Classification of Alzheimer's Disease Using a Self-Smoothing Operator

Juan Eugenio Iglesias, Jiayan Jiang, Cheng-Yi Liu, Zhuowen Tu, and the Alzheimers Disease Neuroimaging Initiative

Laboratory of Neuro Imaging, University of California, Los Angeles
{jeiglesias,chengyiliu}@ucla.edu, {jet.jiang,zhuowen.tu}@loni.ucla.edu

Abstract. In this study, we present a system for Alzheimer's disease classification on the ADNI dataset [1]. Our system is able to learn/fuse registration-based (matching) and overlap-based similarity measures, which are enhanced using a self-smoothing operator (SSO). From a matrix of pair-wise affinities between data points, our system uses a diffusion process to output an enhanced matrix. The diffusion propagates the affinity mass along the intrinsic data space without the need to explicitly learn the manifold. Using the enhanced metric in nearest neighborhood classification, we show significantly improved accuracy for Alzheimer's Disease over Diffusion Maps [2] and a popular metric learning approach [3]. State-of-the-art results are obtained in the classification of 120 brain MRIs from ADNI as normal, mild cognitive impairment, and Alzheimer's.

1 Introduction

Alzheimers Disease (AD) and its preclinical stage, mild cognitive impairment (MCI), are the most common form of dementia in elders. Magnetic resonance imaging (MRI) can provide insight into the relation between AD and the structure of the brain: AD is known to be connected with gray matter loss [4] and with the shape of subcortical structures (especially the hippocampus) [5]. There have been several attempts in the literature to automatically classify a brain MRI as AD, MCI or normal (typically represented by older control subjects, OC). *Chupin et al.* [6] automatically segment the hippocampus and use its volume for the classification. *Vemuri et al.* [7] use support vector machines (SVM) based on tissue densities and a number of covariates (demographics, genotype). *Klöppel et al.* [8] feed a SVM directly with image data after registration (i.e. spatial alignment). *Zhang et al.*[9] use a SVM with cerebrospinal fluid, positron emission tomography and MRI data as features. *Davatzikos et al.* [10] use the distribution of gray matter, white matter and cerebrospinal fluid in registered space. *Desikan et al.* [11] feed the entorhinal cortex thickness, hippocampal volume and supramarginal gyrus thickness to a logistic regression analysis. These works are summarized in Table 1.

In this paper, we approach the OC/MCI/AD classification problem from the perspective of metric learning. Given a number of heterogeneous affinity measures between the data points, the task is to find an enhanced metric which will

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part III, LNCS 6893, pp. 58–65, 2011.

Table 1. Representative methods in the literature of AD classification. For *Chupin et al.* we report the range of results in a number of two-class classification problems.

Method	# Subjects	Classes (prevalences)	Classification rate
Chupin et al. [6]	605	OC (24%), MCI (49%), AD(27%)	60-80%
Vemuri et al. [7]	100	OC (50%), AD (50%)	89%
Klöppel et al. [8]	68	OC (50%), AD (50%)	94%
Zhang et al.[9]	103	OC(50%), AD(50%)	93%
Zhang et al.[9]	150	OC(34%), MCI(66%)	76%
Davatzikos et al. [10]	30	OC (50%), MCI (50%)	90%
Desikan et al. [11]	151	OC (62%), MCI (38%)	90%

ultimately improve the classification rate in a k-nearest neighbor (kNN) framework. Popular distance metric learning methods [12,3], which are mostly supervised, learn a Mahalanobis distance parametrized by a positive semi-definite matrix. However, the performance gain is rather limited because a global linear transform does not suffice to discriminate the data. Nonlinear versions exist, but it is difficult to find a kernel that provides good results. Non-parametric manifold learning techniques such as Isomap [13] do not necessarily provide a better metric, which limits their use in classification. They also have the disadvantage that explicitly estimating the manifold can be difficult and time consuming. Their application to medical image analysis has also been limited [14].

Here we adopt an unsupervised metric learning algorithm: self-smoothing operator (SSO). SSO enhances an input pair-wise affinity matrix similar to a Gram matrix. A smoothing kernel is built from the matrix and used to iteratively propagate the affinity mass between strongly connected neighbors, following the structure of the manifold without having to compute it explicitly. The framework can accommodate semi-supervise learning (i.e. taking advantage of not only labeled but also unlabeled examples to build a classifier [15]): even if unlabeled examples cannot be used in the kNN classification, they can still be considered in the prior diffusion, often bridging gaps between points with the same label. A feature selection method is incorporated into the design of the affinity matrix to improve the results. We apply the proposed framework to the AD classification problem with registration-based and overlap-based similarity measures, comparing the results with metric learning [3] and Diffusion Maps [2].

