COMBINING IMAGING AND CLINICAL DATA IN MANIFOLD LEARNING: DISTANCE-BASED AND GRAPH-BASED EXTENSIONS OF LAPLACIAN EIGENMAPS

Jean-Baptiste Fiot [1,2] *Jurgen Fripp* [2] *Laurent D. Cohen* [1]

[1] CEREMADE, UMR 7534 CNRS Université Paris Dauphine, France
[2] CSIRO Preventative Health National Research Flagship ICTC,
The Australian e-Health Research Centre-BioMedIA,
Royal Brisbane and Women's Hospital, Herston, QLD, Australia

ABSTRACT

Manifold learning techniques have been widely used to produce low-dimensional representations of patient brain magnetic resonance (MR) images. Diagnosis classifiers trained on these coordinates attempt to separate healthy, mild cognitive impairment and Alzheimer's disease patients. The performance of such classifiers can be improved by incorporating clinical data available in most large-scale clinical studies. However, the standard non-linear dimensionality reduction algorithms cannot be applied directly to imaging and clinical data. In this paper, we introduce a novel extension of Laplacian Eigenmaps that allow the computation of manifolds while combining imaging and clinical data. This method is a distance-based extension that suits better continuous clinical variables than the existing graph-based extension, which is suitable for clinical variables in finite discrete spaces. These methods were evaluated in terms of classification accuracy using 288 MR images and clinical data (ApoE genotypes, $A\beta_{42}$ concentrations and mini-mental state exam (MMSE) cognitive scores) of patients enrolled in the Alzheimer's disease neuroimaging initiative (ADNI) study.

Index Terms— Manifold learning, population analysis, image processing, clinical data, Alzheimer's disease

1. INTRODUCTION

Large scale population studies aim to improve the understanding of the causes of diseases, define biomarkers for early diagnosis, and develop preventive treatments. In the context of the Alzheimer's disease (AD), imaging biomarkers, blood biomarkers, cognitive tests, lifestyle and diet biomarkers are all potential sources of information to diagnose the disease as early as possible.

Manifold learning techniques have been used to analyse trends in populations and describe the space of brain images by a low-dimensional non-linear manifold [1, 2]. These studies attempt to describe the space of brain images via a low-dimensional manifold while capturing relevant information with regard to disease diagnosis. Diagnosis classifiers trained on the low-dimension coordinates evaluate the ability to capture this information and separate healthy, mild cognitive impairment and Alzheimer's patients [2].

As most of the current large-scale clinical studies also provide non-imaging information, one would want to be able to use this information to improve the diagnosis classification. However, as the imaging and clinical data are in different spaces, the non-linear dimensionality reduction cannot be applied directly and must be adapted. We introduce a distance-based extension and compare it theoretically to an exisiting graph-based extension [2]. We also evaluate their numerical classification performances on a large dataset from ADNI [3].

2. METHODS

2.1. Population analysis and diagnosis classification from manifold learning

It has been shown that the space of brain images in $\mathbb{R}^M$ can be described by a non-linear manifold $\mathcal{M}$ of intrisic dimension m, with $m \ll M$ [1]. Laplacian eigenmaps (LEM) [4] can be used to compute the low-dimensional representation of the data (Fig. 1). Given a matrix Δ of pairwise distances between n images and a number of nearest neighbours (NN) $k \in \mathbb{N}$, an adjacency-graph $\mathcal{G} =< V, E >$ is computed. Each node represents an image, and weighted edges connecting each image to its k-NN are created. From the weight matrix W, a diagonal matrix D is computed with $d_{ii} = \sum_j w_{ij}$. The graph Laplacian is given by $L = D - W$. Its eigenvectors $\{e_j \in \mathbb{R}^n\}_{1 \leq j \leq m}$ associated to the m smallest non-zero eigenvalues provide the low-dimension coordinates $\{y_i = (e_1^i, \ldots, e_m^i) \in \mathbb{R}^m\}_{1 \leq i \leq n}$. Noting $y = (y_1, \ldots, y_n)^T$, these coordinates are the solutions of the optimization problem:

$$\underset{y^T D y = I}{\operatorname{argmin}} \sum_{ij} w_{ij} ||y_i - y_j||^2 \qquad (1)$$

To evaluate the ability of the dimensionality reduction process to capture relevant information with regard to disease progression, it is possible to use the low-dimension coordinates to train a disease classifier.

$$\underset{\Delta \in \mathbb{R}^{n\times n}}{\text{Distance matrix}} \rightarrow \underset{W\in\mathbb{R}^{n\times n}}{\text{k-NN graph}} \rightarrow \underset{L\in\mathbb{R}^{n\times n}}{\text{Laplacian}} \rightarrow \underset{Y\in\mathbb{R}^{n\times m}}{\text{Coordinates}}$$

Fig. 1: Standard LEM pipeline to compute low-dimension coordinates Y. The distance-based extension modifies Δ, whereas the graph-based extension modifies W.

