Nonlinear Feature Transformation and Deep Fusion for Alzheimer's Disease Staging Analysis

Yani Chen[1], Bibo Shi[1], Charles D. Smith[2], and Jundong Liu[1(✉)]

[1] School of Electrical Engineering and Computer Science, Ohio University, Athens, USA
liu@cs.ohio.edu
[2] Department of Neurology, University of Kentucky, Lexington, USA

Abstract. In this study, we develop a novel nonlinear metric learning method to improve biomarker identification for Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI). Formulated under a constrained optimization framework, the proposed method learns a smooth nonlinear feature space transformation that makes the input data points more linearly separable in SVMs. 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 deep network based feature fusion strategy through stacked denoising sparse autoencoder (DSAE) is adopted to integrate cross-sectional and longitudinal features estimated from MR brain images. Using the ADNI dataset, we evaluate the effectiveness of the proposed feature transformation and feature fusion strategies and demonstrate the improvements over the state-of-the-art solutions within the same category.

1 Introduction

The Alzheimer's Disease Neuroimaging Initiative (ADNI) [1] has provided a wealth of new data including structural and functional MR images to support the research on intervention, prevention and treatments of AD. Significant research efforts have been conducted using ADNI data to identify neuroimage biomarkers for the diagnoses of AD/MCI and various mixed pathologies. There is a pressing need to refine the solutions for patient classification as well as feature extraction, selection and fusion.

Many pattern classification algorithms rely on Euclidean metrics to compute pairwise dissimilarities, with equal weights assigned to the feature components. Replacing Euclidean with a metric learned from the inputs, which is equivalent to learn a feature transformation [2], can often improve the algorithm's performance significantly [2,3]. Depending on the feature space transformation to be sought, metric learning (ML) can be divided into linear and nonlinear groups [3]. Linear models commonly try to estimate a "best" affine transformation to deform the feature space, such that the resulted Mahalanobis distance would well agree with the supervisory information brought by training samples. While easy to use and convenient to optimize, linear models show inherently

L. Zhou et al. (Eds.): MLMI 2015, LNCS 9352, pp. 304–312, 2015.
DOI: 10.1007/978-3-319-24888-2_37

limited expressive power and separation capability in handling data with nonlinear structures. Nonlinear models are usually designed through kernelization or localization of certain linear models. The idea of localization is to build a overall nonlinear metric through combination of multiple piecewise linear metrics that are learned based on either local neighborhoods or class memberships. Although the multi-metric strategies are more powerful in accommodating nonlinear structures, generalizing these methods to fit other classifiers than kNN is not trivial. To avoid non-symmetric metrics, extra cares are commonly needed to ensure the smoothness of the transformed feature space.

Other than learning distance metrics, feature extraction and fusion from the ADNI database is also in great need of further exploration. For structural features extracted from brain MRIs, cortical thickness [4], volumetry of brain strucrures [5,6] and voxel tissue probability maps [7,8] 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 in ADNI, "dynamic" longitudinal information such as the atrophies 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/MCI neuroimage biomarker identification along two directions: 1) feature space transformation through a novel nonlinear ML technique, and 2) extraction and integration of dynamic longitudinal atrophy features into the classification framework. The proposed ML solution is a generalization of linear ML through the application of a deformable geometric model — the thin-plate spline (TPS) - to transform the feature space in SVMs. Toward the integration of longitudinal information, we adopt a deep network model – multi-modal stacked denoising sparse autoencoder (DSAE), with both cross-sectional (baseline) and longitudinal atrophy features extracted from MR brain images.

2 TPS Metric Learning for Support Vector Machines (TML-SVM)

Since learning a metric is equivalent to learn a feature transformation [2], metric learning can be applied to SVM models [9,10]. However, the existing SVM-based ML models employ only linear transformations, limiting their capabilities in dealing with complex data. In this study, we propose a new nonlinear ML solution for SVMs, which is a direct generalization of linear metric learning through the application of deformable geometric models to transform the entire input space. We choose thin-plate splines (TPS) as the transformation model, as TPS are well-known for their remarkable versatility and representation power in accounting for high-order deformations. To our best knowledge, this is the first work that utilizes nonlinear dense transformations, or spatially varying deformation models in metric learning. Next, we will briefly describe the theoretical background of

the TPS in the general context of transformations, followed by the presentation of our proposed ML model.