2 Materials

Brain MRI from 120 subject Brain MRI scans from 120 subjects (age 76.7±6.4 years) are used in this study. The subjects were randomly selected from the ADNI dataset [1] under two constraints: 1) the scans are from the same cross section (12 months after the start of the study); and 2) the three classes (OC,MCI,AD) and the two genders are equally represented. The scans were acquired with T_1-weighted MPRAGE sequences, skull-stripped with BET [16] and fed to BrainParser [17] to automatically extract 56 cortical and subcortical structures.

3 Methods

3.1 Self-Smoothing Operator

SSO is closely related to the Diffusion Maps algorithm [2], which defines diffusion distances between data samples to improve an input pair-wise affinity matrix. It introduces a global diffusion distance metric over data samples. Given the transition kernel H (a row-wise-normalized version of the pair-wise affinity matrix), the diffusion distance between data samples x_i and x_j at step t is defined as:

$$d_t^2(i,j) = ||h_t(i,\cdot) - h_t(j,\cdot)||^2_{1/\phi_0} = \sum_k \frac{1}{\phi_0(k)}(h_t(i,k) - h_t(j,k))^2$$

where $h_t(i,\cdot)$ is the i-th row of H^t, and ϕ_0 is the equilibrium distribution.

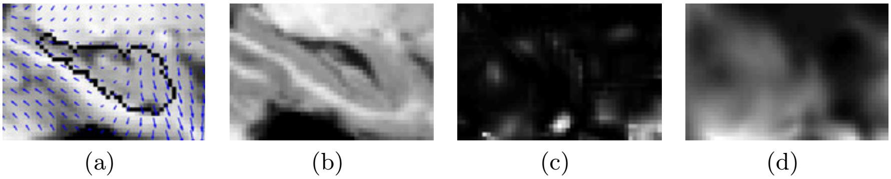
Instead of using an alien notion of diffusion distances between data samples as in Diffusion Maps, we work on the affinity matrix directly, using a self-induced smoothing kernel. Given data samples $\{x_i, \ldots, x_n\}$ and a symmetric affinity function $\vartheta(x_i, x_j) = \vartheta(x_j, x_i) \in [0,1]$, we define the $n \times n$ weight matrix W as $W(i,j) = \vartheta(x_i, x_j)$. SSO diffuses the pair-wise affinities of W along the geometry of the manifold without having to construct it explicitly:

1. Create the diffusion matrix $P = D^{-1}W$, where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^n W(i,j)$.
2. Self diffusion: $W^{(p)} = WP^p$.

In Step 1, the smoothing kernel P that governs the diffusion process in Step 2 is induced from the input similarity matrix. $W^{(p)}$ is not a proper Gram matrix (since $\vartheta(x_i, x_j)$ is not an inner product), so it is in general neither symmetric nor positive semi-definite (PSD), which is not a problem in this application: we simply take the k minimal non-diagonal values of each row as the k nearest neighbors for classification. The only parameter in the algorithm is the step p, which determines the scale at which the data are analyzed. The output $W^{(p)}$ is an updated weight matrix that represents similarity more faithfully than W (as experimentally shown below) and that can be used directly in classification.

3.2 Similarity/Divergence Measures

The affinities $W(i,j)$ can be built from a similarity or divergence function, γ, using a suitable transform. If several $\{\gamma_m\}$, $m = 1, \ldots, M$, are available, W can be a linear combination: $\bar{W} = w_m \sum_{m=1}^M W_m^{(p)}(\gamma_m)$, with $\sum_{m=1}^M w_m = 1$. We use two types of measures in this study: overlap- and registration-based. The first type is based on the Dice overlap of the structures of interest, giving a rough estimate of how similar two binary masks are. First, the centroids of the two instances of the structure to compare (s) are aligned, yielding $Y_{s,i}$ and $Y_{s,j}$. Then, the Dice overlap is computed as: $O(Y_{s,i}, Y_{s,j}) = \frac{2|Y_{s,i} \cap Y_{s,j}|}{|Y_{s,i}|+|Y_{s,j}|} \in [0,1]$. One minus the overlap would be a valid affinity. However, in order to enhance



(a) (b) (c) (d)