2.2. Extended LEM based on distance matrix combination

To combine imaging and clinical data in the manifold learning process, one can define a distance on the clinical data, combine linearly the image-based and clinical-based distance matrices ($\Delta = \Delta_{img} + \lambda\Delta_{clinical}$), and apply the standard LEM algorithm. This extension adds two constraints to the original algorithm: 1) the need for a distance on the clinical data and 2) the need to define a weight for the clinical data. Combining the two distance matrices and applying LEM creates a graph with the same nodes but different edges and weights $\hat{w}_{ij}$ (Fig. 2a and 2b). Using this extension, the optimisation problem becomes:

$$\underset{y^T\hat{D}y=I}{\operatorname{argmin}} \sum_{ij} \hat{w}_{ij} \, ||y_i - y_j||^2$$

2.3. Extended LEM based on adjacency graph extension

An alternative method to combine imaging and clinical data is to extend the adjacency graph by adding extra nodes and edges. One such technique has been presented in [2]. This extension also adds two constraints to the original algorithm: 1) a set of rules to extend the graph (extra nodes and extra weights), 2) the need to define a weight for the clinical data.

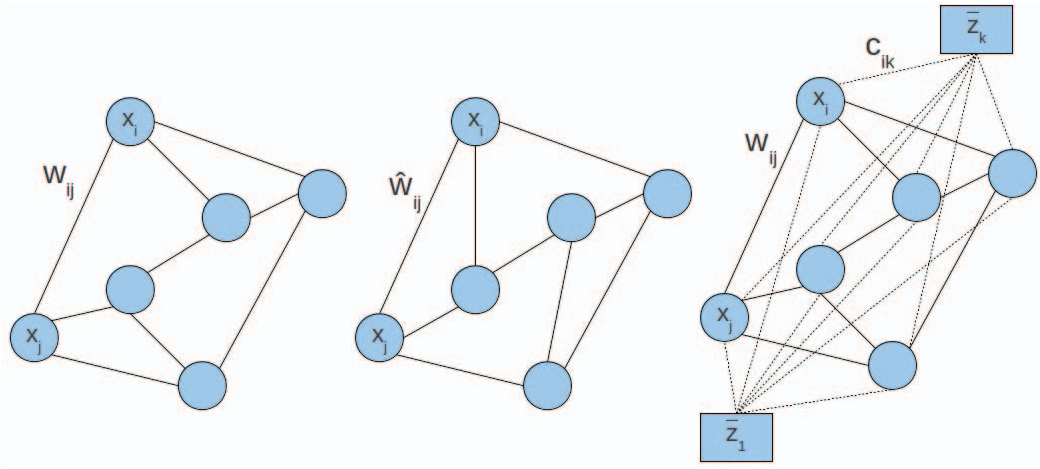
When the clinical variable is in a discrete finite space, such as for ApoE genotype, one node is created for each element of that space. Extra edges are created (with a weight equal to one) between the node of each patient with a particular element to the node of that element (e.g. all the patients with a particular genotype are connected to the node representing that genotype). When the clinical variable is in a continuous space, such as $A\beta_{42}$ concentration or MMSE clinical score, Wolz et al. proposed to partition this continuous space and set the weights as the fuzzy probabilities of belonging to each partition:

$$\forall k \in \{1,\dots,\tilde{n}\}, \forall i \in \{1,\dots,n\}, c_{ik} = \frac{\frac{1}{d(z_i,\overline{z_k})}}{\sum_{k=1}^{\tilde{n}} \frac{1}{d(z_i,\overline{z_k})}}$$

where d is a distance on the space of clinical variable, the $\overline{z_k}$ are the means of the sub-intervals defined using the mininum, maximum and several percentile values. Figure 2c represents the extended graph, which corresponding matrix is written

$$\tilde{W} = \begin{pmatrix} I & \frac{\gamma}{2}C^T \\ \frac{\gamma}{2}C^T & W \end{pmatrix}$$

where I is the identity matrix, W is the weight matrix of the standard LEM on images, and C contains the weights of the extra-edges. A parameter γ is introduced to weight the clinical data versus the imaging data. When extending the graph by $\tilde{n}$ nodes, we are now looking for $Y =$



(a) Standard LEM (b) Distance-based LEM extension (c) Graph-based LEM extension

Fig. 2: Comparison of the graphs in the standard LEM algorithm and in the two extensions. When combining distances matrices, one gets a graph as in 2b with the same nodes as the standard LEM 2a but different edges and different weights. In the graph-based LEM extension, the graph 2c is built from the graph of the standard LEM 2a, then extra new nodes and weights are added.