TPS When utilized to align a set of n corresponding point-pairs $\mathbf{u}_i$ and $\mathbf{v}_i$, $(i = 1, \ldots, n)$, a TPS transformation is a mapping function $f(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ within a suitable Hilbert space $\mathcal{H}$, that matches $\mathbf{u}_i$ and $\mathbf{v}_i$, as well as minimizes a smoothness TPS penalty functional:

$$J_m^d(f) = \int ||\mathcal{D}^m f||^2 \mathrm{d}X = \sum_{a_1+\cdots+a_d=m} \frac{m!}{a_1! \ldots a_d!} \int \cdots \int (\frac{\partial^m f}{\partial x_1^{a_1} \ldots \partial x_d^{a_d}})^2 \prod_{j=1}^{d} \mathrm{d}x_j \quad (1)$$

where $\mathcal{D}^m f$ is the matrix of m-th order partial derivatives of f, with a_k being positive, and $\mathrm{d}X = \prod_{j=1}^{d} \mathrm{d}x_j$, where x_j are the components of $\mathbf{x}$. The classic solution of Eqn. (1) has a representation in terms of a radial basis function (TPS interpolation function),

$$f_k(\mathbf{x}) = \sum_{i=1}^{n} \psi_i G(||\mathbf{x} - \mathbf{x}_i||) + \boldsymbol{\ell}^T \mathbf{x} + c, \quad (2)$$

where $||.||$ denotes the Euclidean norm and $\{\psi_i\}$ are a set of weights for the nonlinear part; $\boldsymbol{\ell}$ 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. (1), is as follows:

$$G(\mathbf{x}, \mathbf{x}_i) = \ G(||\mathbf{x} - \mathbf{x}_i||) \propto \begin{cases} ||\mathbf{x} - \mathbf{x}_i||^{2m-d} \ln ||\mathbf{x} - \mathbf{x}_i||, & \text{if } 2m - d \text{ is even;} \\ ||\mathbf{x} - \mathbf{x}_i||^{2m-d}, & \text{otherwise.} \end{cases} \quad (3)$$

The TPS transformation for point interpolation, as specified in Eqn. (2), can be employed as the geometric model to deform the input space for nonlinear metric learning. Such a transformation would ensure certain desired smoothness as it minimizes the bending energy $J_m^d(f)$ in Eqn. (1). Within the metric learning setting, let $\mathbf{x}$ be one of the training samples in the original feature space $\mathcal{X}$ of d dimensions, and $f(\mathbf{x})$ be the transformed destination of $\mathbf{x}$, also of d dimensions. Through a straightforward mathematical manipulations [11], we can get $f(\mathbf{x})$ in matrix format:

$$f(\mathbf{x}) = L\mathbf{x} + \Psi \begin{pmatrix} G(\mathbf{x}, \mathbf{x}_1) \\ \cdots \\ G(\mathbf{x}, \mathbf{x}_p) \end{pmatrix} = L\mathbf{x} + \Psi \boldsymbol{G}(\mathbf{x}), \quad (4)$$

where L (size $d \times d$) is a linear transformation matrix, Ψ (size $d \times p$) is the weight matrix for the nonlinear parts, and p is the number of anchor points $(\mathbf{x}_1, \ldots, \mathbf{x}_p)$ to compute the TPS kernel. We can use all the training data points as the anchor points. However, in practice, p anchor points are extracted through k-medoids method under the consideration of reducing computational cost.

TML-SVM By utilizing the nonlinear TPS transformation, we formulate our model under the *Margin-Radius-Ratio* bounded SVM paradigm, similarly as in [10]. Given training dataset $\mathcal{X} = \{\mathbf{x}_i | \ \mathbf{x}_i \in \mathbb{R}^d, i = 1, \cdots, n\}$ together with

the class label information $y_i \in \{-1, +1\}$, our proposed TML-SVM aims to simultaneously learn the nonlinear transformation f as described in Eqn. (4) and a SVM classifier, which can be formulated as follows:

$$\begin{aligned} \min_{L,\Psi,\mathbf{w},b} \quad & J = \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{n} \xi_i + C_2 \|\Psi\|_F^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^T f(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1 \ldots n; \text{ (I \& II)} \\ & \|f(\mathbf{x}_i) - \mathbf{x}_c\|^2 \leq 1, \quad \forall i = 1 \ldots n; \text{ (III)} \\ & \sum_{i=1}^{p} \Psi_i^k = 0, \quad \sum_{i=1}^{p} \Psi_i^k \mathbf{x}_i^k = 0, \ \forall k = 1 \ldots d. \text{ (IV)} \end{aligned} \tag{5}$$

