# Distance-informed metric learning for Alzheimer's Disease Staging

Bibo Shi Zhewei Wang Jundong Liu
for the Alzheimers Disease Neuroimaging Initiative*
Schoole of Electrical Engineering and Computer Science
Ohio University, Athens, OH 45701, USA

*Abstract*— **Identifying intermediate biomarkers of Alzheimer's disease (AD) is of great importance for diagnosis and prognosis of the disease. In this study, we develop a new AD staging method to classify patients into Normal Controls (NC), Mild Cognitive Impairment (MCI), and AD groups. Our solution employs a novel metric learning technique that improves classification rates through the guidance of some weak supervisory information in AD progression. More specifically, those information are in the form of pairwise constraints that specify the relative Mini Mental State Examination (MMSE) score disparity of two subjects, depending on whether they are in the same group or not. With the imposed constraints, the common knowledge that MCI generally sits in between of NC and AD can be integrated into the classification distance metric. Subjects from the Alzheimer's Disease Neuroimaging Initiative cohort (ADNI; 56 AD, 104 MCI, 161 controls) were used to demonstrate the improvements made comparing with two state-of-the-art metric learning solutions: large margin nearest neighbors (LMNN) and relevant component analysis (RCA).**

## I. Introduction

Alzheimer's disease (AD) and its early stage, mild cognitive impairment (MCI), affect more than five million elderly people in the US [1]. Identifying reliable biomarkers to characterize different stages of AD would potentially provide objective and early measures for diagnosis and treatment monitoring 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.

Alzheimer's Disease Neuroimaging Initiative (ADNI) [2] provides reliable clinical data including MRI/ PET imaging to support the research on intervention, prevention and treatment of AD. Since the inception of ADNI in 2005, many efforts have been made in the research community to identify neuroimaging biomarkers that can differentiate AD, MCI and normal controls (NC). The common processing flow starts with feature selection, followed by either supervised classification or unsupervised clustering. Features that have been well explored include image modalities such as MRI and/or PET, tests of cerebrospinal fluid (CSF), neurological assessments scores and genetic information. For structural features extracted from brain MRIs, cortical thickness [3], hippocampal volume/shape [4], [5] and voxel tissue probability maps [6], [7] across the whole brain or around certain regions of interest (ROI), are among the popular choices. Various classifiers can then be trained and applied to determine the stage of the patients.

Most of the current studies of AD and MCI simplified the classification problem into two-class classification problems, i.e, AD vs. NC and/or MCI vs. NC. Although multiclass separation can be easily achieved through strategies such as one-vs-one and all-vs-all, the AD spectrum in reality is not a simple combination of three independent classes. The fact that MCI is usually found to be the early stage of AD, is seldom exploited or integrated within the current classification solutions. In addition, while many efforts have been devoted to identify and extract discriminative features, not so has gone into transforming the features or making the feature space better separable with domain knowledge incorporated.

Distance metric learning (DML), 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 side information is often formulated as pairwise constraints, e.g., pairs of similar and dissimilar data points. Many studies have demonstrated that a learned metric can significantly improve the performance in classification, clustering and retrieval tasks.

In this paper, we propose the application of metric learning solutions to transform patients' feature space. Subjects from the ADNI are used as both the training and evaluation sets. Our solution is a novel semi-supervised scheme that incorporates a sequential relationship among AD/MCI/NC classes. More specifically, we add pairwise constraints that specify the relative distance between a pair of patients, in accordance with their respective groups. With those constraints, the common knowledge that MCI generally sits in between of NC and AD can be imposed to update the classifier's distance metric. We name our solution *Distance Informed Metric Learning* (DIML) model. We will apply our approach, together two state-of-the-art supervised DML solutions, relevant component analysis (RCA) [8] and large margin nearest neighbors (LMNN) [9] to demonstrate the improvement that can be possibly made for AD patient classification through metric learning.

*Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

TABLE I. Feature information

| Feature # | Abbreviation | Type | Rank (AD vs NC) | Rank (MCI vs NC) |
|---|---|---|---|---|
| 1 | Ventricle | Volume | 6 | 5 |
| 2 | Hippocampus | Volume | 1 | 1 |
| 3 | Entorhinal | Volume | 2 | 2 |
| 4 | Fusiform | Volume | 7 | 7 |
| 5 | MidTemp | Volume | 5 | 6 |
| 6 | Vent-atropy | Atropy | 4 | 4 |
| 7 | Hipp-atropy | Atropy | 3 | 3 |
| 8 | Ento-atropy | Atropy | 11 | 10 |
| 9 | Fusi-atropy | Atropy | 10 | 11 |
| 10 | MidTemp-atropy | Atropy | 8 | 9 |
| 11 | WholeBrain-atropy | Atropy | 9 | 8 |
| 12 | Hipp-Dice-L | Shape | 16 | 16 |
| 13 | Hipp-Dice-R | Shape | 12 | 13 |
| 14 | Hipp-Dice-LR | Shape | 14 | 14 |
| 15 | Hipp-MI-L | Shape | 15 | 15 |
| 16 | Hipp-MI-R | Shape | 13 | 12 |

## II. METHOD

### A. ADNI data and features

Data used in this work were obtained from the ADNI database [2]. We selected all the subjects for whom the baseline (M0) and 12-month follow-up information (M12), including Hippocampus masks, are available. As a result, 321 subjects are selected: 56 patients with AD, 104 with MCI and 161 normal controls (NC).

To rank and choose the most discriminative features, sixteen candidates are used, which can be grouped into three categories, as shown in Table I. The first category consists of baseline volumes (at M0) for several important subcortical structures, including hippocampus. To minimize individual variations, structure volumes are normalized by total intracranial volume (TIV). The second category includes longitudinal atrophy percentage estimated based on baseline and follow-up volumes. The third category comprises of longitudinal hippocampal shape features, which are computed as the dissimilarity between the baseline (M0) and repeat (M12) hippocampal masks. Dice coefficient and Mutual Information (MI) are used as the indexes to measure the dissimilarities. After feature extraction, weight normalization is conducted to ensure all features are assigned with equal weights. As a result, the sample datasets are encoded into feature vectors $X \in \mathbf{R}^{m\times n}$ $(m = 16, n = 321)$ and subject target labels $Y \in \mathbf{R}^{1\times n}$ ($Y = 1$ for AD, 2 for MCI, 3 for NC).

To reduce the dimensionality and avoid overfitting from irrelevant features, we conduct a routine feature ranking through measuring Pearson Correlation Coefficient between individual features and the target labels [10]. The resulted feature rankings are shown in Table I for AD vs. NC, and MCI vs. NC, respectively.

### B. Metric learning solutions

In this paper, we choose relevant component analysis (RCA) [8] and large margin nearest neighbors (LMNN) [9] as the representatives of the state-of-the-art global and local solutions to demonstrate the power of metric learning. They also serve as motivations and comparisons to our proposed DIML approach.

**RCA** The basic idea of RCA is to make use of class-equivalent pairs to identify the global unwanted variability within the data. The relevant dimensions are estimated by chunklets, and within-chunklet variability is essentially recuded in an effort to assign large weights to relevant dimensions. The optimal solution for RCA can be obtained very efficiently in part due to its closed form expression.

**LMNN** LMNN is arguably the most widely-used local metric learning method. Unlike RCA and other global metric learning methods, LMNN defines the constraints in a local neighborhood, where the "pull force" within the class-equivalent data and the "push force" for the class-inequivalent data (the "imposters") are optimized to lead a balanced trade-off. A tailored numerical solver based on gradient descent and book-keeping strategy is utilized, which enables LMNN to perform efficiently in practice.

**DIML** Many popular metric learning solutions [11], [12], [8], [9], including RCA and LMNN, learn only from binary class-equivalent or inequivalent constraints. In reality, however, similarities tend to have different levels, and binary constraints often can not fully account for many situations occurring in practice.

It is commonly accepted that MCI is a transitional stage from normal aging to AD. In other words, the dissimilarities between NC/MCI and MCI /AD should be both smaller than that of NC/AD. A reliable feature that is well-known to possess the same trend is the Mini Mental State Examination (MMSE) test score. Typically the MMSE scores for normal cognitive aging are $\geq$ 27, while MCI are $(19 - 24)$, and $(< 19)$ for AD. Highly significant correlations between MMSE and hippocampal atrophy were also reported in [13]. Overall, MMSE scores characterize dementia progression very well, and the disparities among different stages can be measured and identified rather robustly and easily. These desired properties make MMSE scores an ideal side information to boost the performance of various classifiers through metric learning. In the following, we will formulate pairwise MMSE score disparities as an additional force to provide a helpful guidance for transforming the feature space.