$(\tilde{y}_1, \ldots, \tilde{y}_{\tilde{n}}, y_1, \ldots, y_n)$, $\tilde{y}_k, y_i \in \mathbb{R}^m$ as a solution of the optimisation problem:

$$\underset{Y^T DY = I}{\operatorname{argmin}} \sum_{ij} w_{ij} \, ||y_i - y_j||^2 + \gamma \sum_{ik} c_{ik} \, ||y_i - \tilde{y}_k||^2$$

3. MATERIAL AND RESULTS

3.1. Data

A dataset of 288 Magnetic Resonance images from 101 patients enrolled in ADNI (`http://www.loni.ucla.edu/ADNI`, [3]) has been used to compare the diagnosis classification performances of the standard LEM algorithm and its two extensions.

As clinical data, ADNI provides the Apolipoprotein E (ApoE) genotype. Three ApoE alleles exist $(\epsilon_2, \epsilon_3, \epsilon_4)$, and since each individual carries two alleles, six ApoE genotypes are possible. The ϵ_4 allele has been shown to increase the risk of developing AD, whereas ϵ_2 decreases this risk [5]. Moreover an $A\beta_{42}$ protein analysis of cerebrospinal fluid (CSF) is provided. A decrease in the concentration of this protein has been shown to be associated with a development of AD [3]. Table 1 summarises the clinical information for the various diagnostics in the dataset.

3.2. Experiments

The 288 images were intensity normalized by histogram equalization to the ICBM152 atlas [6] used as template. All the images were then rigidly registered to the atlas using [7]. Sub-images around the hippocampus were finally extracted. A deformation based distance of this area was used for the image distance matrix. The ApoE genotype was used considering all possible pairs of alleles and considering ApoE carriers as in [2], respectively leading to 6 and 3 extra nodes in the graph-based extension. For the graph-based extension with the continuous clinical variables ($A\beta_{42}$ and MMSE), 3 extra nodes were added as in [2]. Adjacency graphs were k-nearest neighbour graphs ($k = 100$) with edges weights computed using the gaussian kernel with a kernel width equal to the standard deviation of the distance matrix coefficients. LEM was applied with target dimension from 2 to 100. The classifiers used were k-nearest neighbour classifiers ($k = 50$). Training set and test sets were built using a leave 5% out scheme. The optimal target dimension in LEM and optimal λ (resp. γ) were automatically selected from a 20-cross validation on the training set on $\{1, \ldots, 100\} \times \{0.1, 1, 2, 5, 10, 25\}$ (resp. $\{1, \ldots, 100\} \times \{0.01, 0.01, 0.1, 0.5, 1, 2, 5, 10, 25\}$).

3.3. Results

Table 2 presents the classification performance of the standard LEM algorithm using the imaging data, and the two extensions using the combined imaging and clinical data. Using clinical data combined with imaging data improves classification results for both methods compared to the standard LEM on only imaging data. For the discrete clinical variable ApoE genotype, the two extensions have similar performance on this dataset. For the continuous clinical variables $A\beta_{24}$ CSF concentration and MMSE cognitive score, the distance-based extension performs better than the graph-based extension.

4. DISCUSSION

We have presented two extensions of LEM able to perform non-linear dimensionality reduction with data from different spaces, such as imaging and clinical data. Both methods come with two additional constraints. In particular, they both need to set an extra parameter to balance how much weight is given to the clinical information versus the weight of the imaging information.

From a theoretical point of view, the graph-based extension seems more natural when the clinical variable is in a finite discrete space, whereas the distance based extension seems more natural when the clinical data lives in a continuous space. First, when the clinical variable's space is a finite discrete space, it is easy to add one node per possible value and edges with weights equal to one for class memberships. However, using the distance-based extension when the clinical variable is in a discrete space requires to define a distance on that space. Depending on the problem, this can raise difficult questions. In the case of the ApoE genotype, we can for example wonder if creating a distance being equal to one between all pairs of different genotypes is really optimal. Having $d((\epsilon_2, \epsilon_2), (\epsilon_4, \epsilon_4))$ higher than $d((\epsilon_2, \epsilon_2), (\epsilon_2, \epsilon_3))$ would not be absurd given the known biological impact of the ApoE alleles [5]. This example illustrates that the distance-based extension is not necessarily well suited for discrete clinical variables. On the other hand, when the clinical variable lives in a continuous space such as $\mathbb{R}^n$, many distances are commonly associated (e.g. distances from l^p norms $\|x\|_p = \left(\sum_i {x_i}^p\right)^{1/p}$). However, if one wants to use the graph extension technique, it is obviously impossible to add an infinite number of nodes. So the continous space has to be discretized into a finite number of subparts. At this point, using memberships to these subparts would mean that each z value would be considered as being one of the $\overline{z_k}$. To avoid this huge loss of information, Wolz et al. introduced fuzzy memberships. Nonetheless, there is no natural way to select the number of elements of the partition. In their paper, Wolz et al. have a clinical variable in $z \in \mathbb{R}$, they add $\tilde{n} = 3$ extra nodes, and the weights were defined by the minimum of z, its 33% and 67% percentiles and its maximum value, but this choice is rather arbitrary.