Fig. 1. (a) Slice of a sample volume, cropped around the hippocampus. The automatic hippocampal segmentation is outlined in black. The deformation field provided by the registration towards the slice in (b) is superimposed in blue. Summing the curvature (c) and diffusion (d) within the segmentation provides the divergences γ_{curv} and γ_{diff}.

the differences between good and bad matches, we linearly map the interval $[O_{s,min}, 1]$ to $[0, 1]$:

$$W_{over,s}(i,j) = 1 - \frac{O(Y_{s,i}, Y_{s,j}) - O_{s,min}}{1 - O_{s,min}} = \frac{1 - O(Y_{s,i}, Y_{s,j})}{1 - O_{s,min}}$$

A divergence function complementing the Dice coefficient should consider non-linear deformations. Here we use a diffeomorphic registration algorithm [18] to estimate the degree of warping that is required to deform a shape into another. To compare brains i and j, we first register j to i. Then, for structure of interest s, we compute the irregularity of the obtained deformation field $\boldsymbol{u}_{ji}(\boldsymbol{r})$ within the mask $\Omega_{s,i}$ corresponding to s in i ($\boldsymbol{r}$ is the location vector). We use the curvature and diffusion of $\boldsymbol{u}_{ji}(\boldsymbol{r})$ as measures of irregularity:

$$\gamma_{curv}[\boldsymbol{u}(\boldsymbol{r})] = \int_{\Omega_{s,i}} \sum_{\{x,y,z\}} [\triangle u_d(\boldsymbol{r})]^2 d\boldsymbol{r}, \quad \gamma_{diff}[\boldsymbol{u}(\boldsymbol{r})] = \int_{\Omega_{s,i}} \sum_{\{x,y,z\}} \|\nabla_d(\boldsymbol{r})\|^2 d\boldsymbol{r}$$

where the index d loops along the three spatial dimensions. The deformation field for a sample case is shown in Fig. 1a. The integrands γ_{curv} and γ_{diff} are displayed in Fig. 1c and 1d. Finally, the corresponding weight matrices can be computed using a Gaussian function as follows:

$$W_{[\cdot]}(i,j) = \exp\Big(- \big(\gamma_{[\cdot]}\left[\boldsymbol{u}_{ij}\left(\boldsymbol{r}\right)\right] + \gamma_{[\cdot]}\left[\boldsymbol{u}_{ji}\left(\boldsymbol{r}\right)\right]\big)^2 / \left(\mathrm{var}\left(\gamma_{[\cdot]}\right)\right)\Big)$$

where $[\cdot]$ refers to curvature or diffusion, and $\mathrm{var}(\gamma_{[\cdot]})$ is the variance of the divergence γ across the dataset. The weights $W_{[\cdot]}$ are explicitly symmetrized.

3.3 Feature Selection

Assuming that the global weight matrix $\bar{W}$ is a linear combination of matrices based on single features (divergences or similarities), the question is which combination of weights $\boldsymbol{w} = \{w_m\}$ to use. Specifically, we seek to maximize the leave-one-out (LOO) classification rate $\Psi(\boldsymbol{w})$ under the constraints: $0 \leq \boldsymbol{w} \leq 1$ and $\mathbb{1}^t \boldsymbol{w} = 1$. This problem is difficult to solve because Ψ is neither smooth nor convex, and has multiple local maxima. Instead, we further constrain the problem by assuming that only $M' \leq M$ weight matrices are used with equal weights $w_m = M'^{-1}, \forall m$. Then the problem becomes analogous to that of feature selection in machine learning. This is still a hard combinatorial problem, but good approximate solutions can be achieved using a proper selection strategy. Here we use "plus 2 - take away 1" [19]: from an initial empty set, features are greedily added / removed one at the time following the pattern $+, +, -, +, +, -, \ldots$. The final set of features is the one that maximizes $\Psi(\boldsymbol{w})$ throughout the process.

