Published in final edited form as:

Proc IEEE Int Symp Biomed Imaging. 2011 ; : 1101–1104. doi:10.1109/ISBI.2011.5872593.

UNDERSTANDING HETEROGENEITY IN NORMAL OLDER ADULT POPULATIONS VIA CLUSTERING OF LONGITUDINAL DATA

Roman Filipovych*, Susan M. Resnick†, and Christos Davatzikos*

*Section of Biomedical Image Analysis, Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104

†Laboratory of Personality and Cognition, Biomedical Research Center/04B317, 251 Bayview Blvd., Baltimore, MD 21224

Abstract

Populations of healthy older individuals are often highly heterogeneous, as prevalence of various underlying pathologies increases with age. Finding coherent groups of normal older adults may allow to identify subpopulations that are at risk of developing Alzheimer's disease (AD). In this paper, we propose an approach that utilizes longitudinal magnetic resonance imaging (MRI) data to obtain natural groupings of older adult subjects via an unsupervised (i.e., clustering) technique. We develop a k -medoids-like clustering algorithm that simultaneously finds clusters of longitudinal images, as well as weights brain regions in such a way that the obtained clusters are maximally coherent. We propose a cluster-based measure that reflects the individual subject's cognitive decline. The proposed method is unsupervised and is suitable for analyzing AD at its very early stages.

Index Terms

Alzheimer's; MRI; Mild Cognitive Impairment; Cluster Analysis; Longitudinal Image Analysis

1. INTRODUCTION

Assigning images into groups with respect to meaningful characteristics is of significant importance in many applications. Supervised classification approaches aim at deducing a decision function from the labeled training data. Significant work has been done in applying pattern classification algorithms in the context of Alzheimer's disease (AD) and mild cognitive impairment (MCI) [1, 2, 3]. In the supervised analysis, subjects are usually grouped under a common umbrella corresponding to high-level clinical categories (e.g., patients and controls). However, the population of healthy adults is highly heterogeneous as prevalence of various underlying pathologies, including amyloid plaques, neurofibrillary tangles, and cerebrovascular lesions, increases with age. Moreover, the rate of cognitive decline was shown to reflect the risk of developing MCI and AD, and is different for different subjects [4].

Clustering algorithms find natural groupings in the data, and present a promising technique for disentangling heterogeneity that is inherent to many diseases. A guided clustering algorithm was recently proposed in [5], and was applied to a study of normal older adult populations. It was found that cognitively stable subjects form tight clusters, while the subpopulation of cognitively declined subjects is less coherent. However, the approach in [5] does not perform grouping of longitudinal scans. Additionally, while the cluster-based

biomarker from [5] correlates with cognitive measurements, it is preferable to have a measure that is associated with rates of change in cognitive scores, as the rate of change in cognitive evaluations is a more robust measure of cognitive decline.

In this paper we present an approach for clustering longitudinal scans of normal older adults. By extracting image features from a set of predefined regions we transform longitudinal images into multidimensional sequences (i.e., trajectories). Our method uses an adapted version of a robust sequence matching algorithm [6] to establish the notion of similarity between the longitudinal data of any pair of subjects. We then develop a k -medoids-like clustering algorithm that simultaneously finds clusters of longitudinal images, as well as weights brain regions in such a way that the obtained clusters are maximally coherent. Finally, we propose a cluster-based measure that reflects the individual subject's cognitive decline, and correlates well with the clinical data. Our approach is unique in the sense that it allows to obtain a computational biomarker of cognitive decline without relying on the availability of predefined categories in the population.

