was used to determine the best choice of the parameters.

TABLE I
CLASSIFIER PERFORMANCE FOR CLASSIFICATION OF AD

Classifier	Sensitivity	Specificity	Accuracy
SLR	$90.77 \pm 3.67\%$	$80.26 \pm 3.93\%$	$85.26 \pm 1.39\%$
SRSLR	$90.35 \pm 3.73\%$	$80.26 \pm 3.93\%$	$85.26 \pm 1.81\%$
PLR	$85.85 \pm 3.67\%$	$79.85 \pm 4.88\%$	$82.95 \pm 2.23\%$
MLDA	$85.10 \pm 4.38\%$	$79.85 \pm 4.88\%$	$82.95 \pm 2.23\%$

We used PLR as described in section II-A and Maximum uncertainty Discriminant Analysis (MLDA) [19] as comparators with the SLR/SRSLR approaches described. Since PLR is equivalent to SLR without the sparsity penalty, this enables a comparison of SLR/SRSLR with their 'non-sparse' equivalent, so that the effect of the sparsity penalty, both on feature extraction and on classifier performance can be assessed. The single tuning parameter for PLR was estimated using nested cross validation as for SLR/SRSLR. MLDA is similar to Gaussian Linear Discriminant Analysis, but uses an entropy-based approach for stabilizing the estimation of the pooled covariance matrix and does not involve any tuning parameters unlike other regularized LDA approaches. It has been shown to give comparable accuracy to a linear support vector machine in predicting mental state from FMRI [20].

IV. RESULTS

Table I summarizes the performance of SLR, SRSLR, PLR and MLDA across the 10 folds. SLR and SRSLR perform similarly well, with both giving better overall accuracies than PLR and MLDA. In particular, the sensitivities of both SLR and SRSLR for detection of AD subjects are 5% greater than that of the other approaches. The improvements with SLR/SRSLR could be as a result of the feature selection inherent to these methods, which seeks to eliminate noisy variables that harm classifier performance.

We can also see the implications of feature selection for classifier interpretability in Fig. 1, in which we show axial and sagittal views of the weight images estimated for the fourth fold for each classifier, overlaid on the MNI structural brain atlas. The crosshairs are located within the left hippocampus. Negative weights, shown in 'hot' colours, indicate reduced grey matter in AD subjects compared to NC subjects, while 'cool' colours indicate increased grey matter. We can see that both SLR and SRSLR have extracted a small percentage of the $\approx 2 \times 10^5$ voxels fed into the classifier, with negative weights found in clinically relevant regions for AD such as the left hippocampus and left amygdala [2]. The extraction of these regions implies that the feature selection component of SLR/SRSLR is giving physiologically plausible results while simultaneously giving a good classification performance. In comparison, both PLR and MLDA, since they cannot perform feature selection, produce weight images with non-zero values at every voxel. Such classifiers require arbitrary post-processing eg., thresholding, before voxels can be discarded as unimportant for distinguishing between AD and NC subjects. As expected, the weight images produced by SRSLR were in general smoother than that produced by SLR, as can be seen in Fig. 1. The smoother images may be

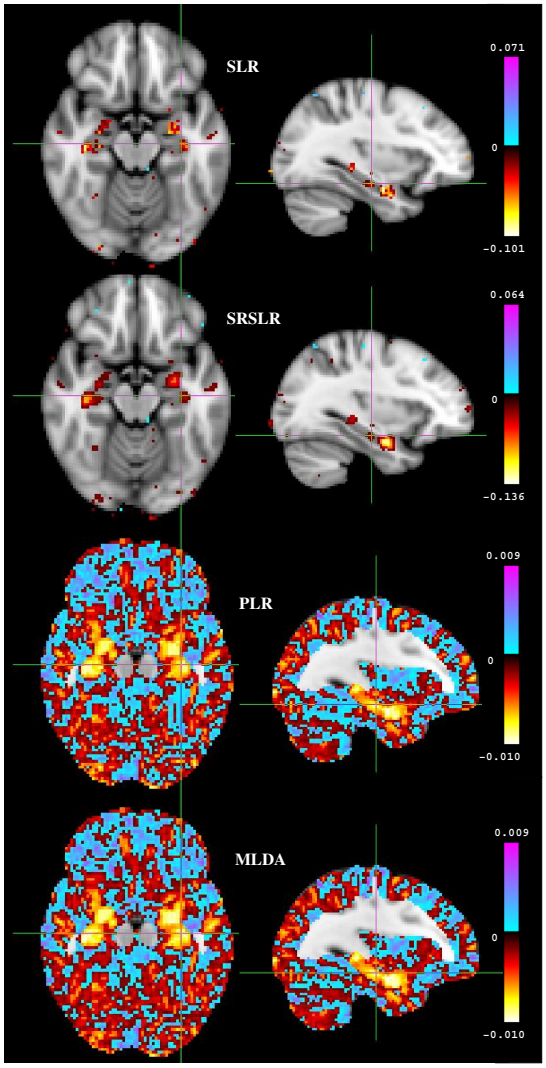

Fig. 1. This figure shows the weight images produced for the fourth fold using SLR, SRSLR, PLR and MLDA. Each of the weight images has been normalized to length 1, and weight values at each voxel are indicated by the colour shown. Both SLR (1655 non-zero weights) and SRSLR (3724 non-zero weights) give sparse weight vectors with negative values, indicating reduced grey matter volume, in clinically relevant regions for AD. PLR and MLDA assign non-zero weights to all 197150 voxels and are less interpretable.

more physiologically plausible, since they better reflect the regional effects of Alzheimer's Disease.

