Symmetric and Multi-Scale Features for Automatic Segmentation of Multiple Sclerosis Lesions using Pattern Classification

M. Battaglini[1], N. De Stefano[1], and M. Jenkinson[2]
[1]Quantitative Neuroimaging Laboratory, University of Siena, Siena, Italy, [2]Clinical Neurology, FMRIB Centre, University of Oxford, Oxford, Oxon, United Kingdom

Introduction: Our aim is to create a fully automated tool for the segmentation of MS lesions that is reliable, repeatable and suitable for large-scale clinical trials. Existing segmentation methods are often tested on unrealistic data (e.g. simulated data) or data from a single scanner, unlike that used in clinical trials. In this study we explore using novel input features with two pattern classification methods (Neural Networks and Random Forests). The particular features are motivated by observations of manual raters and include information from a broader neighbourhood (multi-scale) and measures of the asymmetry of typical lesions. We test our methods with a set of typical multi-site clinical trial data.

Methods: *Dataset*: 27 subjects with T1w and PD/T2w images (0.97x0.97x3mm), from 21 sites, plus manual lesion segmentations (done by 4 different raters). These are divided into a 6 subject training set and a 21 subject testing set. The training set was deliberately made small and heterogeneous (lesion loads from 1.1 to 36.6 cc) to make it feasible to retrain this method – for example, with different modalities, scanners, field strengths, resolutions as well as different manual raters (with potentially different inclusion/exclusion criteria for lesions).

Figure 1: Illustration of Multi-Scale Feature (left=original image; middle=segmentation; right=3x3 multi-scale features)

Features: In addition to standard features as used in Dyrby et al [1] (3x3 neighbourhood of intensities from each image, plus standard-space coordinates) we propose adding multi-scale features consisting of an array of averages within patches of size 3x3 or 9x9 voxels within the 2D slice [see Fig 1]. As well as this we propose a symmetry index: S=2(Ia –Ib)/(Ia + Ib) where Ia and Ib are intensities (or patches) in the ipsi-lateral and contra-lateral hemisphere respectively (mirrored coordinates in the inter-hemispheric plane). Symmetry is calculated for the PD and T2w images, plus at the 3x3 and 9x9 patch scales. The multi-scale features were calculated for the three different image modalities, PD, T2w and T1w, as well as on an artificially-enhanced image we call Pseudo-FLAIR: PsF=(PD*T2w*T1w)/(PD+T2w). To test the utility of these new features we use various different combinations (feature sets) in the experiments: i.e. all the above features (ALL); all except coordinates (NoC); all except symmetry and coordinates (NoSNoC); 3x3 patch symmetry and 3x3 patch multi-scale but no 9x9 multi-scale or 9x9 symmetry (S3+MS3); both 3x3 and 9x9 patch symmetry but only 3x3 patch multi-scale for intensities (S3+S9+MS3); no symmetry and no multi-scale features, but just 3x3 intensities plus coordinates (3x3C) as used in [1]. For example, dimensions of some of the used feature sets were: ALL=165, NoC=162, NoSNoC=108, 3x3+C=39.

Pre-processing: Intensities in each image were initially normalized by dividing them by the 98th percentile of all intensities in that image. T1w images were registered (6 DOF) to T2w/PD images and resampled to this space [5].

Pattern Classification methods: Two classifiers were used: (i) Neural Networks (NN) [2] (Netlab toolbox in MATLAB) with 39 to 165 input nodes (depending on the feature set), 80 hidden nodes and a single sigmoidal output node, giving a value between 0 and 1, where these parameters were partly based on those chosen in Dyrby et al [1] but without any pruning of the network; and (ii) Random Forest (RF) [3], which is a collection of binary decision trees used to vote on the outcome. Each tree is trained independently using a subset of the features and a random set of training examples (voxels) selected by bagging (selection with replacement). The random forest configuration in these experiments used 5 features per tree (selected based on initial experiments), with 500 trees per forest, and with a total of 56618 training examples used in training (number of voxels selected at random from the training set, but with an enforced equal number of lesion and non-lesion voxels – the same set of training data was used for the NN).

