Nonlinear Metric Learning for Alzheimer's Disease Diagnosis with Integration of Longitudinal Neuroimaging Features

Bibo Shi[1]
bs354409@ohio.edu
Yani Chen[1]
yc147311@ohio.edu
Kevin Hobbs[2]
hobbsk@ohio.edu
Charles D. Smith[3]
csmith@mri.uky.edu
Jundong Liu[1]
liu@cs.ohio.edu

[1] School of Electrical Engineering and Computer Science
Ohio University
Athens OH, USA

[2] Department of Biological Sciences
Ohio University
Athens OH, USA

[3] Department of Neurology
University of Kentucky
Lexington KY, USA

Abstract

Identifying neuroimaging biomarkers of Alzheimer's disease (AD) is of great importance for diagnosis and prognosis of the disease. In this study, we develop a novel nonlinear metric learning method to improve biomarker identification for Alzheimer's disease and its early stage Mild Cognitive Impairment (MCI). Formulated under a constrained optimization framework, the proposed method learns a smooth nonlinear feature space transformation that pulls the samples of the same class closer to each other while pushing different classes further away. The thin-plate spline (TPS) is chosen as the geometric model due to its remarkable versatility and representation power in accounting for sophisticated deformations. In addition, a multi-resolution patch-based feature selection strategy is proposed to extract both cross-sectional and longitudinal features from MR brain images. Using the ADNI dataset, we evaluate the effectiveness of the proposed metric learning and feature extraction strategies and demonstrate the improvements over the state-of-the-art solutions within the same category.

1 Introduction

Alzheimer's disease (AD) and its early stage, mild cognitive impairment (MCI), affect more than 35 million people worldwide [1]. Identifying reliable biomarkers to characterize different stages of AD would potentially provide objective and early measures for diagnosis and treatments of this disease. In the past two decades or so, neuroimaging modalities including Magnetic Resonance Imaging (MRI) have emerged as a positive predictive component and become more and more commonly used in this pursuit.

Pages 138.1-138.13
DOI: https://dx.doi.org/10.5244/C.29.138

The Alzheimer's Disease Neuroimaging Initiative (ADNI) [17] provides reliable clinical data including structural and functional MR imaging to support the research on intervention, prevention and treatments of AD. Since the inception of ADNI in 2004, a significant amount of research effort have been conducted on supplementing imaging data by combining it with cerebrospinal fluid (CSF) biomarker levels and genetics information, as well as utilizing a variety of newer classification methods to differentiate patient groups. However, insufficient attention has been given to rationally selecting appropriate metrics (equivalent to transforming the feature space) from the training data that could maximize the power of various classifiers.

Metric learning (ML), the procedure aiming to learn a good distance metric tuned to a particular task with certain side information, would certainly offer a remedy in this regard. The learned metric, tailored to the training input, can significantly improve the performance of many metric-based algorithms, such as *k*NN, *k*-means, and even SVMs [37], in various classification, clustering and retrieval tasks [3, 38].

Learning a metric from the training input is equivalent to learn a feature transformation [3]. Depending on the feature space transformation to be sought, metric learning can be divided into linear and nonlinear groups [38]. Linear models commonly try to estimate a "best" affine transformation to deform the feature space, such that the resulted Mahalanobis distance would very well agree with the supervisory information brought by training samples. Many early works have focused on linear methods because they are easy to use, convenient to optimize and less prone to overfitting [3]. However, when handling data with nonlinear structures, linear models show inherently limited expressive power and separation capability. Nonlinear models are usually designed through kernelization or localization of certain linear models. The idea of kernelization [20, 31] is to embed the input features into a higher dimensional space, with a goal that the data would be more linearly separable under the new space. While kernelization may dramatically improve the performance of linear methods for many highly nonlinear problems, solutions in this group are prone to overfitting [3], and their utilization is inherently limited by the sizes of the kernel matrices [13]. Localization approaches focus on combining multiple local metrics, which were learned based on either local neighborhoods or class memberships. The granularity levels of the neighborhoods vary from per-partition [16, 25], per-class [35] to per-exemplar [24, 33]. Although the multi-metric strategies are usually more powerful in accommodating nonlinear structures, generalizing these methods to fit other classifiers than *k*NN is not trivial. To avoid non-symmetric metrics, extra cares are commonly needed to ensure the smoothness of the transformed feature space. In addition, estimating geodesic distances and group statistics on such metric manifolds are often computationally expensive.