From a numerical point of view, when the graph-based LEM extension is used with a continuous clinical variable, the divisions in the c_{ik} can be sources of numerical instability.

Table 1: Number of patients, ApoE genotypes, mean and standard deviation of $A\beta_{42}$ concentration in CSF and mini-mental state exam (MMSE) cognitive scores are shown for the normal controls (NC), mild cognitive impairment (MCI) and Alzheimer's disease (AD) patients.

Diagnosis	N	ApoE genotype						$A\beta_{42}$	MMSE cognitive score
		(ϵ_2, ϵ_2)	(ϵ_2, ϵ_3)	(ϵ_2, ϵ_4)	(ϵ_3, ϵ_3)	(ϵ_3, ϵ_4)	(ϵ_4, ϵ_4)		
NC	94	0	12	0	65	17	0	210.15 ± 58.15	29.28 ± 1.02
MCI	114	0	2	0	58	46	8	160.48 ± 43.50	26.62 ± 1.92
AD	80	0	0	3	26	37	14	137.53 ± 24.54	21.53 ± 4.74

Table 2: Performance $(\%)$ of the standard LEM algorithm and its two extensions in diagnosis classification.

Data	Algorithm	NC vs MCI	NC vs AD	MCI vs AD
Imaging	LEM	65.6	63.3	61.9
Imaging & ApoE carriers	Distance-based LEM extension	66.7	71.8	66.1
Imaging & ApoE carriers	Graph-based LEM extension	65.8	73.0	64.8
Imaging & ApoE pairs	Distance-based LEM extension	62.3	62.5	66.0
Imaging & ApoE pairs	Graph-based LEM extension	63.8	65.8	65.3
Imaging & $A\beta_{42}$	Distance-based LEM extension	70.7	75.5	67.1
Imaging & $A\beta_{42}$	Graph-based LEM extension	65.2	70.7	65.5
Imaging & MMSE	Distance-based LEM extension	83.2	93.1	67.1
Imaging & MMSE	Graph-based LEM extension	65.8	75.3	68.8

5. CONCLUSION AND PERSPECTIVES

We have introduced a novel extension of LEM able to perform non linear dimensionality reduction while combining imaging data and clinical data which are in different spaces. This distance-based extension leads to a graph with the same nodes as from the standard LEM but with different edges and weights, whereas the previously existing graph-based extension leads to a graph where all the nodes and edges from the standard LEM are kept and extra ones are created. This new distance-based extension is better suited for a continuous clinical data than the graph-based which is well-suited when the clinical variable lives in a finite discrete space.

We have shown that both extensions improve the numerical classification performance compared to the original LEM on a large dataset from ADNI. Performances of both extensions are similar with the discrete ApoE genotype clinical value, and our new distance-based extension have higher classification accuracy with the continuous clinical variables $A\beta_{42}$ CSF concentrations and MMSE clinical scores.

In terms of generalization of the two extensions to other dimensionality reduction algorithms, the existing graph-based extension can potentially be adapted only if the dimensionality reduction process is based on a graph. Our new distance-based extension is more general and can be directly used in any dimensionality reduction algorithm that requires a distance of pairwise distances between all objects as input.

6. REFERENCES

[1] Gerber, S, et al. "Manifold modeling for brain population analysis." *Medical Image Analysis*, vol. 14, no. 5, pp. 643 – 653, 2010.

[2] Wolz, R, et al. "Manifold learning combining imaging with non-imaging information." In *ISBI*, pp. 1637–1640. 2011.

[3] Mueller, SG, et al. "The Alzheimer's Disease Neuroimaging Initiative." *Neuroimaging Clinics of North America*, vol. 15, no. 4, pp. 869 – 877, 2005. Alzheimer's Disease: 100 Years of Progress.

[4] Belkin, M and Niyogi, P. "Laplacian eigenmaps for dimensionality reduction and data representation." *Neural Comput.*, vol. 15, 2003.

[5] Macdonald, A and Pritchard, D. "A mathematical model of Alzheimer's disease and the ApoE gene." *ASTIN Bulletin*, vol. 30, pp. 69–110, 2000.

[6] Mazziotta, J, et al. "A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM)." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 356, no. 1412, pp. 1293–1322, 2001.

[7] Ourselin, S, et al. "Reconstructing a 3D structure from serial histological sections." *Image and Vision Computing*, vol. 19, no. 1-2, pp. 25 – 31, 2001.