4 Experiments and Results

4.1 Experimental Setup

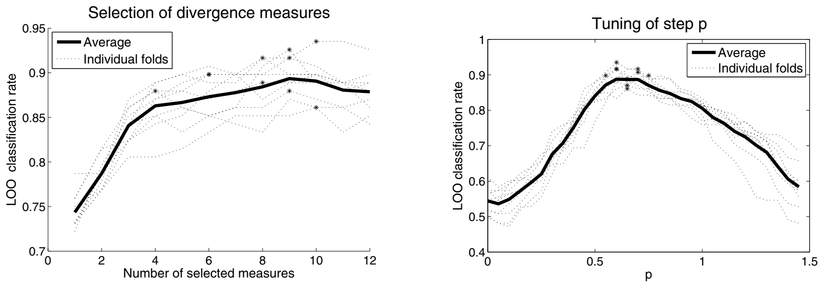
The feature selection was cross-validated (10 folds) to obtain an unbiased estimate of the performance; otherwise features are selected upon the test data. For each fold, a set of features is selected with LOO on the training data. For each candidate set, the scale of the diffusion p is tuned individually using exhaustive search. The selected features and p are used to classify the test data in the fold. The number of neighbors was kept constant ($k = 10$) to limit the computational load of training. Ties are broken by examining subsequent neighbors. Rather than using all the 56 segmented structures in the selection process, only the (left and right) caudate nucleus, hippocampus and putamen are considered (18 features in total). These structures are well-known to be related to AD, and using this reduced set decreases the risk of overfitting.

In testing, an augmented W is created by adding to the original a new row and column for each test sample. We assume that all the test data are simultaneously available, which enables semi-supervised learning: during the diffusion process, the unlabeled test data can increase the performance of the system by making the structure of data easier to follow (only the labeled training data are considered during the kNN classification).

For the sake of comparison, analogous experiments were run using Diffusion Maps and the metric learning approach from [3], which attempts to find the positive definite matrix A that parameterizes the Mahalanobis distance best separating the training data into the different classes. Cross validation was again performed with 10 folds using the training data to select features (same selection strategy) and tune parameters: the matrix A for metric learning and the step t for Diffusion Maps (i.e. the scale of the diffusion). As for SSO, the number of nearest neighbors was fixed throughout the experiments ($k = 10$).

4.2 Results

The impact of feature selection on the performance is illustrated in Fig. 2a. The three most frequently selected features were: 1) diffusion - left hippocampus; 2) curvature - left putamen; and 3) overlap - left caudate. It is not surprising that the top feature is related to the hippocampus, which is known to be strongly connected with AD [5]. The curve in Fig. 2b shows the impact of the diffusion



(a) (b)

Fig. 2. (a) Best LOO classification rate against number of selected features. (b) For the feature subset chosen for each fold: dependence of the classification rate on the diffusion step p. The point $p = 0$ corresponds to the classification rate with no diffusion. In both graphs, the chosen operating points for each fold are marked with asterisks.

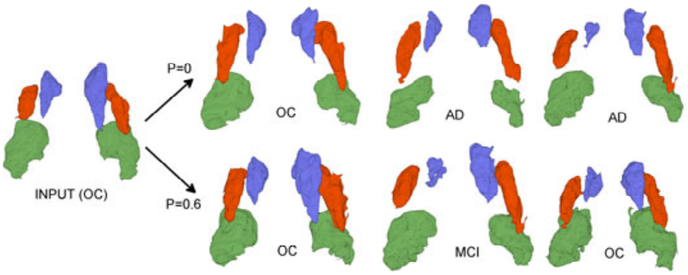


Fig. 3. Three-dimensional rendering of the structures of interest of an OC sample and its nearest neighbors, before ($p = 0$) and after diffusion ($p = 0.6$). The hippocampi are rendered in green, the putamens in red and the caudate nuclei in blue. The diffusion bridges the gaps with other OCs, moving the AD and MCI cases farther away.

on the classification. At first, increasing the scale of the diffusion p has a positive influence on the accuracy, which is boosted from $\sim 55\%$ at $p = 0$ (no diffusion) to $\sim 90\%$ at $p \approx 0.6$. This is illustrated with a sample subject and its nearest neighbors before and after diffusion in Fig. 3. When p becomes too large, data samples start to come too close to one another and the accuracy begins to decrease. Fortunately, the location of the peak is quite stable and the method generalizes well, as shown by the cross validation experiment below.

Tables 2a through 2c display the confusion matrices for metric learning, Diffusion Maps and our approach, respectively. Metric learning performs poorly because the structure of the data is too complex to discriminate the classes using a global linear transform. Diffusion Maps provides decent results: 78% accuracy with no mistakes between OC and AD. Our SSO-based approach makes no OC-AD mistakes either, but preserves the structure of the input similarity better than Diffusion Maps, increasing the accuracy to 89%. There is no noticeable drop in accuracy from the training data (Fig. 2) because cross-validation (LOO) was already used within the feature selection process.