Other than metric learning, feature extraction and selection from the ADNI database is also in great need of further exploration. For structural features extracted from brain MRIs, cortical thickness [19], hippocampal volume/shape [5, 23] and voxel tissue probability maps [9, 21] across the whole brain or around certain regions of interest (ROI), are among the popular choices. Most of them are either cross-sectional features obtained at one point in time, or "static" longitudinal volumetric information acquired at two or multiple time points but only through structural segmentation. In part due to the unavailability of deformation data under ADNI, "dynamic" longitudinal information such as the atrophy over time at various gray matter (GM) areas, which is a major hallmark in the progression of AD, has not been fully utilized in the literature.

In this paper, we propose to improve the quality of AD neuroimage biomarker identification along two directions: 1) feature space transformation through a novel nonlinear ML

technique; 2) extraction and integration of dynamic longitudinal atrophy features into the classification framework. The proposed ML solution is a direct generalization of linear ML through the application of a deformable geometric model – the thin-plate spline (TPS) - to transform the feature space. TPS is chosen due to its remarkable versatility and representation power in accounting for high-order deformations. Unlike the multi-metric solutions, our proposed nonlinear ML method seeks a smooth global feature transformation, which can be applied as a preprocessing step for a variety of classifiers. Toward the integration of longitudinal information, we propose a multi-resolution patch selection strategy, with both cross-sectional (baseline) and longitudinal atrophy features extracted from MR brain images.

The rest of the paper is organized as follows. Section 2 introduces a classic linear ML model [36], which is used as the platform for our proposed TPS-based nonlinear ML model presented in Section 3. Section 4 describes the proposed multi-resolution patch extraction and selection procedure. In Section 5, we present experiments and results to evaluate the components of our model. Finally, section 6 concludes this paper.

2 A Classic Linear ML Model: MMC

In this paper, a pioneer Mahalanobis ML for clustering method (MMC) proposed by Xing *et al*. [36] will be used as the platform to formulate our nonlinear TPS solution. Therefore, we briefly review the concept of MMC here.

Given a set of training data instances $\mathcal{X} = \{\mathbf{x}_i | \ \mathbf{x}_i \in \mathbb{R}^m, i = 1, \cdots, n\}$, where n is the number of training samples, and m is the number of features that a data instance has, the goal of ML is to learn a "better" metric function $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to the problem of interest with the information carried by the training samples. Mahalanobis metric is one of the most popular metric functions used in existing ML algorithms [11, 15, 16, 18, 27, 34], which is defined by $D_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T M (\mathbf{x}_i - \mathbf{x}_j)}$. The control parameter $M \in \mathbb{R}^{m \times m}$ is a square matrix. In order to qualify as a valid (pseudo-)metric, M has to be positive semi-definite (PSD), denoted as $M \succeq 0$. As a PSD matrix, M can be decomposed as $M = L^T L$, where $L \in \mathbb{R}^{k \times m}$ and k is the rank of M. Then, $D_M(\mathbf{x}_i, \mathbf{x}_j)$ can be rewritten as follows:

$$D_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T L^T L (\mathbf{x}_i - \mathbf{x}_j)} = \sqrt{(L\mathbf{x}_i - L\mathbf{x}_j)^T (L\mathbf{x}_i - L\mathbf{x}_j)}. \tag{1}$$

Eqn. (1) explains why learning a Mahalanobis metric is equivalent to learning a linear transformation function and computing the Euclidean distance over the transformed data domain. With the side information embedded in the class-equivalent constraints $\mathcal{P} = \{(\mathbf{x}_i, \mathbf{x}_j) | \ \mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class$\}$ and class-nonequivalent constraints $\mathcal{N} = \{(\mathbf{x}_i, \mathbf{x}_j) | \ \mathbf{x}_i$ and $\mathbf{x}_j$ belong to different classes$\}$, MMC formulated the problem of ML into the following convex programming problem:

$$\min_M \quad J(M) = \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{P}} D_M^2(\mathbf{x}_i, \mathbf{x}_j) \quad \text{s.t.} \quad M \succeq 0, \quad \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{N}} D_M^2(\mathbf{x}_i, \mathbf{x}_j) \geq 1. \tag{2}$$

The objective function aims at improving the subsequent classification for the data via minimizing the sum of distances between similar training data, while keeping the sum of distances between dissimilar ones large. Note that, besides the PSD constraint on M, an additional constraint on the training samples in $\mathcal{N}$ is needed to avoid trivial solutions for the optimization. To solve this optimization problem, the projected gradient descent method is used, which projects the estimated matrix back to the PSD group whenever it is necessary.