2. METHOD

2.1. Longitudinal scans as multidimensional trajectories

A principal challenge in medical image analysis is attributed to the very high dimensionality of the data and relatively small sample size. By registering brain images to a common template with a predefined number of labeled anatomical regions of interest (ROIs), and by obtaining a measurement at each ROI, the dimensionality of the original data can be reduced to a manageable size. More formally, imaging evaluations of a given subject can be represented as a multidimensional trajectory $\mathcal{S} = ((s_{t,r}))$, where $t = 1, \dots, T$ denotes the t -th imaging evaluation, and $r = 1, \dots, R$ indicates the ROI at which the measurement was taken. The number of dimensions in $\mathcal{S}$ corresponds to the number of ROIs, while the length of the trajectory corresponds to the number of imaging evaluations. Figure 2 provides a schematic example of a multidimensional trajectory representation of subject's imaging evaluations.

2.2. Similarity measure based on trajectory matching

The task of comparing longitudinal imaging data involves several major challenges. First, some subjects may have different number of imaging evaluations, which would make comparison of longitudinal data on an image-by-image basis inappropriate. Second, due to flaws in image acquisition and processing, the imaging measurements may be very noisy. We address these issues by using the robust Least Common Subsequence (LCSS) matching scheme introduced in [6]. It is specifically designed to disregard trajectory points that are associated with more-than-tolerable noise. For two high-dimensional trajectories $X = (x_1, \dots, x_p)$ and $Y = (y_1, \dots, y_q)$, and for an integer δ and a real number $0 < \epsilon < 1$, the LCSS scheme can be defined as follows:

$$LCSS_{\delta,\epsilon}^{p,q} = \begin{cases} 0 & \text{if } p=0 \text{ or } q=0 \\ 1 + LCSS_{\delta,\epsilon}^{p-1,q-1} & \text{if } \|x_p, y_q\| < \epsilon \text{ and } |p - q| \leq \delta \\ \max (LCSS_{\delta,\epsilon}^{p-1,q}, LCSS_{\delta,\epsilon}^{p,q-1}) & \text{otherwise.} \end{cases} \quad (1)$$

Here, $LCSS_{\delta,\epsilon}^{i,j} = LCSS_{\delta,\epsilon}((x_1, \dots, x_i), (y_1, \dots, y_j))$, and $\|\cdot, \cdot\|$ is the distance between the points x_p and y_q . The parameter ϵ is the threshold on the maximum distance between a pair of matched points, and δ controls the maximum number of consecutive points without a match. $LCSS$ matches only points that are very similar, and discards points that are significantly affected by noise. The similarity between the two trajectories is then measured as

$$D_{\delta,\varepsilon}(X, Y) = 1 - \frac{LCSS_{\delta,\varepsilon}^{p,q}}{\min(p, q)}. \quad (2)$$

Our formulation is different from that in [6] where the parameter ε controlled the maximum absolute difference between each of the vector components.

2.3. Dimensions-weighted clustering

In the context of clustering multidimensional trajectories, the k -medoids clustering algorithm can be stated as the algorithm that attempts to solve the following minimization problem:

$$\min \sum_{k=1}^K \sum_{n=1}^N I_{\mathcal{J}^n \in \hat{\mathcal{J}}^k} LCSS(\hat{\mathcal{J}}^k, \mathcal{J}^n) \quad (3)$$

where $(\mathcal{J}^1, \dots, \mathcal{J}^N)$ are multidimensional trajectories corresponding to the subjects in the population, K is the number of clusters, and $I_{\mathcal{J}^n \in \hat{\mathcal{J}}^k}$ is the function that indicates whether trajectory $\mathcal{J}^n$ belongs to the k -th cluster. $\hat{\mathcal{J}}^k$ denotes medoid associated with the k -th cluster.

Given that the solution space of Problem 3 is large and multimodal, k -medoids algorithm is prone to produce inconsistent and sub-optimal results that may be unreliable and misleading for biological interpretation. This in part is due to the presence of irrelevant dimensions in the data. Intuitively, only a subset of anatomical regions are associated with clinical decline, and the ability to determine these regions typically increases the reliability of the pattern recognition algorithms. Our task is therefore to determine volumetric regions of interest (i.e., dimensions) that allow to obtain most coherent clustering results.