Even though the results reported in Table 1 were achieved on other datasets, it is illustrative to compare them to ours. *Chupin et al.'s* study, the only one considering the three-class problem, reports considerably lower accuracy than this work. To compare our results with the methods which classify OC vs. AD, we assume that only OC / AD are fed to the classifier and that the samples classified as MCI are relabeled to either OC or AD. Another option would be to remove the MCI cases from the training data, but that would have a negative impact on the results (the diffusion would be guided by less data). Our approach provides 96.25% or 97.5% accuracy (depending on the relabeling criterion), slightly higher than the best reported results in the literature (*Klöppel et al.*, 94%). In order to compare our approach with methods that discriminate OC from MCI, we assume that only OC and MCI cases are fed to the classifier, and the cases for which the estimated class is AD are relabeled as MCI. In that case, the accuracy is 91.25%, comparable to *Davatzikos et al.* and *Desikan et al.* (90%).

Table 2. Confusion matrices: (a) metric learning, (b) Diffusion Maps, (c) the proposed method. The global accuracies are 50%, 78% and 89%. GT stands for "ground truth".

(a)

GT\Method	OC	MCI	AD
OC	20	11	9
MCI	14	18	8
AD	2	16	22

(b)

GT\Method	OC	MCI	AD
OC	29	11	0
MCI	5	29	6
AD	0	4	36

(c)

GT\Method	OC	MCI	AD
OC	38	2	0
MCI	5	32	3
AD	0	3	37

5 Discussion and Future Work

A nearest neighbor classifier based on registration and overlap features and enhanced by a self-smoothing operator has been presented in this study. SSO propagates the similarity between data samples along the manifold in which the data lie. The updated affinity measure can be used in a nearest neighbor framework to classify brains as AD, MCI or OC, achieving state-of-the-art results. The main disadvantage of the method is that, when a new case is presented to the system, computing the corresponding new row in the affinity matrix requires nonrigid registration to all the training cases, which is very time consuming (the SSO algorithm itself only takes a fraction of a second). Exploring its application to other disease patterns, testing features that are faster to compute and improving the design and combination of features remain as future work.

Acknowledgments. This work was funded by NSF career award IIS-0844566 and ONR grant N000140910099.

References

1. Alzheimer's Disease Neuroimaging Initiative, `http://adni.loni.ucla.edu`
2. Coifman, R., Lafon, S.: Diffusion maps. Applied and Computational Harmonic Analysis 21(1), 5–30 (2006)
3. Davis, J., Kulis, B., Jain, P., Sra, S., Dhillon, I.: Information-theoretic metric learning. In: Proceedings of ICML, pp. 209–216 (2007)
4. Karas, G., et al.: A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. Neuroimage 18(4), 895–907 (2003)
5. Du, A., et al.: Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. Journal of Neurology, Neurosurgery & Psychiatry 71(4), 441–447 (2001)
6. Chupin, M., et al.: Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19(6), 579–587 (2009)
7. Vemuri, P., et al.: Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. Neuroimage 39(3), 1186–1197 (2008)
8. Klöppel, S., et al.: Accuracy of dementia diagnosis–a direct comparison between radiologists and a computerized method. Brain 131, 2969–2974 (2008)
9. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. NeuroImage 55(3), 856–867 (2011)
10. Davatzikos, C., et al.: Detection of prodromal Alzheimer's disease via pattern classification of MRI. Neurobiology of Aging 29(4), 514–523 (2008)
11. Desikan, R., et al.: Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. Brain 132(8), 2048 (2009)
12. Xing, I., Ng, A., Jordan, M., Russell, S.: Distance metric learning with application to clustering with side-information. Advances in Neural Information Processing Systems, 505–512 (2002)
13. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323
14. Gerber, S., Tasdizen, T., Fletcher, P., Joshi, S., Whitaker, R.: Manifold Modeling for Brain Population Analysis. Medical Image Analysis 14(5), 643–653 (2010)
15. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
16. Smith, S.: Fast robust automated brain extraction. Human Brain Mapping 17(3), 143–155 (2002)
17. Tu, Z., et al.: Brain anatomical structure segmentation by hybrid discriminative/generative models. IEEE Transactions on Medical Imaging 27(4)
18. Avants, B., Epstein, C., Grossman, M., Gee, J.: Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12(1), 26 (2008)
19. Stearns, S.: On selecting features for pattern classifiers. In: Proc. of the 3rd Int. Joint Conf. on Pattern Recognition, pp. 71–75