AUTOMATIC CLASSIFICATION OF ALZHEIMER'S DISEASE VS. FRONTOTEMPORAL DEMENTIA: A SPATIAL DECISION TREE APPROACH WITH FDG-PET

N. Sadeghi[1], *N. L. Foster*[2], *A. Y. Wang*[2], *S. Minoshima*[3], *A. P. Lieberman*[4], *T. Tasdizen*[1]

[1]School of Computing, University of Utah, Salt Lake City, UT 84112
[2]Center for Alzheimer's Care, Imaging and Research, University of Utah, Salt Lake City, UT 84112
[3] School of Medicine, University of Washington, Seattle, WA 98195
[4] Department of Pathology, University of Michigan, Ann Arbor, MI 48109

ABSTRACT

We introduce a novel approach for the automatic classification of FDG-PET scans of subjects with Alzheimers Disease (AD) and Frontotemporal dementia (FTD). Unlike previous work in the literature which focuses on principal component analysis and predefined regions of interest, we propose the combined use of information gain and spatial proximity to group cortical pixels into empirically determined regions that can best separate the two diseases. These regions are then used as attributes in a decision tree learning framework. We demonstrate that the proposed method provides better classification accuracy compared to other methods on a group of 48 autopsy confirmed AD and FTD patients.

Index Terms— Brain imaging, decision tree, FDG-PET, Alzheimer's Disease, Frontotemporal dementia.

1. INTRODUCTION

Distinguishing with confidence among different neurodegenerative diseases that share the same behavioral symptoms is a challenging problem in clinical diagnosis. Positron emission tomography (PET) image analysis has great potential to aid clinical diagnosis in this respect. In PET imaging, a radioactive tracer isotope incorporated into a metabolically active molecule such as fluorodeoxyglucose (FDG) is injected into the subject. This allows the imaging of metabolic activity for glucose. Since brain energy normally is completely dependent upon glucose, FDG-PET accurately reflects brain function. FDG-PET imaging promises a better accuracy rate compared to qualitative judgments required to clinically distinguish different types of dementia such as Alzheimers disease (AD) vs. frontotemporal dementia (FTD) by providing quantitative localization of metabolic activity. However, analysis of these images beyond visual interpretation is time intensive and not routinely used in clinical settings. Therefore, computational methods that can expedite analysis using more precise quantitative methods are of great interest.

This study was supported in part by the University of Utah Center for Alzheimer's Care and NIH grants AG22394 and AG08671.

An important challenges in automatic diagnosis from PET images is the design of a robust classifier. Brain images contain a very large number of pixels that can be used as attributes in a classifier. On the other hand, the training set size is typically very small. This discrepancy can result in overfitting the classifier to the training data [1]. One approach for addressing this problem is region of interest (ROI) analysis. These methods consider anatomically defined ROIs, e.g. the frontal cortex, to compute a small number of summary measures [2, 3]. One drawback is the requirement of precise prior knowledge of the localization of expected abnormalities for each disease. Abnormalities may also not be unique to an anatomical area and may span portions of multiple ROIs resulting in reduced discrimination power. Principal component analysis (PCA) has also been used to project the image down to a few attributes [4]. However, PCA is not good at capturing complex, non-linear relationships in high-dimensional spaces. Furthermore, PCA is prone to errors if portions of the image data are missing. For instance, parts of the cerebellum can be missing from PET images due to non-optimal patient placement in the scanner. In such cases, these pixels must be manually excluded before PCA can be applied [4].