f is in the form of Eqn. (4); Ψ^k is the kth column of Ψ; $\mathbf{x}^k$ is the kth component of $\mathbf{x}$. Besides the components for the traditional soft margin SVMs, another component $\|\Psi\|_F^2$, the squared Frobenius norm of Ψ, is added to the objective function as a regularizer to prevent overfitting. C_1 and C_2 are two trade-off hyper-parameters. The first two nonequivalent constraints (I and II) are the same as used in traditional SVMs. The third nonequivalent constraint (III) is a unit-enclosing-ball constraint, which forces the radius of minimum-enclosing-ball to be unit in the transformed space and avoids trivial solutions. $\mathbf{x}_c$ is the center of all samples. The last two equivalent constraints (IV) are used to maintain the properties for TPS transformation at infinity.

To solve this optimization problem, we propose an efficient EM-like iterative minimization algorithm by updating $\{\mathbf{w}, b\}$ and $\{L, \Psi\}$ alternatively. The details of this algorithm can be found in the supplementary material (http://media.cs.ohio.edu/mlmi2015_supplementary.pdf).

3 Neuroimage Data and Feature Extraction

The neuroimage data used in this work were obtained from the ADNI database [1]. 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).

Recently, patch-level neuroimage features extraction and fusion [8,12] 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). Different from the existing work, we propose a strategy that utilizes patch extraction and deep network based feature fusion with longitudinal brain atrophy, which is one of the pathological hallmarks of AD, as an addition information source. Our proposed framework 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 deep network based feature fusion step to learn a fused feature representation from the extracted patch pools.

Patch Extraction: After spatially normalized into an International Consortium for Brain Mapping template (with the dimensions reduced to $79 \times 79 \times 95$ and the voxel sizes to $2 \times 2 \times 2\ mm^3$), each baseline M0-MRI was segmented into three brain tissues: gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). We choose the spatially normalized GM tissue densities from the baseline MRIs as the cross-sectional information source in our work. 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 $5 \times 5 \times 5$ are then used to sort the patch seeds. Based on their ascending order, we select the first 100 class-discriminative patches in a greedy manner with the condition that no candidate patch pair should have more than 50% overlapping volume. The corresponding patch-wise average GM densities consist a cross-sectional feature vector.

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 [13] is utilized to generate the deformation vector fields. 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 patch extraction as for the cross-sectional GM features, resulting in a set of 3D local patches along with the local average DM of each patch.

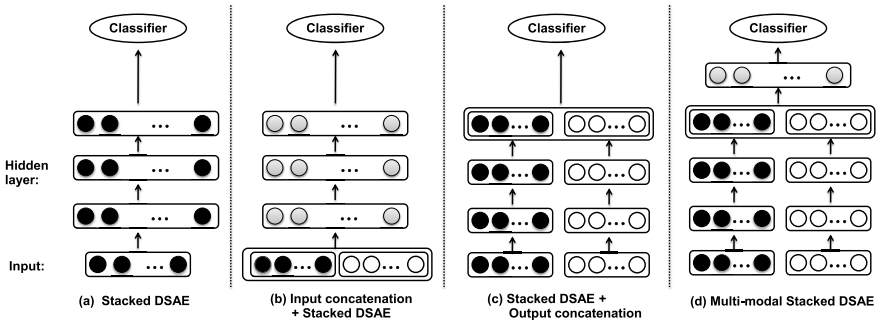

Fig. 1. Deep network structures of stacked DSAE, and three fusion strategies.

Feature Fusion: In our experiments, the above patch extraction steps return 100 discriminative patches for GM and DM respectively. Ideally, the two types of features should be fused with a reduced dimensionality. To this end, deep neural networks [4] provide very powerful solutions. Deep networks have been utilized in several recent AD/MCI works [6,12,14], with the same goal to learn a latent and compressed representation of the input feature vectors. Stacked Auto-encoder [6], Restricted Boltzman machine [12] and convolutional networks [14] are among the choices that have been examined. In this paper, we adopt a different model — stacked denoising sparse auto-encoder (DSAE), a direct

combination of both denoising and sparse auto-encoder [15,16]. This choice is based on the nature of the GM/DM features in our model. While easy to obtain, the GM/DM vectors contain many non-discriminative components, or in other words, high noise level.

On top of the stacked DSAE, several strategies are available to fuse different types of features, as shown in Fig. 1.(b)∼ (d): (b) shows the most intuitive way that concatenates different types of feature in the input layer, and learns a single deep neural network, as used in [6]; (c) learns separate deep neural networks for each feature type, and concatenates the output layers; (d) adds one more fully connected fusion layer on top of (c). In this paper, we choose the last strategy, the so-called multi-modal stacked DSAE as the solution to learn a discriminative fused feature representation.