3 Metric Learning through TPS (ML-TPS)

Instead of using a linear transformation L as in MMC, we choose to deform the feature space through a radial basis function – thin-plate spline (TPS). The TPS is the high-dimensional analog of the cubic spline in one dimension, and was first used in surface reconstruction research as an interpolation tool. In m dimensions, the idea of TPS is to choose a function $f(\mathbf{x})$ that exactly goes through the data points $(\mathbf{x}_i, y_i)$ (i.e., $y_i = f(\mathbf{x}_i)$) and minimizes the bending energy, $E[f] = \int_{R^m} |\mathcal{D}^2 f|^2 \mathrm{d}X$, where $\mathcal{D}^2 f$ is the matrix of second-order partial derivatives of f, and $\mathrm{d}X = \mathrm{d}x_1 ... \mathrm{d}x_m$, where x_j are the components of $\mathbf{x}$. The Euler-Lagrange equation for $E[f]$, which specifies the necessary condition the minimizing function should satisfy, is the *biharmonic equation*

$$\Delta^2 f = \sum_{k=1}^{m} \sum_{l=1}^{m} f_{x_k x_l}^2 = 0. \tag{3}$$

The classic solution for Eqn. (3) has a representation in terms of a radial basis function,

$$f(\mathbf{x}) = \sum_{i=1}^{n} w_i G(||\mathbf{x} - \mathbf{x}_i||) + \mathbf{b}^T \mathbf{x} + c, \tag{4}$$

where $||.||$ denotes the Euclidean norm and $\{w_i\}$ are a set of weights for the nonlinear part; $\mathbf{b}$ and c are the weights for the linear part. The corresponding radial distance kernel of TPS, which is the Green's function to solve Eqn. (3), is as follows:

$$G(\mathbf{x}, \mathbf{x}_k) = G(||\mathbf{x} - \mathbf{x}_k||) \propto \begin{cases} ||\mathbf{x} - \mathbf{x}_k||^2 \ln ||\mathbf{x} - \mathbf{x}_k||, & \text{if } m \text{ is even positive;} \\ ||\mathbf{x} - \mathbf{x}_k||, & \text{otherwise.} \end{cases} \tag{5}$$

For more details about TPS, we refer readers to [8, 32].

The TPS transformation for data interpolations, as specified in Eqn. (4), can be employed as the geometric model to deform the feature spaces to achieve nonlinear metric learning. Such transformation would ensure certain desired smoothness as it minimizes the bending energy $E[f]$ of the transformation. Within the ML setting, let $\mathbf{x}$ be one of the training samples in the original feature space $\mathcal{X}$ of m dimensions, and $f(\mathbf{x})$ be the transformed destination of $\mathbf{x}$, which is still of m dimensions. Through a straightforward mathematical manipulations [4], we can get $f(\mathbf{x})$ in matrix format:

$$f(\mathbf{x}) = \mathbf{x} \cdot B + \begin{pmatrix} G(\mathbf{x}, \mathbf{x}_1) \\ \cdots \\ G(\mathbf{x}, \mathbf{x}_p) \end{pmatrix} \cdot W = \mathbf{x} \cdot B + \vec{G} \cdot W, \tag{6}$$

where B (size $m \times m$) is the linear transformation matrix, corresponding to L^T in Mahalabonis metric, and W is the weight matrix for the nonlinear parts. $\mathbf{x}_p$ are the anchor points used to compute the TPS kernel. Usually, we can use all the training data points as the anchor points. However, in practice, p anchor points are extracted via different methods to describe the whole input space under the consideration of computational cost. In this study, k-medoids method is adopted, similar as in [33].

The goal of our ML solution is still pulling the similar subjects closer while pushing dissimilar subjects apart, directly through a TPS nonlinear transformation f as described in

Eqn. (6). This can be achieved through the following constrained optimization:

$$\begin{aligned} \min_{B,W} \quad & J = \sum_{\mathbf{x}_i,\mathbf{x}_j \in \mathcal{P}} \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|^2 + \lambda \|W\|_F^2 \\ \text{s.t.} \quad & \sum_{\mathbf{x}_i,\mathbf{x}_j \in \mathcal{N}} \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|^2 \geq 1; \\ & \sum_{i=1}^{p} W_i^k = 0, \quad \sum_{i=1}^{p} W_i^k \mathbf{x}_i^k = 0, \ \forall k = 1 \ldots m. \end{aligned} \tag{7}$$

Compared with MMC, another component $\|W\|_F^2$, the squared Frobenius norm of W, is added to the objective function as a regularizer to prevent overfitting. λ is the weighting factor to control the importance of two components. Similar as in MMC, the nonequivalent constraint $\sum_{\mathbf{x}_i,\mathbf{x}_j \in \mathcal{N}} \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|^2 \geq 1$ is to impose a scaling control to avoid trivial solutions. The other two equivalent constraints with respect to W is to ensure that the elastic part of the transformation is zero at infinity [26]. W^k is the kth column of W, and $\mathbf{x}^k$ is the kth component of $\mathbf{x}$.