We propose a new, automated decision tree learning approach which can take into account the spatial distribution of the attributes as well as their information gain during the training phase. This method results in empirically determined cortical regions which offer good discrimination between AD and FTD, and are large enough to be robust to noise and over-fitting problems. Unlike the ROI approaches [2, 3], this method allows the discovery of arbitrarily shaped regions of abnormality and is not limited by anatomical definitions. Furthermore, unlike attributes obtained from PCA [4], which have weighted contributions from all image pixels, the proposed approach clearly defines regions. This is important from the point of view of clinical practice where defining regions that are affected in certain disease is a question of interest. In the context of structural imaging, a method for finding brain regions that exhibit correlated structural changes in a given population and are spatially consistent was introduced

in [5]. Our work focuses on PET images which requires a learning approach focused on changes in image intensity values rather than structural changes. Furthermore, we use a decision tree framework for finding a hierarchy of regions of interest compared to the support vector machine (SVM) classifier used in [5]. The rest of this paper will discuss specifics of the proposed method and validation results using a group of 48 autopsy-confirmed AD and FTD patients.

2. PREPROCESSING

As a preprocessing step, we use the stereotactic surface projection (SSP) package [6] to warp images into the common Talairach coordinate system [7] allowing for pixel-by-pixel comparisons among the group of subjects. Since AD and FTD are diseases of the gray matter, we are concerned with metabolic activity in the gray matter rather than the entire brain. SSP also extracts metabolic activity in the cerebral cortex using a predefined list of approximately 16000 brain surface pixels. Finally, since each patient has a different global metabolic rate, and the scaling of the image values depends on the specific scanner used, the data needs to be normalized. Normalization using global average metabolism is problematic because it is affected by disease. We normalize by the metabolic activity in the pons located on the brain stem which is known to be relatively spared by AD and FTD [6, 4].

3. METHODS

Even after the extraction of locations limited to the cerebral cortex, there are still a very large number of pixels that can be used as attributes in a classifier. If a learning algorithm chooses attributes based solely on their discriminatory power between AD and FTD, the resulting classifier is likely to use only a few pixels and over-fit to the training data. We describe a method for locating areas of the cerebral cortex, i.e. collections of pixels, that are as large as possible without a significant loss of discriminative power.

3.1. Metabolic activity in local regions

Let $\mathcal{C}$ denote the set of pixel locations on the cerebral cortex. Ideally, we want to choose the largest subset of $\mathcal{C}$ which also has maximum discrimination power between the disease classes of interest. This problem is computationally intractable due to the large number of combinatorial possibilities that arise in defining a subset. However, we can easily find approximations to maximally large regions with high discrimination power gain by taking into account the spatial distribution of the pixels in $\mathcal{C}$. Given a location $x \in \mathcal{C}$ and a distance threshold R, we define

$$S(x, R) = \{x' \in \mathcal{C} : |x - x'| \leq R\}, \quad (1)$$

which is the local cortical region centered at x. The total metabolic activity in region $S(x, R)$ is computed as

$$A(x, R) = \sum_{x' \in S(x,R)} M(x'), \quad (2)$$

where $M(x')$ is the normalized metabolic activity at pixel x'.

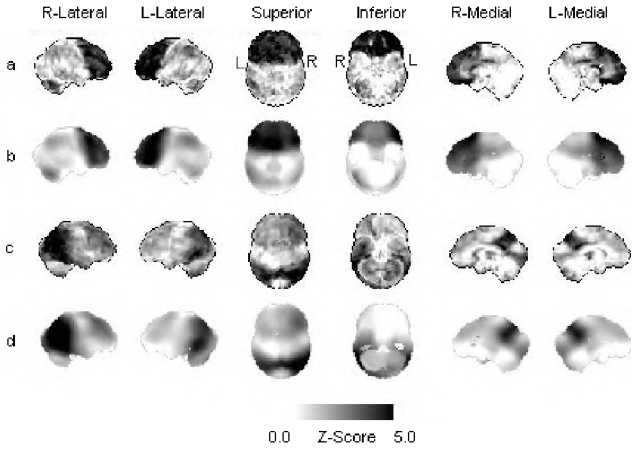

Fig. 1. Z-scores for a FTD case computed with (a) R=0, (b) R=8, and for an AD case with (c) R=0 and (d) R=8 shown as surface projections. Often the distinction between AD and FTD is not as apparent as shown here; some patients who are affected with AD also show reduction of metabolism in frontal and temporal regions of the brain.