4 Experiments and Results

In this section, we evaluate the proposed nonlinear feature transformation introduced by TML-SVM, as well as the integrated feature representation obtained via multi-modal stacked DSAE, 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), sensitivity (SEN), and specificity (SPE).

4.1 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"– features learned using stacked DSAE from only GM patches, "DM only"–features learned using stacked DSAE from only DM patches, and "Fused GM & DM"– features learned using multi-modal stacked DSAE from both GM and DM patches, are evaluated based on three performance measures, ACC, SEN, and SPE. To reduce the potential bias introduced by any particular classifier, here we utilize softmax regression model as the classifier, which is also regarded as "the classifier for stacked auto-encoder" [6]. Unlike other classifiers, softmax regression model makes "fine-tuning" deep networks for stacked DSAE and multi-modal stacked DSAE straightforward.

To better compare the classification performance, we run each experiment 10 times with different random 5-fold splits (three folds for training, one fold for validation, and one fold for testing). Similarly as in [6], we use three hidden layers for DSAE, with the numbers for the three layer's hidden nodes selected from $[100, 300, 500, 1000] - [50, 100] - [10, 20, 30]$ (bottom to up); the hyper-parameters for sparsity control and denoising corruption are both set to 0.2. For the fusion layer in multi-modal DSAE, the number of hidden nodes is selected from $[3, 5, 10, 20]$. The classification results based on the three different features, averaging over the 10 runs, are summarized in Table 1. It is evident that the idea of combining longitudinal and baseline features paid off – "Fused GM &

DM" feature has generally improved the classification performance over the two single feature types, "GM only" and "DM only", with the highest ACC, SEN and SPE for both AD vs. NC and MCI vs. NC.

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

		AD versus NC			MCI versus NC		
Classifier	Feature	ACC(%)	SEN(%)	SPE(%)	ACC(%)	SEN(%)	SPE(%)
Softmax	GM only	81.80	75.32	86.75	74.71	72.63	76.75
	DM only	80.49	75.13	84.56	70.21	68.46	71.93
	Fused GM & DM	**86.55**	**85.88**	**87.05**	**77.78**	**76.71**	**78.85**

4.2 Comparisons of Different Feature Fusion Strategies

The second set of experiments is to test the effectiveness of our adopted multi-modal stacked DSAE in improving AD/MCI versus NC classifications, with three other feature fusion strategies compared: 1) traditional "PCA based" strategy – concatenate the feature from GM and DM patches, and use Principle Component Analysis (PCA) to reduce the dimension with 99% variances kept; 2) "Input concatenation + Stacked DSAE", as shown in Fig. 1.(b); 3) "Stacked DSAE + Output concatenation", as shown in Fig. 1.(c); 4) our adopted multi-modal Stacked DSAE. We adopt the same performance measures (ACC, SEN, SPE), experimental setting (5-fold splits with 10 runs), and softmax regression classifier as in Section 5.1. The classification results of each strategy for AD/MCI versus NC are summarized in Table 2. As we can see from the results, the adopted multi-modal stacked DSAE has the best overall classification performance with the highest ACC, SEN for both AD vs. NC and MCI vs. NC. Although the 'Stacked DSAE + Output concatenation" strategy results in a slightly higher SPE values than multi-modal stacked DSAE, it is at the cost of sacrificing SEN values.

Table 2. Four feature fusion strategies for AD vs. NC and MCI vs. NC classifications.

		AD versus NC			MCI versus NC		
Classifier	Fusion strategy	ACC(%)	SEN(%)	SPE(%)	ACC(%)	SEN(%)	SPE(%)
Softmax	PCA based	79.77	73.30	84.72	53.71	53.96	53.46
	Input concat. + Stacked DSAE	80.82	76.42	84.12	75.26	72.89	77.57
	Stacked DSAE + Output concat.	85.58	81.32	**88.83**	75.46	71.82	**79.01**
	Multi-modal Stacked DSAE	**86.55**	**85.88**	87.05	**78.20**	**77.77**	78.60

4.3 Comparisons of TML-SVM with Other Classifiers

The last set of experiments is to test the effectiveness of the nonlinear feature transformation introduced by our proposed TML-SVM classifier in improving AD/MCI versus NC classifications. We compare TML-SVM against two other

classifiers without feature transformation: the softmax regression and the traditional SVM. For all the three classifiers, the same multi-modal Stacked DSAE are used to obtain the fused feature representation. It is worth noting that only the deep network in softmax regression model is fine-tuned. For SVM, the slackness coefficient C is selected from $\{2^{-5} \sim 2^{15}\}$. TML-SVM has three hyper-parameters to be tuned: the number of anchor points p and the tradeoff coefficients C_1 and C_2. For p, we empirically set it to 30% of the training samples; for C_1 and C_2, we select them from $\{2^{-5} \sim 2^{15}\}$ and $\{5^{-5} \sim 5^{25}\}$ respectively. We still adopt the same experimental setting and performance measures, and report the results averaged from 10 runs in Table 3.