Due to the nonlinearity of TPS, it is difficult to analytically solve this nonlinear constrained problem. Alternatively, we can use a gradient based constrained optimization solver [1] to get a local minimum for Eqn. (7) . The complexity of our ML-TPS model is dominated by the computation of the TPS kernel, which is $O(p * n^2)$, as well as the rate of convergence of the chosen gradient based optimizer. n is the number of training samples, and p is the number of anchor points.

4 Neuroimage Data and Feature Extraction

The neuroimage data used in this work were obtained from the ADNI database [17]. We consider only the subjects for whom the baseline (M0) visits and 12-month follow-up (M12) T1-weighted MRIs, together with their *MIDAS Whole Brain Masks*, are all available. As a result, 338 subjects were selected: 94 patients with AD, 121 with MCI and 123 normal controls (NC). More detailed information, including the demographics and clinical evaluations of the subjects, is available in the supplementary material.

Recently, patch-level neuroimage features extraction and fusion [22, 30] have been used in producing excellent performance for AD/MCI/NC classifications. The features utilized in their work are cross-sectional, extracted from the baseline MRIs and Positron emission tomography images (PETs). In this paper, we propose a multi-resolution patch extraction strategy with longitudinal brain atrophy, which is one of the pathological hallmarks of AD, as an addition information source. Fig. 1 illustrates an overview of the schematic diagram of our proposed framework, which consists of two main steps. The first step is the extraction of class-discriminative patches from both baseline and longitudinal MRIs; the second step is a wrapper feature selection [12] to select a most discriminative subset from the patch pools.

4.1 Patch Extraction

To facilitate the ensuing patch-level operations, the T1-weighted MRIs (at both M0 and M12) were first normalized into an International Consortium for Brain Mapping template through Statistical Parametric Mapping [10], with the dimensions reduced to $79 \times 79 \times 95$ and the

[1] We use a SQP based constrained optimizer "fmincon" in Matlab Optimization Toolbox.

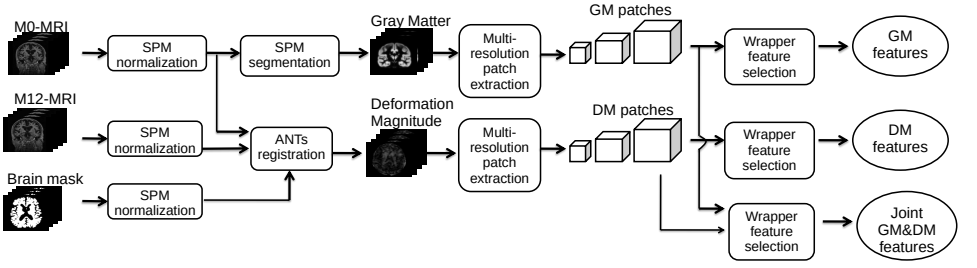


Figure 1: Flowchart of the proposed two-step patch based feature extraction and selection strategy.

voxel sizes to $2 \times 2 \times 2$ mm^3. Then, each baseline M0-MRI was segmented into three brain tissues: gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). As GM is more related to AD and MCI pathologies than WM and CSF [22], we choose the spatially normalized GM tissue densities from the baseline MRIs as the cross-sectional information source in our work.

With the GM segmentation in place, we adopt a patch extraction procedure similar to that in [22] to generate our baseline features. A voxel-wise t-test is first performed based on the group labels, i.e., AD vs. NC and MCI vs. NC. Voxels with statistically significant group difference (with the p-value smaller than 0.05) are identified as the seeds for patch extraction. The mean p values in the seed voxels' enclosing patches of size $w \times w \times w$ are then used to sort the patch seeds. Based on their ascending order, we select class-discriminative patches in a greedy manner with the condition that no candidate patch pair should have more than 50% overlapping volume. Unlike [22] where patches were extracted with a single size (i.e., $w = 11$ of voxels sized $4 \times 4 \times 4$ mm^3), we adopt a multi-resolution strategy with three different patch sizes: $w = 11, 17, 23$, aiming to capture more useful information at different scales for AD/MCI diagnosis. Finally, after three rounds of patch selection with different w, we obtain a set of 3D local patches at three levels of resolutions. The corresponding patch-wise average GM densities, denoted as $P_{GM} = \{P^1_{GM}, \cdots, P^k_{GM}, \cdots, P^{K_1}_{GM}\}$, make a cross-sectional feature vector, where K_1 is the total number of legitimate patches.

