MRI Brain Scan Classification According to The Nature of The Corpus Callosum

Ashraf Elsayed[1]
a.el-sayed@liv.ac.uk
Frans Coenen[1]
coenen@liv.ac.uk
Marta García-Fiñana[2]
m.garciafinana@liv.ac.uk
Vanessa Sluming[3]
vanessa.sluming@liv.ac.uk

[1] Department of Computer Science, University of Liverpool
[2] Centre for Medical Statistics and Health Evaluation, University of Liverpool
[3] School of Health Sciences, University of Liverpool

Abstract

Two mechanisms for classifying Magnetic Resonance Image (MRI) brain scans according to the nature of the corpus callosum are described. The first mechanism adopts an approach founded on the concept of graph mining whereby MRI scans are represented in terms of frequently occurring sub-graphs across the data set, the second is founded on a time series representation coupled with a Case Based Reasoning (CBR) approach to classification. The two mechanisms are evaluated through application to a set of MRI scans describing musicians and non-musicians. In both cases a high degree of accuracy is obtained.

1 Introduction

This paper describes and compares two approaches to classifying (catagorising) MRI brain scans according to the nature of the corpus callosum, a structure of the mammalian brain that connects the two hemispheres; a graph mining based approach and a time series analysis based approach. Both approaches, although operating in very different manners, are essentially supervised learning mechanisms whereby a pre-labelled training set is used to build a "classifier" which can be applied to unseen data. The first approach uses a tree based representation for the corpus callosum, one tree per image. A graph mining technique is then applied to identify frequently occurring sub-graphs (sub-trees). The identified set of trees are then used to describe the image set so that it is described in terms of a set of attributes, each of which equates to a frequently occurring sub-tree. A decision tree algorithm is then applied to this attribute set to build a classifier to be applied to "unseen" data. The second approach is founded on a time series representation coupled with a Case Based Reasoning mechanism. The features of interest are represented as time series, one per image. These time series are then stored in a Case Base (CB) which can be used to categorise unseen data. The unseen data is compared with the categorisation on the CB using a Dynamic Time Warping (DTW) based similarity checking mechanism, the categorisation associated with the most similar time series (case) in the CB is then adopted as the categorisation for the unseen data.

2 Application Domain

The work described in this paper is directed at the classification of MRI brain scan data according to the corpus callosum. This is a highly visible structure in MRI scans whose function is to connect the left and right hemispheres of the brain, and to provide the communication conduit between these two hemispheres. Figure 1 gives an example MRI scan, the corpus callosum is located in the centre of the image. A related structure, the *fornix* is also indicated. The fornix often "blurs" into the corpus callosum and thus presents a particular challenge in the context of the segmentation of these images.

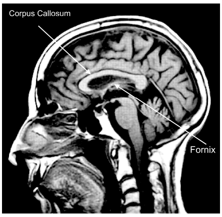

Figure 1: corpus callosum in a midsagittal brain MR image.

The corpus callosum is of interest to medical researchers for a number of reasons. The size and shape of the corpus callosum have been shown to be correlated to sex, age, neurodegenerative diseases and various lateralized behaviour in people. It is also conjectured that the size and shape of the corpus callosum reflects certain human characteristics (such as a mathematical or musical ability). Several medical studies indicate that the size and shape of the corpus callosum, in humans, are correlated to sex and age [8], brain growth and degeneration [4], handedness [2] and various types of brain dysfunction [5].

3 Graph Based Approach

The proposed graph based classification process commences with segmentation and registration to isolate the corpus callosum in each image. The pixel represented corous callosum is then tesselated into homogenous sub-regions. Tessellation entails the recursive decomposing of an identified Region Of Interest (ROI), into quadrants. The tesselation continues until either sufficiently homogenous quadrants are identified or some user specified level of granularity is reached. The result is then stored in a quadtree data structure such that each root node represents a *tile* in the tesselation. Nodes nearer the root of the tree represent larger tiles than nodes further away. Thus the tree is "unbalanced" in that some root nodes will cover larger areas of the ROI than others. The advantage of the representation is thus that it maintains information about the relative location and size of groups of pixels (i.e. the shape of the corpus callosum).