While $A(x, R)$ can be used directly as an attribute, metabolic values are typically converted to z-scores using a database of normals [4]. Z-scores measure how many standard deviations away a given attribute value is from its expected mean and allow a more standardized comparison. For instance, z-scores larger than 3 are generally deemed significant regardless of the specific application. Given a group of normal controls, we compute the mean $\mu(x, R)$ and standard deviation $\sigma(x, R)$ of $A(x, R)$ at all $x \in \mathcal{C}$. Then the z-score at location x for the *k'th* subject is computed as

$$Z_k(x, R) = \left(\mu(x, R) - A_k(x, R)\right) / \sigma(x, R). \quad (3)$$

Figure 1 illustrates the attributes $Z_k(x, R)$ for an AD case and a FTD case in the form of six surface projections: right/left lateral, superior/inferior and right/left medial. Z-scores computed with $R = 0$ correspond to single pixel regions as defined by Equation (1) and are shown in Figure 1 (a) and (c). Z-scores computed with $R = 8$, Figure 1 (b) and (d), illustrate the spatial smoothing effect of using local cortical regions. While some information is lost, a classifier that is less susceptible to over-fitting can be constructed by using attributes computed over these larger regions. Furthermore, comparing areas rather than comparing individual pixels reduces the effects of any registration error in warping of the brain to the

Talairach coordinate system. In fact, the parameter R plays an important role in the performance of the algorithm as will be discussed in Section 4.

3.2. Region growing with information gain

The local cortical regions are defined in Section 3.1 as collections of pixel on the cortical surface that are within a certain distance of a central point. While attributes computed over these regions are expected to be more robust compared to attributes computed from single pixels, further improvements can be realized by finding the most discriminative such regions and merging them into a larger, arbitrarily shaped region. We describe this process next.

Areas of significant hypometabolism in a subject are reflected by high z-score values. Therefore, we can create binary attributes $Y(x, R)$ by thresholding the corresponding Z-score variables $Z(x, R)$ defined in Section 3.1. Then, the value of $Y_k(x, R)$ determines whether pixel x in the *k'th* subject is considered abnormal. In the context of classification, information gain measures how well a given attribute separates the training examples into their target classes. Each binary attribute $Y(x, R)$ separates a set of training examples Ω into two mutually exclusive and exhaustive subsets $\Omega_t = \{k : Y_k(x, R) = true\}$ and $\Omega_f = \{k : Y_k(x, R) = false\}$. Then information gain for $Y(x, R)$ given the set Ω is defined as

$$G(x, R, \Omega) = H(\Omega) - \left(\frac{|\Omega_t|}{|\Omega|} H(\Omega_t) + \frac{|\Omega_f|}{|\Omega|} H(\Omega_f) \right), \quad (4)$$

where $H(\Omega)$ and $|\Omega|$ denote the entropy and the size of set Ω, respectively. Let $P_\Omega(l)$ denotes the fraction of examples which belong to class l in set Ω, then entropy is defined as

$$H(\Omega) = - \sum_{l=AD,FTD} P_\Omega(l) \log_2 P_\Omega(l). \quad (5)$$

A set has a minimum entropy of 0 if all its members belong to the same class. We compute the information gain for all the attributes $Y_k(x, R)$ and find the maximum information gain, G_{max}. Then we can define a new arbitrarily shaped region that has near optimal information gain as

$$N(R, \Omega) = \cup_{x:G(x,R,\Omega) \geq \gamma G_{max}} S(x, R) \quad (6)$$

where γ is a parameter that determines how close $Y_k(x, R)$'s information gain has to be to be included in the larger region. A typical value for γ is 0.95. Similar to $A(x, R)$ in Equation (2), the total metabolic activity for the new region is then computed as a sum over $N(R, \Omega)$ and is also thresholded at T to generate a binary-valued attribute.