By weighting each dimension with a nonnegative value we can transform the original multidimensional trajectory as $\mathcal{J} = \mathcal{J}(\mathbf{w})$, where $\mathbf{w} = (w_1, \dots, w_R)$ are the weights, and the values in the multidimensional trajectory $\mathcal{J}(\mathbf{w})$ have the form $\mathcal{J}(\mathbf{w}) = ((w_s s_{r,d}))$. The coherency of the obtained clusters can then be measured as

$$\mathcal{F}(\mathbf{w}) = \sum_{k=1}^K \sum_{n=1}^N I_{\mathcal{J}^n \in \hat{\mathcal{J}}^k} LCSS(\hat{\mathcal{J}}^k(\mathbf{w}), \mathcal{J}^n(\mathbf{w})) \quad (4)$$

The problem of finding simultaneously the clusters and the optimal weights can be formulated as follows:

$$\min_{\mathbf{w}} \mathcal{F}(\mathbf{w}), \quad \text{s.t. } \|\mathbf{w}, \mathbf{w}\| = 1, \quad (5)$$

where the constraint $\|\mathbf{w}, \mathbf{w}\| = 1$ in (5) encourages solutions with only a small number of dimensions having large weights.

Finding an optimal solution of Problem 5 is a challenging task. We therefore propose a steepest descend algorithm that finds a locally optimal solution of Problem 5. Our clustering procedure is detailed in Algorithm 1 and consists of the following two main steps: (1) Minimize $\mathcal{F}$ for fixed weights $\mathbf{w}$; and (2) Given a current solution to Problem 5, find dimension r along which $\mathcal{F}(\mathbf{w})$ decreases fastest with respect to $\mathbf{w}$, and update $w_r = w_r + \alpha$, where α is a small number.

Algorithm 1**Dimensions-weighted clustering****Input:** $\{\mathcal{S}^1, \dots, \mathcal{S}^n\}$: Longitudinal data in the form of multidimensional trajectories K : Desired number of clusters**Output:** $\{\hat{\mathcal{J}}^1, \dots, \hat{\mathcal{J}}^K\}$: Medoids of the obtained clusters

- 1 Select weights $w_r = \frac{1}{R}$, and randomly initialize medoids $\{\hat{\mathcal{J}}^1, \dots, \hat{\mathcal{J}}^K\}$.
- 2 Use standard k -medoids algorithm initialized with $\{\hat{\mathcal{J}}^1, \dots, \hat{\mathcal{J}}^K\}$ to find $\{\hat{\mathcal{S}}_{new}^1, \dots, \hat{\mathcal{S}}_{new}^K\}$ that minimize $\mathcal{C}$; Set $\{\hat{\mathcal{S}}_{new}^1, \dots, \hat{\mathcal{S}}_{new}^K\} \rightarrow \{\hat{\mathcal{S}}^1, \dots, \hat{\mathcal{S}}^K\}$.
- 3 For fixed $\{\hat{\mathcal{J}}^1, \dots, \hat{\mathcal{J}}^K\}$, find dimension r along which $\mathcal{C}(\mathbf{w})$ decreases fastest with respect to $\mathbf{w}$.
- 4 Set $w_r = w_r + \alpha$, and normalize $\frac{w_r}{\|\mathbf{w}\|} \rightarrow w_r$.
- 5 Repeat steps 2,3 and 4 until convergence.

3. MATERIALS

The Baltimore Longitudinal Study of Aging (BLSA) is a prospective longitudinal study of aging. Its neuroimaging component [4], currently in its 16th year, has followed a number of healthy older individuals (age 55–85 years at enrollment) with annual or semi-annual imaging and clinical evaluations. In this paper, we considered 94 normal BLSA individuals that had at least seven evaluations. From the battery of neuropsychological tests administered to participants in conjunction with each imaging evaluation, we selected the total score from the Mini-Mental State Exam (MMSE) to assess mental status, the immediate free recall score on California Verbal Learning Test (CVLT), and the long-delay free recall score on CVLT, to assess verbal learning and immediate and delayed recall [4].