Our longitudinal features are obtained based on the estimated voxel deformations matching the baseline and follow-up MRIs for each subject. A diffeomorphic registration method provided via ANTs package [2] is utilized to generate the deformation vector fields. To minimize the effect of the soft-tissue shifts outside the brains, a dilated *MIDAS Whole Brain Mask* for each subject is used to specify the registration area for ANTs. We then calculate the magnitude (or length) of the deformation vector at each voxel, and a 3D scalar field of deformation magnitudes (DM) is obtained. Based on the DM fields, which show the longitudinal atrophy, we conduct the same multi-resolution patch extraction as for the cross-sectional GM features, resulting in a set of 3D local patches at three levels of resolutions along with the local average DM of each patch, denoted as $P_{DM} = \{P^1_{DM}, \cdots, P^k_{DM}, \cdots, P^{K_2}_{DM}\}$.

4.2 Wrapper Feature Selection

In our experiments, the above patch extraction steps return more than 1000 discriminative patches for GM and DM each. To reduce the dimensionality of the feature vectors and avoid overfitting from redundant information, we propose to conduct a wrapper feature selection

[12] with a greedy forward searching from the extracted GM and DM patch pools. Starting from an empty set, the greedy-forward wrapper method iteratively adds a new patch each time that would lead to the largest improvement in classification on test dataset, until the classification performance over the current set starts to degrade. While the wrapper feature selection does not limit the choice of the wrapped classifier, they have to be chosen consistently to work well. Since our proposed nonlinear metric learning model is under the nearest neighbor paradigm, we choose kNN (with $k = 1$) as the wrapped classifier, which also very well reflects the intrinsic structures of the data samples. The wrapper feature selections are conducted from the pools of P_{GM}, P_{DM} and their union, resulting in three different types of features: "GM only", "DM only", and "Joint GM & DM", which can be directly fed into various of classifiers, including our ML-TPS model.

5 Experiments and Results

The effectiveness of our ML-TPS model and the feature extraction strategy is evaluated in this section, through two binary classification problems: AD vs. NC, and MCI vs. NC. The performance of various classification solutions is compared based on three measures: classification accuracy (ACC), i.e., the proportion of correctly classified subjects among the whole test set; sensitivity (SEN), i.e., the proportion of correctly classified AD (or MCI) patients; and specificity (SPE), i.e., the proportion of correctly classified normal controls. In the end, we also compare our method with three state-of-the-art AD/MCI diagnosis solutions [6, 22, 39] that also use T1-weighted MRIs from the ADNI database.

5.1 Comparisons of Different Features

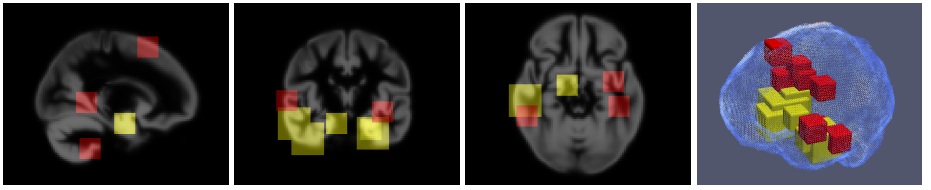
The first set of experiments is to investigate the efficacy of different features in distinguishing AD and MCI from normal controls. Specifically, the three types of features, i.e., "GM only", "DM only", and "Joint GM & DM" in Section 4 are evaluated based on three performance measures, ACC, SEN, and SPE. Both kNN and a kernel support vector machine (kSVM) are utilized for classification to reduce the potential bias introduced by any particular classifier. To better compare the classification performance, we run each experiment 100 times with different random 3-fold splits (two folds for training, one fold for testing). We choose $k = 1$ for kNN, and a Gaussian kernel for kSVM. The two hyper-parameters C and σ in the kSVM are tuned via 3-fold inner cross validation (CV) respectively from $\{2^{-15} \sim 2^{15}\}$.

		AD versus NC			MCI versus NC		
Classifier	Feature	ACC(%)	SEN(%)	SPE(%)	ACC(%)	SEN(%)	SPE(%)
kNN	GM only	85.9	78.7	91.1	77.8	73.6	82.0
	DM only	84.3	80.9	87.0	78.1	74.9	81.5
	Joint GM & DM	**88.4**	**84.7**	**91.6**	**79.3**	**76.5**	**82.2**
kSVM	GM only	85.2	80.2	**88.9**	74.7	72.8	76.5
	DM only	82.6	80.4	84.3	70.4	65.0	75.7
	Joint GM & DM	**87.1**	**85.7**	88.1	**75.3**	**73.1**	**77.5**

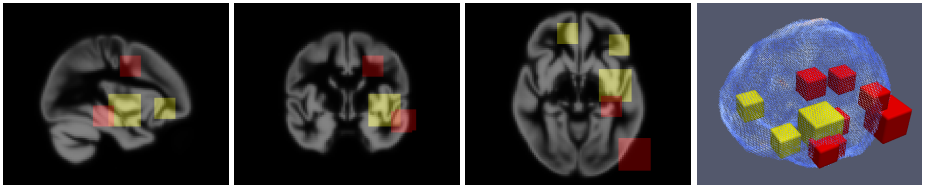
Table 1: Comparisons of the three different features for AD vs. NC and MCI vs. NC classifications. Boldface denotes the best performance for each classifier.

