CADDementia based on structural MRI using Supervised Kernel Metric Learning

D. Cárdenas-Peña, D. Collazos-Huertas, and G. Castellanos-Dominguez

Signal Processing and Recognition Group
Universidad Nacional de Colombia
Campus La Nubia, km 7 via al Magdalena, Manizales-Colombia
{dcardenasp, dfcollazosh, cgcastellanosd}@unal.edu.co

Abstract. Recently, computer-aided diagnosis tools for dementia using Magnetic Resonance Imaging scans have been successfully proposed to support reliable researches on intervention, prevention, and treatments of Alzheimer's disease. However, is necessary to improve the performance of classification machines. As an alternative, a supervised kernel framework for learning metrics that enhances conventional machines and supports the diagnosis of dementia is proposed. Therefore, the proposed metric learning produces discriminative features aimed at improving the discrimination neurological classes. The testing stage is carried out using the ADNI dataset to train the commonly used supervised classification machine, namely, support vector machine (SVM). Attained classification results (57.6% average accuracy) prove that our framework is able to discriminate dementia patients.

1 Introduction

For Alzheimer's Disease (AD), the use of structural magnetic resonance imaging (MRI) data became a suitable alternative to develop computer-aided diagnosis (CAD) tools due to its wide availability and non-invasiveness [2]. Nonetheless, the most of researches has been focusing on discriminating pathologies with a variety of classification tools from neuroimaging data, genetic information, and other biomarkers. Hence, insufficient attention has been given to build appropriate metrics from the training data that could maximize the performance of several classifiers [4].

Some approaches that introduce a metric learning stage into the MRI discrimination process are the following: In the case of linear models, [6] stacks PCA matrices and logistic regressors in a multi-layer architecture. However, the generative properties of the resulting machine highlight over the discriminative ones. In addition, when handling data distributions with nonlinear structures, linear models show inherently limited performance and class separation capability. On the other hand, the most popular nonlinear models are built through kernel-based methods. In [7] combines three different biomarkers using a simple-while-effective multiple-kernel learning for improving the SVM-based classification of AD and MCI. However, optimization of kernel weighting is carried out by a grid search, which is very time consuming when the number of features and samples gets large [3].

In order to enhance the MRI discrimination, we introduce a kernel-based metric learning framework for supporting the dementia diagnosis task. The proposed approach searches for projections into more discriminative spaces so that the resulting data distribution resembles as much as possible the label distribution. Hence, we incorporate kernel theory for assessing the affinity between projected data and available labels through the Center Kernel Alignment (CKA) criterion. To this end, we use morphological measurements (volume, area, and thickness) computed by the widely used `FreeSurfer` suite [5]. The proposed approach is tested on MRI data discrimination using dementia categories (namely, Normal Control (NC), Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD)).

2 Proposed Algorithm Description

2.1 Centered kernel alignment

Kernel functions are bivariate measures of similarity, which are based on the inner product between samples embedded in a Hilbert space. For an input feature space $\mathscr{X}$, a kernel $\kappa_X{:}\mathscr{X}\times\mathscr{X}\to\mathbb{R}^+$ is a positive-definite function that defines an implicit mapping $\varphi_X{:}\mathscr{X}\to\mathscr{H}_X$, aiming to embed a data point $x\in\mathscr{X}$ into the element $\varphi_X(x)\in\mathscr{H}_X$ of some Reproducing Kernel Hilbert Space (RKHS) (noted as $\mathscr{H}_X$). Within a supervised learning framework, a kernel $\kappa_L{:}\mathscr{L}\times\mathscr{L}\to\mathbb{R}^+$ is also introduced that acts over the target space $\mathscr{L}$ to account for the attribute labeling information so that κ_L defines the implicit mapping $\varphi_L(l){:}\mathscr{L}\to\mathscr{H}_L$. Due to each function ($\kappa_X$ and κ_L) reflects a different notion of similarity extracted from a distinct sample set, the concept of alignment between mappings can be introduced to measure the degree of agreement between the input and target kernels. To unify both tasks into a coherent optimization problem, we employ the Centered Kernel Alignment (CKA) that assesses the kernel affinity through the expected value of their normalized inner product over all data points as follows:

$$\rho\left(\kappa_X,\kappa_L\right)=\frac{\mathbb{E}_{xx'll'}\left\{\bar{\kappa}_X\left(x,x'\right)\bar{\kappa}_L\left(l,l'\right)\right\}}{\sqrt{\mathbb{E}_{xx'}\left\{\bar{\kappa}_X^2\left(x,x'\right)\right\}\mathbb{E}_{ll'}\left\{\bar{\kappa}_L^2\left(l,l'\right)\right\}}}, \tag{1}$$

where notation $\mathbb{E}_z\left\{\cdot\right\}$ stands for the expected value of the random variable z, $\bar{\kappa}_Z\left(z,z'\right)$ is the centered version of the kernel function

$$\kappa_Z\left(z,z'\right){=}\left(\varphi_Z(z)-\bar{\varphi}_Z\right)^\top\left(\varphi_Z(z')-\bar{\varphi}_Z\right) \tag{2}$$

being $\bar{\varphi}_Z\in\mathscr{H}_Z$ the expected value of the data distribution on $\mathscr{H}_Z$.

In practice, the characterizing kernel matrices, $\boldsymbol{K}_X\in\mathbb{R}^{N\times N}$ and $\boldsymbol{K}_L\in\mathbb{R}^{N\times N}$, are extracted from a provided input dataset $\boldsymbol{X}\in\mathbb{R}^{N\times P}$, holding samples $\boldsymbol{x}_n\in\mathbb{R}^P$, along with its corresponding target vector $\boldsymbol{l}{=}\{l_n\subset\mathbb{Z}{:}n\in[1,N]\}\in\mathbb{Z}^N$. Hence, the empirical estimate for the CKA value can be computed as follows:

$$\hat{\rho}\left(\bar{\boldsymbol{K}}_X, \bar{\boldsymbol{K}}_L\right) = \frac{\langle \bar{\boldsymbol{K}}_X, \bar{\boldsymbol{K}}_L\rangle_{\mathtt{F}}}{\sqrt{\langle \bar{\boldsymbol{K}}_X, \bar{\boldsymbol{K}}_X\rangle_{\mathtt{F}} \langle \bar{\boldsymbol{K}}_L, \bar{\boldsymbol{K}}_L\rangle_{\mathtt{F}}}}, \tag{3}$$

where notation $\langle\cdot,\cdot\rangle_{\mathtt{F}}$ stands for the matrix-based Frobenius product, and $\bar{\boldsymbol{K}}{=}(\boldsymbol{\varphi}_Z - \bar{\boldsymbol{\varphi}}_Z)^\top(\boldsymbol{\varphi}_Z - \bar{\boldsymbol{\varphi}}_Z)$ is the centered kernel matrix (associated with $\bar{\kappa}_Z(,))$ computed as $\bar{\boldsymbol{K}}{=}\bar{\boldsymbol{I}}\boldsymbol{K}\bar{\boldsymbol{I}}$, being $\mathbf{1}{\in}\mathbb{R}^{N\times 1}$ the all-ones vector, $\boldsymbol{I}$ the identity matrix, and $\bar{\boldsymbol{I}}{=}\left[\boldsymbol{I} - \mathbf{1}\mathbf{1}^\top\!/N\right]$.

Importantly, since the alignment estimates the agreement between $\mathscr{X}$ and $\mathscr{L}$ spaces through their statistical dependence $\rho{\in}[0,1]$, then, the larger the value of CKA, the more similar the distributions of the input and target data.

2.2 Supervised metric learning for classification

The CKA dependence using the Mahalanobis metric learning is developed for the commonly used SVM approach that are fed into a Gaussian kernel optimization.

Gaussian Kernel Optimization for Classification: In general, the Gaussian kernel is preferred in pattern classification applications since it aims at finding an RKHS with universal approximating ability, not to mention its mathematical tractability. Nonetheless, to account for the variance of each space when measuring the pairwise distance between samples $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$, the Gaussian kernel relies on the generalized Euclidean metric that is parameterized by a linear projection matrix $\boldsymbol{W}$ in the form:

$$\kappa_X\left(\boldsymbol{W},\sigma\right) = \exp\left(-\left(\boldsymbol{x}_n - \boldsymbol{x}_{n'}\right)\boldsymbol{W}\boldsymbol{W}^\top\left(\boldsymbol{x}_n - \boldsymbol{x}_{n'}\right)^\top/2\sigma^2\right) \tag{4}$$

where $\sigma{\in}\mathbb{R}^+$ is the kernel bandwidth that rules the observation window within the similarity distance is assessed.