Image processing

MR images were preprocessed following mass-preserving shape transformation framework [7]. Each skull-stripped MR brain image was first segmented into gray matter (GM), white matter, and cerebrospinal fluid, by a brain tissue segmentation method proposed in [8]. Afterwards, each tissue-segmented brain image was spatially normalized into a template space, by using a high-dimensional image warping method [9]. The total tissue mass is preserved in each region during the image warping, which is achieved by increasing the respective density when a region is compressed, and vice versa. Tissue density maps were generated in the template space, reflecting local volumetric measurements corresponding to GM. Finally, using a template image with 101 manually labeled ROIs, we computed average GM tissue densities as the ROI features.

4. RESULTS

It was suggested earlier in [5] that the population of healthy older adults may be composed of a dense subpopulation of cognitively stable subjects, and a rather dispersed subpopulation of cognitively declined individuals. We therefore limit our analysis to a two-cluster problem.

Additionally, due to the suboptimal nature of Algorithm 1 it is possible to obtain different clusters for different initializations. We performed clustering 50 times and selected the pair of medoids that was found most frequently among the clustering results. The parameters of the *LCSS* were selected such that no more than three evaluations were discarded during matching.

Figure 2 shows the voxel-wise group differences of the baseline images corresponding to the subjects in the obtained two clusters. Several regions of relatively reduced volumes of GM in Cluster 2 compared to Cluster 1 are evident, including the hippocampus, amygdala, some additional medial temporal lobe GM, the insular cortex (especially the superior temporal gyrus), posterior cingulate and precuneous, and orbitofrontal cortex. The inverse contrast in Figure 2 shows increased periventricular gray tissue in Cluster 2, likely due to leukoariosis in subjects from Cluster 2.

Cluster-based level of pathology

While the obtained clusters give an insight into the structure of the normal adult population, it is also desirable to have a measure that would reflect cognitive decline of an individual subject. As Cluster 2 has significantly larger atrophy, we identify Cluster 2 as the “less normal” cluster, and Cluster 1 as the “more normal” cluster. Given the medoid of the cluster of subjects with higher pathology (i.e., less normal) $\hat{\mathcal{J}}^L$, and the medoid of the cluster of more normal subjects $\hat{\mathcal{J}}^M$, a measure that reflects the level of pathology in a brain described by a multidimensional trajectory $\mathcal{S}$ can be set as follows:

$$\mathcal{D}(\mathcal{S}) = \text{LCSS}(\hat{\mathcal{J}}^M, \mathcal{S}) - \text{LCSS}(\hat{\mathcal{J}}^L, \mathcal{S}) \quad (6)$$

As the result, we expect cognitively more stable individuals to have lower level of pathology. Having calculated the level of pathology for every longitudinal scan in the population, we can assess the validity of the proposed measure by analyzing the correlations between the measure and clinical evaluations. Figure 3 shows the relationship between our cluster-based measure of pathology and age. The measure exhibits relatively strong correlation with age, with the correlation coefficient being $r = 0.45$.

Comparison of groups corresponding to quantitative scores that are not sufficiently dissimilar may not be appropriate, as the quantitative measures are most likely not to be perfectly discriminative. Therefore, in order to understand the meaning of the level of pathology in Equation 6, we selected the groups that are far apart with respect to the obtained quantitative measure. One group was formed out of individuals having level of pathology values in the upper quartile (i.e., upper 25%). The subjects with the lower 75% of the level of pathology were placed into the second group.

The results in Table 1 indicate that the subpopulation with high level of pathology shows significantly higher rate of cognitive decline. This yet again suggests that the clustering approach identified a subpopulation that is relatively less cognitively stable. Moreover, subjects with high values of the level of pathology have significantly worse CVLT scores at the last available evaluation, as well as significantly lower mean CVLT scores. Additionally, if the dimensions are not weighted, the measure of pathology obtained using the results of the standard *k*-medoids clustering algorithm in Equation 3 typically show less significant correlation with the rate of cognitive decline (i.e., second row in Table 1).