The classification results based on the three different features, averaging over the 100 runs, are summarized in Table 1. It is evident that the idea of combining longitudinal and

baseline features paid off – "Joint GM & DM" feature has generally improved the classification performance over the two single feature types, "GM only" and "DM only", for both *k*NN and kSVM. Note that, since we used *k*NN as the wrapped classifier in the feature selection step, the performance of *k*NN in this set of experiments is generally better than the kernel SVM. We believe the performance of kSVM can be further improved if it is used as the wrapped classifier.



(AD versus NC)



(MCI versus NC)

Figure 2: Visualization of selected patches in the "Joint GM & DM" feature for AD vs. NC and MCI vs. NC classifications. The columns from left to right are: sagittal, coronal, axial and 3D views; yellow and red colors indicate they are originally GM or DM patches respectively.

To better interpret the features, we further visualize the "Joint GM & DM" patches that are selected for AD vs. NC and MCI vs. NC classifications. Fig. 2 shows the selected cubic patches, which originally belong to either the GM (yellow) or the DM (red) patch pools. It can be observed that the most discriminative brain areas detected by our strategy include hippocampus, parahippocampal gyrus, entorhinal cortex, and amygdala, which are consistent with the findings in the literature [6, 22, 39]. Furthermore, we find that the overlapped areas of the GM (yellow) and DM patches (red) are quite small, which somewhat indicates the two types of features, when combined together, are rather complementary and working cooperatively in identifying the cross-sectional and longitudinal disparities among patient groups.

5.2 Comparisons of ML-TPS with other ML methods

The second set of experiments is to test the effectiveness of our proposed nonlinear metric learning model in improving AD/MCI versus NC classifications [2]. Specifically, we compare the improvements over the baseline classifier made by ML-TPS against five state-of-the-art metric learning methods: Neighborhood Components Analysis (NCA) [18], Information-Theoretic Metric Learning (ITML) [7], Large Margin Nearest Neighbor classification (LMNN) [34], multi metric LMNN (mm-LMNN) [35] and Parametric Local Metric Learning (PLML)

[2]We also conducted a set of experiments on the UCI repositories to show the effectiveness of our nonlinear ML-TPS, and the experimental results are included in the supplementary materials.

[33] methods. As mentioned before, ML methods estimate feature transformations, and can be used as the preprocessed step for any metric-based classifiers and also SVM [37]. Here, we choose kNN (with $k = 1$) as the baseline classifier, and use the "Joint GM & DM " feature for classification, due to their superior performance over other features (shown in Section 5.1).

	AD versus NC			MCI versus NC		
Classifier	ACC(%)	SEN(%)	SPE(%)	ACC(%)	SEN(%)	SPE(%)
kNN	88.4	84.7	91.6	79.3	76.5	82.2
NCA + kNN	84.1	79.1	87.8	75.2	74.9	75.5
ITML + kNN	86.7	82.4	90.0	76.4	77.3	75.0
LMNN + kNN	84.2	78.2	88.9	74.7	75.5	74.0
mm-LMNN + kNN	84.0	80.6	86.5	76.5	76.3	76.7
PLML+ kNN	83.4	79.4	86.6	71.5	69.3	73.9
ML-TPS + kNN	**90.5**	**84.9**	**94.1**	**81.6**	**79.5**	**83.7**

Table 2: Performance comparision of ML-TPS with other ML methods for AD vs. NC and MCI vs. NC classifications. Boldface denotes the best performance for each measure.

We adopt the same performance measures (ACC, SEN, SPE) and experimental setting (3-fold splits with 100 runs) as in Section 5.1. The hyper-parameters of NCA, ITML, LMNN and mm-LMNN are set by following [7, 18, 34, 35] respectively. PLML has a number of hyper-parameters, so we follow the suggestion of [33]: use a 3-fold CV to select α_2 from $\{0.01 \sim 1000\}$, and set the other hyper-parameters by its default. In the proposed ML-TPS model, there are two hyper-parameters: the number of anchor points p and the weighting factor λ. For p, we empirically set it to 30% of the training samples; for λ, we select it through CV from $\{5^{-5} \sim 5^{25}\}$. The classification results of each method for AD/MCI versus NC are summarized in Table 2.