A weighted frequent sub-graph mining technique was developed to identify commonly occuring sub-trees within the tree represented image set. The weightings were calculated according to the proximity of individual nodes to the root node in each tree. This weighting concept was built into a variation of the well known gSpan algorithm [9]. The algorithm operates in a depth first search manner, level by level, following a "generate, calculate support, prune" loop. Candidate sub-graphs are pruned if their *support* (frequency of occurrance across the graph set) is below a user defined "support threshold". Note that the lower the threshold the greater the number of frequent sub graphs that will be identified. Space restrictions preclude further detailed discussion of this weighted sub-graph mining algorithm here, however, interested readers are referred to [6].

The identified sub-trees (graphs) thus form the fundamental elements of a *feature space*. Experiments conducted by the authors have revealed that, for many image sets, the graph mining process can identify a great many frequent sub-graphs; more than required for the desired categorisation. Therefore a feature selection strategy is applied so that only those sub-tree that serve as good discriminators are retained. A straightforward wrapper method was adopted whereby a decision tree generator was applied to the feature set. Sub-trees (features) included as "choice points" in the decision tree were selected, while all remaining features were discarded. For the work described here, the well established C4.5 algorithm [7] was adopted. On completion of the feature selection process each image is described in terms of a binary-valued feature vector indicating the selected features (sub-trees) that appear in the image. Once the image set has been represented in this manner any appropriate classifier generator may be applied. For additional information regarding the graph based approach, including the tesselation process, interested readers are referred to [3].

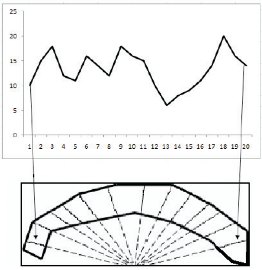

Figure 2: Conversion of corpus callosum into time series.

4 Time Series Based Approach

As in the case of the graph based approach, the time series based approach commences with the segmentation and registration of the input images. The next step is to derive a time series according to the boundary line circumscribing the corpus callosum. The time series is generated using an ordered sequence of N vectors radiating out from a single reference point. The derived time series is then expressed as a series of values (one for each of the N vectors) describing the size (length) of intersection of the vector with the ROI. The representation thus maintains the structural information (shape and size) of the corpus callosum. It should also be noted that N is often variable due to the differences of the shape and size of the individual ROI within the image data set.

With respect to the corpus callosum application the time series generation procedure is illustrated in Figure 2. The midpoint of the lower edge of the Minimum Bounding Rectangle (MBR) was selected as the reference point. The vectors were derived by rotating an arc about the reference point pixel by pixel, hence the value of N will very across the image set. In this manner a time series curve may be generated of the form described in the top half of Figure 2 where the X-axis represents the vector (arc) number, and the Y-axis the "pixel-distance" where the vector intersects the ROI (corpus callosum).

Each time series is then conceptualised as a *proto-type* or case contained in a Case Base (CB), to which a Case Based Reasoning (CBR) mechanism can be applied. Thus, given an unseen record, the record can be classified according to the "best match" discovered in the CB. The CBR community has proposed many techniques to identify the desired best match.

In this paper the authors advocate a Dynamic Type Warping (DTW) time series analysis technique for comparing curves [1]. The advantage offered is that DTW is able to find the optimal alignment between two time series Q and C, of length n and m respectively. The DTW-distance between the two time series Q and C is $D(M,N)$ was calculated as follows:

$$D(i,j) = d(q_i, c_j) + \min\{D(i-1,j-1), D(i-1,j), D(i,j-1)\} \quad (1)$$

Backtracking along the minimum cost k^{th} index pairs $w(i,j)_k$ starting from (m,n) yields the DTW *warping path*.