Given a set of data instances $X = \{x_i | x_i \in \mathbf{R}^m, i = 1, \cdots, n\}$ and a prior distance matrix $D \in \mathbf{R}^{n\times n}$ for each instance pair $(x_i, x_j)$, our goal is the learn an optimal Mahalanobis matrix $M \in \mathbf{R}^{m\times m}$ that linearly transforms the original data, and at the same time preserves the prior pairwise distance in each local neighborhood. In this study, the "known" distance $D_{i,j}$ between subjects $i$ and $j$ is computed as the disparity of their group-wise MMSE scores. More specifically, we calculate the means and standard deviations for the three patient groups, and $D_{i,j}$ is set to the difference of the respective mean MMSE scores if $i$ and $j$ belong to different groups, and the group std otherwise.

With this setup, the optimal transformation $M$ can be obtained through the minimization of the following objective function:

$$\min_{M} \quad J(M) = \sum_{x_i \in X} \sum_{x_j \rightsquigarrow x_i} \|D_M(x_i, x_j)^2 - D_{i,j}^2\|^2 \qquad (1)$$
$$\text{s.t.} \quad M \succeq 0.$$

where $x_j \rightsquigarrow x_i$ denotes that $x_j$ is a neighbor of $x_i$, and $D_{i,j}$ is the prior distance between them. $D_M(x_i, x_j) = \sqrt{(x_i - x_j)^T \mathbf{M}(x_i - x_j)}$ is the Mahalanobis distance to be optimized. The similarity constraints are enforced in a local neighborhood setting, which is similar to that in LMNN. Comparing with global settings that consider all data instances, this approach has the advantage of being able to reduce the kNN error rate when handling data sets with multimodal support [9].

To solve this optimization problem, we adopt the same projected gradient approach utlized in [11], [12], [14]. Differentiaing $J(M)$ with respect to $M$, we get:

$$\frac{\partial J}{\partial M} = \sum_{x_i \in X} \sum_{x_j \rightsquigarrow x_i} 2(D_M(x_i, x_j)^2 - D_{i,j}^2)(x_i - x_j)(x_i - x_j)^T \qquad (2)$$

Accordingly $M$ is updated at each iteration $h$ by:

$$M^{h+1} = M^h - \mu \frac{\partial J}{\partial M} \qquad (3)$$

where $\mu$ is the step size. To ensure the positive semi-definite constraint $M \succeq 0$, a full eigenvalue decomposition is required at each iteration to project the obtained $M$ onto the convex set $C = \{M : M \succeq 0\}$.

## III. Experiments and Results

To evaluate our proposed DIML model for the AD staging problem, three sets of experiments were conducted using the ADNI data sets. We started with binary classifications (AD vs. NC and MCI vs. NC), followed by ternary classification experiments to separate AD/MCI/NC simultaneously. Comparisons were made with (1) kNN: k-nearest-neighbor classification using Euclidean metric (no metric learning involved); (2) RCA [8] and (3) LMNN, with default setting as in [9]. A leave-10%-out 10-fold cross-validation paradigm is adopted through each experiment.

### A. Binary AD vs. NC classification

Based on the feature ranking information from Table I, experiments for each of the four methods were conducted separately in different levels by using only the top 2 features (feature #2 and #3), top 4 (#2, #3, #7, and #6), and all of them. For each experiment, we calculated the sensitivity (SEN), specificity (SPE), positive predictive value (PPV), and negative predictive value (NPV). Since we are using 10-fold cross-validation, the number of subjects in each group is not the same. Thus, we use the balanced accuracy (Bala. ACC. = SEN/2 + SPE/2) to compare the performance between different classifiers.

The binary classification results for AD vs. NC are summarized in Table II, where the highest number in each performance measure is highlighted. From the "Bala ACC" column, it's evident that our DIML has the best overall accuracy than the other three methods. In [15], ten state-of-the-art AD patient classification solutions were evaluated and compared with common data sets. To further validate our DIML's performance, the average performance from the ten methods for AD/NC classification experiments was computed and included at the bottom row of Table II. As different data sets and feature sets were used in [15] and this study, direct comparison would not be possible. Nevertheless, we want to point out that all of the five performance measures obtained from DIML are above the average of the ten methods, which could be an indirect indication how well our method performs.