3.3. Decision tree learning

We use the ID3 algorithm to learn a decision tree classifier in a supervised setting [1]. Given a training set Ω, we construct an attribute as described in Section 3.2 to use at the root node of the decision tree. This attribute splits Ω into two subsets. If a subset has examples of both classes (AD and FTD), a new intermediate node is created and an attribute is selected in the same manner as the root node, but using only the training examples in that subset. The decision tree learning stops when all nodes have only one class associated with them.

4. RESULTS AND VALIDATION

We tested the proposed method on a set of 48 autopsy confirmed cases (34 AD, 14 FTD). A normal control group of 33 subjects was used for computing z-scores. Figure 2 illustrates the decision tree learned using all 48 cases in the training set. This tree has three decision nodes, each associated with an arbitrarily shaped region found using the methods discussed in Section 3. The region shown in Figure 3(a) corresponds to the posterior temporo-parietal cortex which is greatly affected in AD. Notice that *node 1* (the root node) of the decision tree in Figure 2 classifies subjects as having AD if the hypometabolism in this region is significant ($z > 3$). The regions shown in Figure 3(b) and (c) correspond to the frontal and anterior temporal areas, typically affected to a larger extent in FTD than in AD. As expected, the decision tree classifies subjects as having FTD if there is significant hypometabolism in either of these regions. The regions selected by the proposed approach conforms to our expectations from a neurological point of view to a large extent. However, it was not immediately clear why the algorithm showed a preference for right vs. left anterior temporal regions. One explanation is the small sample size that is left after two previous branchings of the tree. This small sample size might not be able to support a robust decision resulting in over-fitting; hence a neurologically unexplained preference for right vs. left anterior temporal regions.

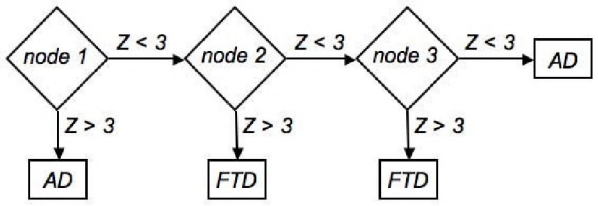

Fig. 2. Decision tree for R=8 and z-score threshold T=3.

Leave-one-out cross validation was performed for evaluating the accuracy of the proposed approach on the set of 48 autopsy confirmed cases. One of the 48 subjects is left out of the training set and the resulting classifier is tested on this left-out subject, the experiment is repeated 48 times leaving each subject out once. The accuracy of the algorithm was evaluated for different settings of the neighborhood radius (R) and z-score threshold (T) parameters, see Table 1. Notice that R=0

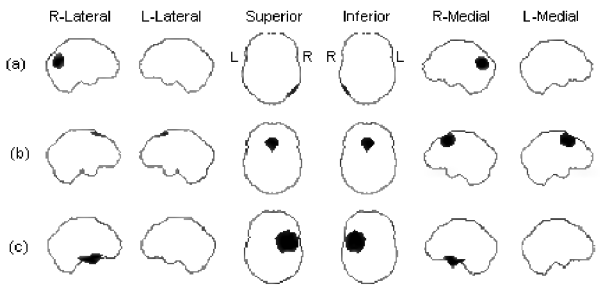

Fig. 3. Regions corresponding to the decision tree nodes in Figure 2: (a) *node 1*, (b) *node 2* and (c) *node 3*.

	R=0	R =4	R=8	R=12	R=16
D. tree T=2	75%	71%	78%	67%	75%
D. tree T=3	81%	90%	94%	88%	80%
D. tree T=4	87%	92%	73%	78%	59%
Boosting T=3	87%	83%	90%	83%	77%

Table 1. Classification accuracy as a function of R and T.