5 Evaluation

To evaluate and compare the two proposed approaches a data set used comprised 106 brain MRI scans was used. The data set comprised two equal categories (classes), 53 images per category, namely musicians and non-musicians. There is significant evidence, amongst the medical community, that traits such as musical ability, influence the shape and size of the corpus callosum. It should be noted that a visual inspection of the MRI images does not indicate any discernible distinction between the two categories. Table 1 shows the Ten Cross Validation (TCV) classification results obtained using the proposed techniques. The columns labelled GB (Graph Based) and TSB (Time Series Based) indicate the classification accuracy obtained in each case. With respect to the GB approach a quad tree depth of 6 coupled with a 30% threshold support produced the best classification accuracy. Table 2 shows the *confusion matrix* for the best result using GB approach listed in Table 1. This gives a precision of 96.15%, a sensitivity of 94.34% and a specificity of 96.23%. A corresponding confusion matrix for the best result using the time series approach is unecessary.

Table 1: TCV Classification accuracy (%) for musicians using GB and TSB approaches

Test set ID	GB	TSB
1	92.45	91
2	96.23	100
3	95.28	91
4	93.4	100
5	97.17	100
6	94.34	100
7	97.17	100
8	95.28	100
9	96.23	100
10	95.28	100
Average	95.28	98.2
SD	1.54	3.8

Table 2: Confusion matrix for best graph based approach

	Pos.	Neg.	Totals
True	50	3	53
False	2	51	53
Totals	52	54	106

Table 3 gives some fiuther average TCV results obtained using the GB approach but with a variety of quad-tree depths and support thresholds. The best result for each depth of quad-tree is indicated in **bold** font. Inspection of the two Tables (1 and 3) demonstrate that the overall classification accuracy (100%) of the TSB approach improves on the GB approach. Although both algorithms perform well.

Table 3: TCV Classification accuracy (%) using graph based ROIBIC

Levels	Support Threshold (%)							
	20	30	40	50	60	70	80	90
4	**70.75**	69.81	68.87	71.70	68.87	61.32	52.83	50.94
5	**90.57**	83.96	80.19	85.85	80.19	81.13	80.19	70.75
6	85.85	**95.28**	84.91	83.96	90.57	83.96	77.36	75.47
7	83.80	85.85	**89.62**	86.79	87.74	75.47	76.42	78.30

5.1 Conclusions

Two approaches to ROI Based Image Classification, founded on graph mining and time series analysis respectively, have been described. The work was directed at MRI brain scan data, and illustrated by considering MRI scan classification according to the nature of the corpus callosum featured within these images. High accuracy results are reported for both approahes. However, the approach has general applicability. The research team are also interested in alternative methods of pre-processing MRI data, and mechanism for the post-processing of results to provide explanations for specific classifications. The latter is seen as particularly significant in the context of medical research involving MRI scan data, such as in the case of the presented application.

References

[1] D. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *Proc. AAAI-94 workshop on Knowledge Discovery in Databases*, pages 359–370, 1994.

[2] P. Cowell, A. Kertesz, and V. Denenberg. Multiple dimensions of handedness and the human corpus callosum. *Neurology*, 43:2353–2357, 1993.

[3] A. Elsayed, F. Coenen, C. Jiang, M. García-Fi nana, and V. Sluming. Corpus callosum mr image classification. In *Proc. AI'2009, Springer*, pages 333–346, 2009.

[4] H. Hampel, S. Teipel, G. Alexander, B. Horwitz, D. Teichberg, M. Schapiro, and S. Rapoport. Corpus callosum atrophy is a possible indicator of region and cell type-specific neuronal degeneration in alzheimer disease. *Archives of Neurology*, 55:193–198, 1998.

[5] G. Hynd, J. Hall, E. Novey, D. Eliopulos, K Black, J. Gonzalez, J. Edmonds, C. C. Riccio, and M. Cohen. Dyslexia and corpus callosum morphology. *Archives of Neurology*, 52:32–38, 1995.

[6] C. Jiang and F. Coenen. Graph-based image classification by weighting scheme. In *Proc. AI'2008, Springer*, pages 63–76, 2008.

[7] R. Quinlan. *C4.5: A program for machine learning*. Morgan Kaufmann, 1993.

[8] D. Salat, A. Ward, J. Kaye, and J. Janowsky. Sex differences in the corpus callosum with aging. *Journal of Neurobiology of Aging*, 18:191–197, 1997.

[9] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In *Proc. ICDM'02: 2nd IEEE Conf. Data Mining*, pages 721–724, 2002.