5. CONCLUSION

In this paper, we presented a method for disentangling heterogeneity of healthy adult populations by clustering longitudinal images. We proposed a measure of individual's brain pathology that is associated with cognitive decline. Our method is unique in that it allows to obtain an individual-based biomarker of cognitive decline in an unsupervised manner. As such, our approach is applicable to populations that lack clearly defined clinical categories.

While we identified the less normal cluster in a post-hoc manner, the process can be automated by seeding the population with subjects that have known labels as it was done in [5], where the less normal cluster was identified as the one that contained most of the seeded MCI subjects.

Acknowledgments

This study was supported in part by NIH funding sources N01-AG-3-2124 and R01-AG14971, and by the Intramural Research Program of the National Institute on Aging, NIH.

REFERENCES

1. Duchesne S, Bocti C, De Sousa K, Frisoni GB, Chertkow H, Collins DL. Amnesic mci future clinical status prediction using baseline mri features. *Neurobiol Aging*. 2010; vol. 31(no. 9):1606–1617. [PubMed: 18947902]
2. Kloppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, Fox NC, Jack CR, Ashburner J, Frackowiak RSJ. Automatic classification of mr scans in alzheimer's disease. *Brain*. 2008; vol. 131(no. 3):681–689. [PubMed: 18202106]
3. Misra C, Fan Y, Davatzikos C. Baseline and longitudinal patterns of brain atrophy in mci patients, and their use in prediction of short-term conversion to ad: Results from adni. *NeuroImage*. 2009 Feb; vol. 44(no. 4):1415–1422. [PubMed: 19027862]
4. Driscoll I, Davatzikos C, An Y, Wu X, Shen D, Kraut M, Resnick SM. Longitudinal pattern of regional brain volume change differentiates normal aging from MCI. *Neurology*. 2009; vol. 72(no. 22):1906–1913. [PubMed: 19487648]
5. Filipovych, Roman; Resnick, Susan M.; Davatzikos, Christos. Semi-supervised cluster analysis of imaging data. *NeuroImage*. 2010; vol. In Press, Corrected Proof.
6. Vlachos M, Hadjieleftheriou M, Gunopulos D, Keogh E. Indexing multi-dimensional time-series with support for multiple distance measures. *ACM SIGKDD*. 2003:216–225.
7. Davatzikos C, Genc A, Xu D, Resnick SM. Voxelbased morphometry using the ravens maps: methods and validation using simulated longitudinal atrophy. *Neuroimage*. 2001 Dec; vol. 14(no. 6):1361–1369. [PubMed: 11707092]
8. Pham, Dzung L.; Prince, Jerry L. Adaptive fuzzy segmentation of magnetic resonance images. *IEEE Trans. Med. Imaging*. 1999; vol. 18(no. 9):737–752. [PubMed: 10571379]
9. Shen DG, Davatzikos C. Hammer: Hierarchical attribute matching mechanism for elastic registration. *IEEE Trans. Med. Imag*. 2002 Nov; vol. 21(no. 11):1421–1439.

Fig. 1. Multidimensional trajectory representation. Images are registered to a common template. For a given ROI, a feature is extracted from all scans of the individual. The sequences of extracted ROI features form a multidimensional trajectory representation of the subject's imaging evaluations.

Fig. 2.
Voxel-wise differences in base-line images of the subjects in the two clusters. The colormap shows t -statistic and was thresholded at $p < 0.001$

Fig. 3.
Correlation between measure of pathology and age.

Table 1

Relationship between scores and level of pathology.

	CVLT List A Sum	CVLT Long Delay Free	MMSE
Rate of change	0.086	0.031	0.037
no weighting	0.097	0.009	0.201
Mean Scores	0.008	0.014	0.189
First Visit Scores	0.033	0.178	0.351
Last Visit Scores	0.003	0.039	0.156

p-values of one-sided *t*-test obtained for the groups with level of pathology in upper 25% vs. the group with lower 75%.