corresponds to starting from single pixels and as expected performs poorly because the resulting regions can fit the training data exactly but fail to generalize well to test cases (overfitting). On the other hand, using very large R diminishes the discriminative power of the resulting regions which can not fit the boundaries of the cortical areas that actually separate AD and FTD cases. Analyzing the performance of the algorithm with respect to the threshold parameter, we observe that $T = 2$ performs poorly because it was not selective enough in terms of difference from normal controls (too many binary attributes with value 1). Choosing $T = 3$ gave the best result when $R \geq 8$ while $T = 4$ was a better choice for $R = 0, 4$. This can be explained by observing that $T = 4$ is overly selective (very few binary attributes with value 1) and performs poorly for large regions which almost never have such high z-scores. To summarize, the choice $T = 3$ offers the best compromise with the particular choice of $T = 3, R = 8$ giving the best overall result. We also experimented with automatically choosing optimal T values for each node of the decision tree based on maximum information gain criteria as is typically done when using decision trees with continuous-valued attributes. However, the results were worse than fixing the threshold at $T = 3$. This is probably due to over-fitting problems that arise from the extra degrees of freedom introduced by this approach.

We compared our spatial decision tree learning approach to using *Adaboost* [8] with classifiers that use the the local cortical region attributes defined in Section 3.1. The results for boosting using regions with various radii and $T = 3$ are shown in the last row of Table 1 and, in general, are worse than the results obtained with the approach introduced in this paper. Our results can also be compared to previously reported experiments in the literature with the same autopsy confirmed dataset [4]. The visual rating of six neurologists using the SSP z-score images was reported to be 89%. Using principal component analysis (PCA) to project the data to a lower-dimensional space followed by learning a linear classifier in this space gave results in the range 80-85% depending on the dimensionality used in PCA. Slightly better results, as high as 90%, were obtained using Partial Least Squares, which takes into account target classifications [4].

5. CONCLUSIONS

We have discussed a method for the automatic classification of brain images of AD and FTD subjects. The dynamic region selection of decision tree based on the information gain is a powerful tool, especially when compared to ROI analysis which requires prior knowledge. This method also makes it possible to locate areas of abnormalities for specific type of dementias, something that is not easily seen in PCA or in other classifiers such as neural networks. Furthermore, due to the grouping of adjacent pixels, it is less sensitive to overfitting or image registration errors. The results obtained using the proposed method are very encouraging in classifying FTD and AD subjects. We are planning to do perform further validation on other AD/FTD databases. Another research direction is to apply the same paradigm to other diseases which are hard to diagnose clinically due to similar symptoms.

6. REFERENCES

[1] T. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[2] H. R. Hooper, A. J. McEwan, B. C. Lentle, T. L. Kotchon, and P. M. Hooper, "Interactive three-dimensional region of interest analysis of HMPAO SPECT brain studies," *J Nucl Med*, vol. 31, no. 12, pp. 2046–2051, 1990.

[3] P. Charpentier *et. al*, "Alzheimer's disease and frontotemporal dementia are differentiated by discriminant analysis applied to 99mTc HmPAO SPECT data," *J Neurol Neurosurg Psychiatry*, vol. 69, pp. 661–663, 2000.

[4] Higdon *et. al*, "A comparison of classification methods for differentiating fronto-temporal dementia from Alzheimer's disease using FDG-PET imaging," *Statistics in Medicine*, vol. 23, pp. 315–326, 2004.

[5] Y. Fan and D. Shen C. Davatzikos, "Classification of structural images via high-dimensional image warping, robust feature extraction, and svm," in *MICCAI, LNCS*, 2005, vol. 3749, pp. 1–8.

[6] S. Minoshima, A. F. Kirk, R. A. Koeppe, N. L. Foster, and D. E. Kuhl, "A diagnostic approach in Alzheimer's disease using three-dimensional streeotactic surface projections of fluorine-18-FDG PET," *J Nucl Med*, vol. 36, pp. 1238–1248, 1995.

[7] J. Talairach and P. Tournoux, *Co-planar Stereotaxic Atlas of the Human Brain*, Thieme Medical Publishers, 1988.

[8] Yoav Freund and Robert E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *E. Conf on Computational Learning Theory*, 1995, pp. 23–37.