Table 3. Comparisons of three different classifiers for AD vs. NC and MCI vs. NC classifications.

	AD versus NC			MCI versus NC		
Classifier	ACC(%)	SEN(%)	SPE(%)	ACC(%)	SEN(%)	SPE(%)
Softmax regression	86.55	85.88	87.05	78.20	77.77	78.60
SVM	85.76	82.11	88.54	76.83	74.60	79.04
Our proposed TML-SVM	**88.98**	**87.42**	**90.17**	**81.66**	**78.09**	**85.16**

As evident, our proposed TML-SVM has the best classification performance with the highest ACC, SEN, SPE for both AD vs. NC and MCI vs. NC. Especially, the improvements made by TML-SVM over the baseline classifier SVM is significant, which means adding the nonlinear feature transformation is effective in leading to a more separable feature space. Also, it is worth pointing out that the deep networks used in SVM and TML-SVM are not fine-tuned as in softmax regression model, and we believe the performance of our TML-SVM can be further improved if fine-tuning is utilized.

5 Conclusions

Our proposed AD/MCI vs. NC diagnosis solution consists of two major components: feature transformation through TML-SVM and feature fusion based on multi-modal stacked DSAE. TML-SVM learns a globally smooth deformation for the input space, and it is the first work that utilizes nonlinear dense transformations, or spatially varying deformation models in metric learning. Multi-modal stacked DSAE integrates longitudinal atrophy with baseline cross-sectional information, and it can be easily generalized to fuse other types of features. For AD/MCI vs. NC classification, some recent works [6,8,12] reported rather high classification rates through extracting features from different types of medical images (mainly MRIs and PETs) and sophisticated multi-classifier decision fusion schemes. To explore features from other data modalities and to enhance our TML-SVM with multi-kernelization are the directions of our ongoing efforts.

References

1. Jack, C.R., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging **27**(4), 685–691 (2008)
2. Bellet, A., Habrard, A., Sebban, M.: A survey on metric learning for feature vectors and structured data (2013). arXiv preprint arXiv:1306.6709
3. Yang, L., Jin, R.: Distance metric learning: A comprehensive survey, vol. 2. Michigan State Universiy (2006)
4. Klöppel, S., et al.: Automatic classification of mr scans in alzheimer's disease. Brain **131**(3), 681–689 (2008)
5. Chupin, M., et al.: Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation. Neuroimage **46**(3), 749–761 (2009)
6. Suk, H.-I., et al.: Latent feature representation with stacked auto-encoder for ad/mci diagnosis. Brain Structure and Function **220**(2), 841–859 (2013)
7. Fan, Y., et al.: Compare: classification of morphological patterns using adaptive regional elements. IEEE Transactions on Medical Imaging **26**(1), 93–105 (2007)
8. Liu, M., Zhang, D., Shen, D.: Hierarchical fusion of features and classifier decisions for alzheimer's disease diagnosis. Human brain mapping **35**(4), 1305–1319 (2014)
9. Xu, Z., et al.: Distance metric learning for kernel machines (2012). arXiv:1208.3422
10. Zhu, X., et al.: Learning similarity metric with svm. In: IJCNN (2012)
11. Chui, H., Rangarajan, A.: A new point matching algorithm for non-rigid registration **89**(2–3), 114–141 (2003)
12. Suk, H.-I., et al.: Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. NeuroImage **101**, 569–582 (2014)
13. Avants, B.B., et al.: Advanced normalization tools (ants). Insight J., 1–35 (2009)
14. Gupta, A., Ayhan, M., Maida, A.: Natural image bases to represent neuroimaging data. In: Proceedings of the 30th ICML, pp. 987–994 (2013)
15. Vincent, P., et al.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th ICML, pp. 1096–1103. ACM (2008)
16. Bengio, Y.: Learning deep architectures for ai. Foundations and trends® in Machine Learning **2**(1), 1–127 (2009)