TABLE II. Binary classification results: AD vs. NC.

| Method | Feature | Bala. ACC. | SEN | SPE | PPV | NPV |
|---|---|---|---|---|---|---|
| KNN | Top 2 | 81.35% | 71.33% | 91.36% | 83.64% | 86.64% |
| | Top 4 | 82.59% | 72.00% | 93.18% | 85.50% | 86.35% |
| | ALL | 76.67% | 56.33% | **97.00%** | **94.64%** | 81.33% |
| RCA | Top 2 | 80.89% | 70.33% | 91.45% | 85.38% | 85.83% |
| | Top4 | 82.39% | 74.33% | 90.45% | 80.45% | 87.72% |
| | ALL | 81.46% | 70.67% | 92.24% | 85.28% | 86.03% |
| LMNN | Top 2 | 83.23% | 73.00% | 93.45% | 87.31% | 87.21% |
| | Top4 | 83.11% | 72.67% | 93.55% | 87.90% | 87.06% |
| | ALL | 81.93% | 69.67% | 94.18% | 88.83% | 85.77% |
| DIML | Top 2 | **83.55%** | 73.00% | 94.09% | 88.17% | 87.39% |
| | Top4 | 83.09% | **75.00%** | 91.18% | 83.64% | **88.01%** |
| | ALL | 82.52% | 72.67% | 92.36% | 84.83% | 86.86% |
| Ave. Perf. of [15] | | 80% | 71% | 89% | 85% | 79% |

### B. Binary MCI vs. NC classification

Similar binary classification experiments for MCI vs. NC were conducted, with results shown in Table III. The best performance is again produced by our DIML. One may notice that the SPE value is rather low (less than 70%) for all four methods, which is in part because the MCI subjects used in our experiments include both MCIc patients (MCI who converted to AD in the following) and MCInc (MCI who hadn't converted to AD) as defined in [15]. The heterogeneity within MCI subjects increases the difficulty to obtain a more accurate classification result. MCI/NC classification wasn't conducted in [15]. A similar work [13] that uses 12-month follow-up data sets reported $63\%$ for classification accuracy, $59\%$ for sensitivity and $71\%$ for specificity.

### C. Ternary AD/MCI/NC classification

Multiclass classification can commonly be decomposed into several binary classification tasks through one-vs-all (OVA), or all-vs-all (AVA) strategies, and subsequently solved by binary classifiers. However, this approach neglects the intrinsic connections among different classes, and lacks the capability of distinguishing different degrees of similarities. RCA and LMNN can handle multiclass data, but their internal implementations are both based on OVA that

TABLE III. Binary classification results: MCI vs. NC.

| Method | Feature | Bala. ACC. | SEN | SPE | PPV | NPV |
|---|---|---|---|---|---|---|
| KNN | Top 2 | 64.19% | 72.65% | 55.73% | 71.96% | 57.17% |
| | Top 4 | 59.01% | 67.10% | 50.91% | 68.65% | 48.49% |
| | ALL | 62.63% | 67.90% | 57.36% | 71.95% | 54.60% |
| RCA | Top 2 | 62.91% | 72.72% | 53.09% | 71.06% | 56.03% |
| | Top4 | 62.42% | 68.93% | 55.91% | 72.43% | 53.02% |
| | ALL | 61.23% | 71.54% | 50.91% | 69.15% | 55.90% |
| LMNN | Top 2 | 65.05% | 71.36% | 58.73% | 72.83% | 57.68% |
| | Top4 | 64.2% | 67.67% | 60.64% | 72.32% | 57.77% |
| | ALL | 64.20% | 71.58% | 56.82% | 72.29% | 57.56% |
| DIML | Top 2 | **71.56%** | **77.57%** | **65.55%** | **77.59%** | **69.25%** |
| | Top4 | 66.48% | 69.60% | 63.36% | 74.93% | 58.85% |
| | ALL | 64.94% | 75.15% | 54.73% | 72.54% | 59.40% |

only uses binary class-equivalent or inequivalent information. Our DIML, on the other hand, focuses on the integration of side information residing in the prior distance matrix. The multiclass membership information is implicitly preserved.

Ternary classification experiments of AD/MCI/NC were conducted using kNN, RCA, LMNN and DIML. Since all four methods can handle multiclass data, the classification results were generated without OVA or AVA for further processing. To evaluate multiclass classifiers, SEN, SPE, PPV and NPV are no longer applicable. Instead, we construct the cobweb graph based on the resultant confusion ratio matrices from the four methods, which provides a quick way to visualize classifier performance [16]. The results for 4 methods' performance, as well as chance performance are shown in Fig. 1 (due to space constraint, only the results from top2 features experiments are shown). A polygon within the chance performance hexagon indicates a better than chance performance. From Fig. 1, we can see that all four classification perform much better than chance for most cases, except MCI→ NC and MCI → AD, where they are comparable or worse than chance performance. Among the four methods, DIML's performance polygon always takes the inner bound, which implies the best ternary classification performance.