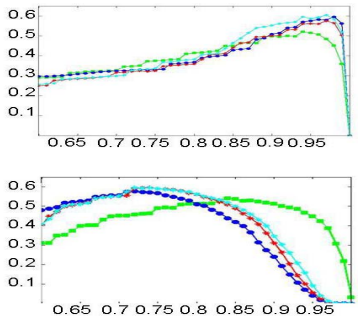

Figure 2: Average DICE scores, across all test subjects, versus threshold for the different classifiers (NN top; RF bottom) - colours given in text.

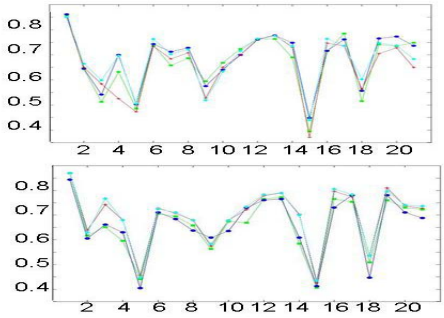

Figure 3: Best DICE scores for each subject in the testing set (NN top; RF bottom) – colours given in text.

Post-processing: A brain mask was applied to segmentation output and individual lesions (as defined by a threshold on the output value and a connected-component algorithm) were rejected if the bordering voxels consisted of less than 30% white matter (as defined by a voxel classification [4] on the T1w image , registered to the T2w/PD images [5]).

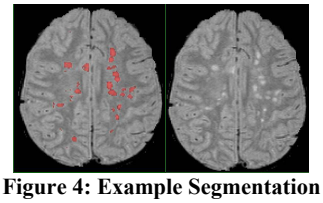

Figure 4: Example Segmentation

Experiments: An initial set of experiments was done using a Leave-One-Out (LOO) test on the 6 training subjects. These results were used to select the best set of features to explore further. Subsequent tests used the testing set with the differently trained classifiers and quantified the results using DICE (Similarity Index) to compare the thresholded classifier output with the manual labels.

Results: Initial LOO tests selected the following as the best classifiers: S3+S9+MS3+C/NN (light blue); S3+MS3+C/NN (blue); S3+MS3/NN (red); Full/RF (light blue); NoSNoC/RF (blue); NoC/RF (red). These were also compared to the base case of 3x3+C (green), as used in [1]. Fig 2 shows DICE averaged across all the test subjects as a function of the threshold applied to the classifier output (since each classifier gives an output between 0 and 1). This is done for each feature set/classifier combination. Fig 3 shows the best DICE values obtained for each subject in the testing set for each feature set/classifier. There is substantial variation between subjects but some very good segmentations were achieved. Statistical tests with a non-parametric rank sign test (paired data) showed significant differences when adding symmetry or multi-scale features in RF, but not for coordinates. For NN there was a significant result for all feature sets versus the base case (3x3+C) as well as a significant difference when adding the multi-scale symmetry feature, but not when adding coordinates.

Discussion: Results show a statistically significant improvement in DICE by using the multi-scale and symmetry features with either a Neural Network or Random Forest classifier. Overall DICE varied significantly between subjects but averaged to around 60%, which is not as high as some studies, but greater than obtained by Dyrby et al [1]. However, it is extremely difficult to compare results between studies due to differences in subjects, image acquisitions, single vs multi-site data, single raters, etc. We believe that this real clinical data, acquired from multiple sites and segmented by different raters, makes this task very challenging. Nonetheless we have achieved DICE consistent with state-of-the-art methods, without requiring very costly pruning of the Neural Networks, which can take weeks of computational time versus a couple of hours for our training. In addition, we applied no exclusion criteria to the images. The advantages of the new features were proved statistically and showed to hold for two different types of classifier, demonstrating their general utility.

References: 1. Dyrby et al, Neuroimage, 41, 2008; 2. Bishop, Neural Networks for Pattern Recognition, OUP; 3. Breiman et al, Machine Learning, 45, 2001; 4. Zhang et al, IEEE TMI, 20(1), 2001; 5. Jenkinson et al, Med Im Anal, 5(2), 2001.