As we can see from the results, our ML-TPS has the best classification performance with the highest ACC, SEN, SPE for both AD vs. NC and MCI vs. NC, which means it has improved the overall performance of the baseline kNN classifier. Especially for AD vs. NC classification, it reduces the error rate of kNN from 11.6% to 9.5%. This improvement is quite significant, considering the *in vivo* diagnostic error rate for AD is believed to be around $8 \sim 10\%$ [14]. It is worth noting that the other five ML methods all fail to improve the performance of kNN on this AD/MCI versus NC classification task. We believe it is because the underlying linear or piecewise linear feature transformations they adopt are not powerful enough to account for the complicated data patterns in AD/MCI vs NC, where the group boundaries are fuzzy and highly nonlinear. While the results in Table 2 show otherwise, the five competing ML solutions, in general, can greatly improve the classification rates over kNN. This has been demonstrated in many machine learning studies. Experiments on the popular UCI repositories, enclosed in the supplementary materials, provide a side evidence.

5.3 Comparisons with state-of-the-art AD staging methods

Numerous solutions [6, 22, 29, 39] have been proposed in the literature for AD/MCI/NC patient classification. Some very recent works [28, 29] reported rather high classification rates through the applications of multi-modality information integration (mainly MRIs and PETs) and sophisticated multi-classifier decision fusion schemes. However, direct comparisons of the methods are often not feasible, unless common subjects, datasets and modalities are employed, as in the evaluation project conducted by Cuingnet *et al*. [6].

Method	Study Size	Feature	Classifier	AD versus NC			MCI versus NC		
				ACC	SEN	SPE	ACC	SEN	SPE
Cuingnet *et al*. [6]	475	Voxel-wise GM	SVM	88.6	81.0	**95.0**	81.2	73.0	**85.0**
Zhang *et al*. [39]	202	93 ROI GMs	SVM	86.2	**86.0**	86.3	72.0	78.5	59.6
Liu *et al*. [22]	652	Patch-wise GM	SVM	86.4	83.9	88.6	79.4	79.2	79.5
Proposed method	338	Joint GM&DM	ML-TPS+kNN	**90.5**	84.9	94.1	**81.6**	**79.5**	83.7

Table 3: Comparision of the proposed method with other existing methods for AD vs. NC and MCI vs. NC classifications. Boldface denotes the best performance for each measure.

In light of this, we choose three recent solutions, which are very close in nature to our model, as the competing methods: 1) voxel-wise GM densities based method [19] which obtained the best performance among the ten methods evaluated in [6]; 2) 93-region GM densities method [39], and 3) single classifier patch-wise GM method in [22]. Similar to our method, they all use MR images as the sole information source and rely on certain single classifier for classification. The comparison results are shown in Table 3. It is remarkable that our model has the highest accuracies for both classification of AD vs. NC and MCI vs. NC, especially considering that the underlying classifier in our model is the extremely simple 1NN and the size of our studied set is relatively small. This would serve as another side evidence for the power of the proposed combination of nonlinear metric learning with dynamic longitudinal features.

6 Conclusion

In this paper, we have proposed a nonlinear metric learning method together with a multi-resolution patch based feature extraction strategy for MR brain image based diagnosis of AD and MCI. The proposed nonlinear metric learning learns a globally smooth nonlinear transformation for the feature space, which generalizes the linear model that can be used to improve various classifiers. The integration of the longitudinal atrophy information is carried out within the proposed feature extraction step, which largely improves the classification of AD/MCI versus NC when working with the baseline information. The geometric model used in this paper is thin-plate spline, and it can be extended to other radial distance functions. To explore other types of geometric models, as well as different ways to integrate the longitudinal feature is the direction of our future efforts.

References

[1] Alzheimer's Association et al. 2013 alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 9(2):208, 2013.

[2] Brian B Avants, Nick Tustison, and Gang Song. Advanced normalization tools (ants). *Insight J*, pages 1–35, 2009.

[3] A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

[4] Haili Chui and Anand Rangarajan. A new point matching algorithm for non-rigid registration. 89(2–3):114–141, 2003.

[5] Marie Chupin, Alexander Hammers, Rebecca SN Liu, Olivier Colliot, J Burdett, Eric Bardinet, John S Duncan, Line Garnero, and Louis Lemieux. Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation. *Neuroimage*, 46(3):749–761, 2009.