In terms of the projection matrix, the formulation of the CKA-based optimizing function in Eq (3) can be integrated into the following kernel-based learning problem:

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}}\left\{-\log\left(\hat{\rho}\left(\bar{\boldsymbol{K}}_X(\boldsymbol{W}), \bar{\boldsymbol{K}}_L\right)\right)\right\}, \tag{5a}$$

$$= \arg\min_{\boldsymbol{W}}\left\{\log\left(\operatorname{tr}\left(\boldsymbol{K}_X(\boldsymbol{W})\bar{\boldsymbol{I}}\boldsymbol{K}_L\bar{\boldsymbol{I}}\right)\right) - \tfrac{1}{2}\log\left(\operatorname{tr}\left(\boldsymbol{K}_X(\boldsymbol{W})\bar{\boldsymbol{I}}\boldsymbol{K}_X(\boldsymbol{W})\bar{\boldsymbol{I}}\right)\right)\right\}, \tag{5b}$$

where the logarithm function is used for mathematical convenience.

Therefore, the first term in Eq (5b) assesses the similarity between input and target kernels while the second one works as a regularization term minimizing the norm of the input kernel.

3 Dataset and Preprocessing

For training the proposed metric learning framework, the ADNI dataset was employed. Specifically, a subset of 3304 MRI scans are considered from 896 subjects aged from 55 to 90 years (1048 NC, 1433 MCI, and 823 AD). Provided images are split into two subsets. The first one holds 30% of the subjects and is devoted to a blindfolded assessment of the performance framework. The remaining 70% of subjects is employed for framework parameter tuning, which is carried by a 5-fold cross-validation scheme to guarantee that all images of the same subject are assigned to a single group of data analysis (i.e., a validation fold or the test subset).

The set of structural MRIs is automatically pre-processed via the widely used `FreeSurfer` software package [1] that computes the needed morphological measurements with suitable test-retest reliability across scanner manufacturers and field strengths. As a result, an input feature matrix $\boldsymbol{X}$ with size N=3304 and P=310 is built using the features from each MRI. Namely, 69 features of Cortical Volumes (CV), 37 features of Subcortical Volumes (SV), and 68 features of Thickness Average (TA), Thickness Std (TS) and Surface Area (SA) set.

4 Results

For the sake of evaluation, the proposed methodology of training is contrasted against the classification results for dementia diagnosis results in [1]. To this end, the accuracy a, true-positive rate $\{\tau_{HC}, \tau_{MCI}, \tau_{AD}\}$, and area under the ROC (Receiver operating curve) AUC are recomputed following the same evaluation scheme, consisting in bootstrapping the test set with 1000 resamples to estimate the average and 95% confidence interval of each measure.

Table 1 report the obtained results of conventional and CKA-enhanced SVM. The darker cells denote the best performance in terms of each evaluation criteria and their respective confidence intervals (CI). As seen in the last row, the CKA-based metric learning improves every one of the evaluation measures. Therefore, it follows that the either case of the CKA metric learning gives rise to the classification performance.