## IV. Conclusions

In this paper, we have described a distance-informed metric learning (DIML) solution for the AD staging problem. The novelty of our approach lies in the fact that it generalizes the binary class-equivalent or inequivalent constraints in traditional metric learning solutions to allow different levels of similarities among the data points. Currently, such similarities are input as a prior pairwise distance matrix, and the distances among ADNI subjects are specified as the disparity of their group-wise MMSE scores. To integrate other type of domain knowledge and explore other feature types would be the direction of our future efforts.

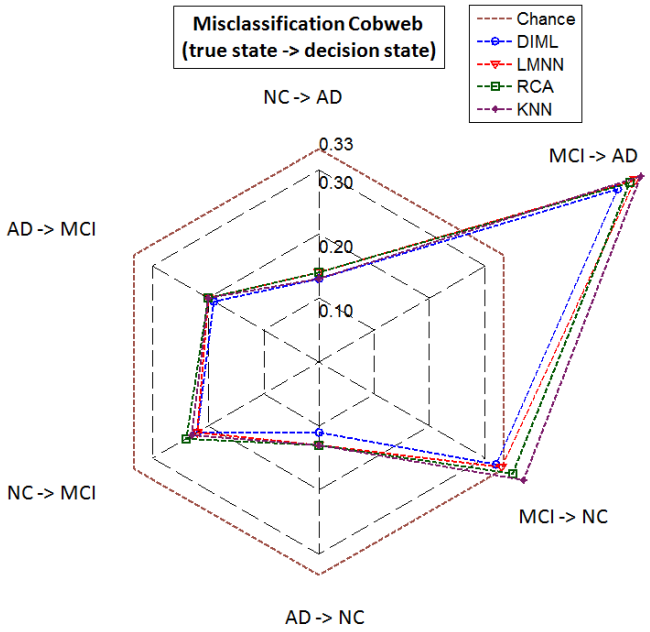


Fig. 1: Cobweb Graph - the misclassification performance for AD/MCI/NC ternary classification.

## References


[1] Alzheimer's Association et al., "2013 alzheimer's disease facts and figures.," *Alzheimer's & dementia: the journal of the Alzheimer's Association*, vol. 9, no. 2, pp. 208, 2013.

[2] C. Jack et al., "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of MRI*, vol. 27, no. 4, pp. 685–691, 2008.

[3] S. Klöppel et al., "Automatic classification of mr scans in alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[4] M. Chupin et al., "Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation," *Neuroimage*, vol. 46, no. 3, pp. 749–761, 2009.

[5] S. Liu et al., "Neuroimaging biomarker based prediction of alzheimer's disease severity with optimized graph construction," in *ISBI2013*. IEEE, 2013, pp. 1336–1339.

[6] Y. Fan et al., "Compare: classification of morphological patterns using adaptive regional elements," *Medical Imaging, IEEE Transactions on*, vol. 26, no. 1, pp. 93–105, 2007.

[7] Z. Lao et al., "Morphological classification of brains via high-dimensional shape transformations and machine learning methods," *Neuroimage*, vol. 21, no. 1, pp. 46–57, 2004.

[8] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons.," in *NIPS*, 2003.

[9] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[10] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[11] E. P. Xing et al., "Distance metric learning with application to clustering with side-information," *NIPS*, pp. 521–528, 2003.

[12] S. JacobGoldberger and R. GeoffHinton, "Neighbourhood components analysis," *NIPS*, 2004.

[13] R. Wolz et al., "Measurement of hippocampal atrophy using 4d graph-cut segmentation: application to adni," *NeuroImage*, vol. 52, no. 1, pp. 109–118, 2010.

[14] Y. Hong et al., "Learning a mixture of sparse distance metrics for classification and dimensionality reduction," in *ICCV2011*. IEEE, 2011, pp. 906–913.

[15] R. Cuingnet et al., "Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database," *Neuroimage*, pp. 766–781, 2011.

[16] A. Patel et al., "Comparison of three-class classification performance metrics: a case study in breast cancer cad," in *Medical imaging*, 2005, pp. 581–589.