[6] Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali, Olivier Colliot, Alzheimer's Disease Neuroimaging Initiative, et al. Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. *neuroimage*, 56(2):766–781, 2011.

[7] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

[8] Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977.

[9] Yong Fan, Dinggang Shen, Ruben C Gur, Raquel E Gur, and Christos Davatziko. Compare: classification of morphological patterns using adaptive regional elements. *Medical Imaging, IEEE Transactions on*, 26(1):93–105, 2007.

[10] K.J. Friston, J.T. Ashburner, S.J. Kiebel, T.E. Nichols, and W.D. Penny. *Statistical parametric mapping: the analysis of funtional brain images*. Elsevier/Academic Press, 2007.

[11] Amir Globerson and Sam Roweis. Metric learning by collapsing classes. In *Nips*, volume 18, pages 451–458, 2005.

[12] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[13] Yujie He, Wenlin Chen, Yixin Chen, and Yi Mao. Kernel density metric learning. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 271–280. IEEE, 2013.

[14] Chris Hinrichs, Vikas Singh, Jiming Peng, and Sterling Johnson. Q-mkl: Matrix-induced regularization in multi-kernel learning with applications to neuroimaging. In *Advances in neural information processing systems*, pages 1421–1429, 2012.

[15] Steven CH Hoi, Wei Liu, Michael R Lyu, and Wei-Ying Ma. Learning distance metrics with contextual constraints for image retrieval. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2072–2078. IEEE, 2006.

[16] Yi Hong, Quannan Li, Jiayan Jiang, and Zhuowen Tu. Learning a mixture of sparse distance metrics for classification and dimensionality reduction. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 906–913. IEEE, 2011.

[17] Clifford R Jack, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

[18] Sam Roweis Jacob Goldberger and Ruslan Salakhutdinov Geoff Hinton. Neighbourhood components analysis. *NIPS'04*, 2004.

[19] Stefan Klöppel, Cynthia M Stonnington, Carlton Chu, Bogdan Draganski, Rachael I Scahill, Jonathan D Rohrer, Nick C Fox, Clifford R Jack, John Ashburner, and Richard SJ Frackowiak. Automatic classification of mr scans in alzheimer's disease. *Brain*, 131(3):681–689, 2008.

[20] James T Kwok and Ivor W Tsang. Learning with idealized kernels. In *ICML*, pages 400–407, 2003.

[21] Zhiqiang Lao, Dinggang Shen, Zhong Xue, Bilge Karacali, Susan M Resnick, and Christos Davatzikos. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. *Neuroimage*, 21(1):46–57, 2004.

[22] Manhua Liu, Daoqiang Zhang, and Dinggang Shen. Hierarchical fusion of features and classifier decisions for alzheimer's disease diagnosis. *Human brain mapping*, 35 (4):1305–1319, 2014.

[23] Sidong Liu, Weidong Cai, Lingfeng Wen, and Dagan Feng. Neuroimaging biomarker based prediction of alzheimer's disease severity with optimized graph construction. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 1336–1339. IEEE, 2013.

[24] Yung-Kyun Noh, Byoung-Tak Zhang, and Daniel D Lee. Generative local metric learning for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 1822–1830, 2010.

[25] Deva Ramanan and Simon Baker. Local distance functions: A taxonomy, new algorithms, and an evaluation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(4):794–806, 2011.

[26] Karl Rohr, H Siegfried Stiehl, Rainer Sprengel, Thorsten M Buzug, Jürgen Weese, and MH Kuhn. Landmark-based elastic registration using approximating thin-plate splines. *Medical Imaging, IEEE Transactions on*, 20(6):526–534, 2001.

[27] Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. In *NIPS*, 2003.

[28] Heung-Il Suk and Dinggang Shen. Deep learning-based feature representation for ad/mci classification. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pages 583–590. Springer, 2013.

[29] Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. *NeuroImage*, 101 (0):569 – 582, 2014.

[30] Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. *NeuroImage*, 101:569–582, 2014.

[31] Lorenzo Torresani and Kuang-chih Lee. Large margin component analysis. *Advances in neural information processing systems*, 19:1385, 2007.

[32] Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

[33] Jun Wang, Alexandros Kalousis, and Adam Woznica. Parametric local metric learning for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 1601–1609, 2012.

[34] Kilian Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems*, 18:1473, 2006.

[35] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

[36] Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. Distance metric learning with application to clustering with side-information. *NIPS*, pages 521–528, 2003.

[37] Zhixiang Xu, Kilian Q Weinberger, and Olivier Chapelle. Distance metric learning for kernel machines. *arXiv preprint arXiv:1208.3422*, 2012.

[38] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2, 2006.

[39] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Multimodal classification of alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3):856–867, 2011.