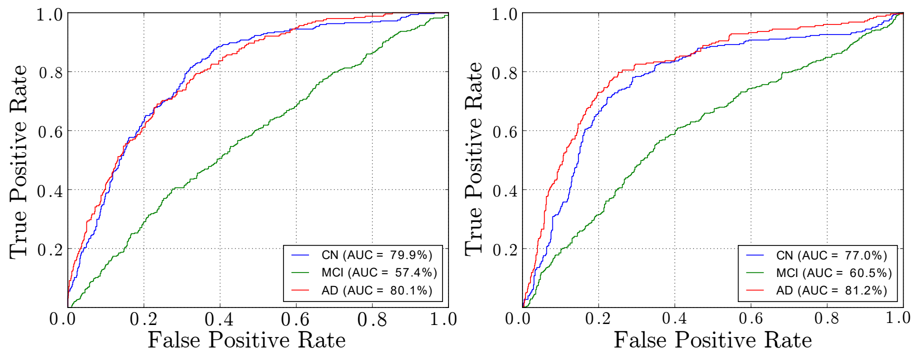
	a (CI)	τ_{HC} (CI)	τ_{MCI} (CI)	τ_{AD} (CI)
SVM	53.7 (50.4-56.7)	58.6 (53.3-63.6)	46.5 (42.1-50.9)	59.9 (53.5-66.4)
ML + SVM	57.6 (54.3-60.7)	62.7 (57.5-67.8)	54.2 (49.4-58.4)	57.1 (50.8-63.3)

Table 1: Accuracy performance measures following the validation scheme in [1]. First row displays the performance before CKA-based metric learning. Bottom row shows the performance after CKA.

Obtained class-wise Receiver Operating Characteristic (ROC) curve for ADNI dataset is depicted in Fig. 1. The Fig. 1 shows a lower area under the curve for the second class, as the Table 1 shows the lowest accuracy for that class. Both facts imply that MCI subjects are the most difficult to classify, which can be

[1] freesurfer.nmr.mgh.harvard.edu

due to the wide spread class distribution. From a morphological perspective, the low accuracy in MCI subjects can be related to nature of such class. Since MCI is an intermediate class between Healthy and Alzheimer's Disease classes, those subjects tend to be more misdiagnosed than the ones belonging to NC and AD.



(a) Baseline SVM   (b) ML + SVM

Fig. 1: Obtained ROC curves for testing MRIs in the ADNI dataset.

5 Conclusion

A supervised metric learning is introduced to support MRI classification. The proposed learning decodes discriminant information based on the maximization of the similarity between the input distribution and the corresponding target (diagnosis classes), aiming at enhancing the class separability. Furthermore, an SVM is trained using the metric learning framework for classifying three dementia categories (HC, MCI and AD). Evaluation of the proposed metric learning framework is carried out on the well-known ADNI dataset, where several morphological measurements are extracted using `FreeSurfer` to represent each MRI scan. Experimental results show that our proposed CKA improves the performance in terms of the classification accuracy and the true positive fraction of each neurological class. In particular, the ML+SVM classifier achieves the best performance (average 57.6%), and the baseline SVM reaches competitive results (53.7%). As future work, we plan to analyze other kinds of image representation strategies aiming at finding their relevance for class discrimination assessed by the CKA criterion. Finally, we note that the class-wise performance can be parameterized by the introduced kernel function in the target space so that a larger similarity of a particular class should increase its true positive rate.

References

1. Bron, E.E., Smits, M., Van Der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Orellana, C.M., Meijboom, R., et al.: Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: The caddementia challenge. NeuroImage **111** (2015) 562–579
2. Jack, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., et al.: Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. The Lancet Neurology **12**(2) (2013) 207–216
3. Liu, F., Zhou, L., Shen, C., Yin, J.: Multiple kernel learning in the primal for multimodal alzheimer's disease classification. IEEE journal of biomedical and health informatics **18**(3) (2014) 984–990
4. Shi, B., Chen, Y., Hobbs, K., Smith, C.D., Liu, J.: Nonlinear metric learning for alzheimer's disease diagnosis with integration of longitudinal neuroimaging features. In: BMVC. (2015) 138–1
5. Tustison, N.J., Cook, P.A., Klein, A., Song, G., Das, S.R., Duda, J.T., Kandel, B.M., van Strien, N., Stone, J.R., Gee, J.C., et al.: Large-scale evaluation of ants and freesurfer cortical thickness measurements. Neuroimage **99** (2014) 166–179
6. Wachinger, C., Reuter, M., Initiative, A.D.N., et al.: Domain adaptation for alzheimer's disease diagnostics. Neuroimage **139** (2016) 470–479
7. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Initiative, A.D.N., et al.: Multimodal classification of alzheimer's disease and mild cognitive impairment. Neuroimage **55**(3) (2011) 856–867