V. CONCLUSIONS

We have applied sparse logistic regression and spatially regularized sparse logistic regression to classify AD and NC subjects from high-resolution structural MRI. Both methods were able to automatically select clinically relevant regions for AD while simultaneously performing the classification with better accuracies than Penalized Logistic Regression and Maximum uncertainty Linear Discriminant Analysis. The incorporation of a smoothing penalty in Spatially Regularized Sparse Logistic Regression gave smoother weight images than Sparse Logistic Regression although the accuracies of both methods were similar. In the future, we intend to apply these methods to different classification problems such as distinguishing between MCI (Mild Cognitive Impairment) and NC subjects from structural MRI.

REFERENCES

[1] N. Fox, P.Freeborough, and M. Rossor, "Visualisation and quantification of rates of atrophy in alzheimer's disease," *Lancet*, vol. 348, pp. 94–97, 1996.

[2] P. Thompson, K. Hayashi, G. D. Zubicaray, A. Janke, S. Rose, J. Semple, D. Herman, M. Hong, S. Dittmer, D. Doddrell, and A. Toga, "Dynamics of gray matter loss in alzheimer's disease," *Journal of Neuroscience*, vol. 23, pp. 994–1005, 2003.

[3] S. Kloppel, C. Stonnington, C. Chu, B. Draganski, R. Scahill, J. Rohrer, N. Fox, C. Jack, J. Ashburner, and F. Frackowiak, "Automatic classification of mr scans in alzheimer's disease," *Brain*, vol. 131, pp. 681–689, 2008.

[4] P. Vemuri, J. Gunter, M. Senjem, J. Whitwell, K. Kantarci, D. Knopman, B. Boeve, R. Petersen, and C. Jack, "Alzheimer's disease diagnosis in individual subjects using structural mr images:validation studies," *Neuroimage*, vol. 39, pp. 1186–1197, 2008.

[5] C. Davatzikos, S. Resnick, X. Wu, P. Parmpi, and C. Clark, "Individual patient diagnosis of ad and ftd via high-dimensional pattern classification of mri," *Neuroimage*, vol. 41, pp. 1220–1227, 2008.

[6] C. Plant, S. Teipel, A. Oswald, C. Bohm, T. Meindl, J. Mourao-Miranda, A. Bokde, H. Hampel, and M. Ewers, "Automated detection of brain atrophy patterns based on mri for the prediction of alzheimer's disease," *Neuroimage*, vol. 50, pp. 162–174, 2010.

[7] S. Ryali, K. Supekar, D. Abrams, and V. Menon, "Sparse logistic regression for whole-brain classification of fmri data," *Neuroimage*, vol. 51, pp. 752–764, 2010.

[8] L. Grosenick, S. Greer, and B. Knutson, "Interpretable classifiers for fmri improve prediction of purchases," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 16, pp. 539–548, 2008.

[9] B. Ng, A. Vahdat, G. Hamarneh, and R. Abugharbieh, "Generalized sparse classifiers for decoding cognitive states in fmri," in *Proc. of MICCAI Workshop on Machine Learning in Medical Imaging*, 2010, pp. 108–115.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

[11] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of Royal Statistical Society: Series B*, vol. 67, pp. 301–320, 2005.

[12] J. Zhu and T. Hastie, "Classification of gene microarrays by penalized logistic regression," *Biostatistics*, vol. 5, pp. 427–443, 2004.

[13] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, pp. 1–22, 2010.

[14] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, pp. 957–968, 2005.

[15] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. Stadian, "Clinical diagnosis of alzheimer's disease: Report of the nincds-adrda work group under the auspices of department of health and human services task force on alzheimer's disease," *Neurology*, vol. 34, pp. 939–944, 1984.

[16] (2009) Statistical parametric mapping version 8. [Online]. Available: http://www.fil.ion.ucl.ac.uk/spm/software/spm8/

[17] J. Ashburner and K. Friston, "Unified segmentation," *Neuroimage*, vol. 26, pp. 839–851, 2005.

[18] J. Ashburner, "A fast diffeomorphic image registration algorithm," *Neuroimage*, vol. 38, pp. 95–113, 2007.

[19] C. Thomaz, F. Duran, G. Busatto, D. Gillies, and D. Rueckert, "Multivariate statistical differences of mri samples of the human brain," *Journal of Mathematical Imaging and Vision*, vol. 29, pp. 95–106, 2007.

[20] J. Sato, A. Fujita, C. Thomaz, G. Martin, J. Mourao-Miranda, M. Brammer, and E. Amaro, "Evaluating SVM and MLDA in the extraction of discriminant regions for mental state prediction," *Neuroimage*, vol. 46, pp. 105–114, 2009.