Joint High-Order Multi-Task Feature Learning to Predict the Progression of Alzheimer's Disease

Lodewijk Brand[1], Hua Wang[1(✉)], Heng Huang[2], Shannon Risacher[3], Andrew Saykin[3], and Li Shen[3,4] for the ADNI

[1] Department of Computer Science, Colorado School of Mines, Golden, CO, USA
lbrand@mines.edu, huawangcs@gmail.com
[2] Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA
heng.huang@pitt.edu
[3] Department of Radiology and Imaging Sciences, Department of BioHealth Informatics, Indiana University, Indianapolis, IN, USA
{srisache,asaykin}@iupui.edu
[4] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA
Li.Shen@pennmedicine.upenn.edu

Abstract. Alzheimer's disease (AD) is a degenerative brain disease that affects millions of people around the world. As populations in the United States and worldwide age, the prevalence of Alzheimer's disease will only increase. In turn, the social and financial costs of AD will create a difficult environment for many families and caregivers across the globe. By combining genetic information, brain scans, and clinical data, gathered over time through the Alzheimer's Disease Neuroimaging Initiative (ADNI), we propose a new *Joint High-Order Multi-Modal Multi-Task Feature Learning* method to predict the cognitive performance and diagnosis of patients with and without AD.

Keywords: Alzheimer's disease · Multi-modal · Longitudinal · Tensor

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative condition characterized by the progressive loss of memory and cognitive functions. The *Alzheimer's Association*

H. Wang—To whom all correspondence should be addressed.

ADNI—Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (ad-ni.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: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11070, pp. 555–562, 2018.
https://doi.org/10.1007/978-3-030-00928-1_63

recently released a report [1] in which they described various societal costs of AD in the United States. They found that in 2017 the total spending of caring for individuals with AD surpassed $259 billion. In addition, they report that 1 in 10 people aged 65 or older suffer from some form of Alzheimer's dementia. Given the widespread effects of AD on patients, their families, and caregivers, it is important that the scientific community investigates methods that can accurately predict the progression of AD.

Following the body of work done through the Alzheimer's Disease Neuroimaging Initiative (ADNI), we present a new joint regression and classification model, inspired by our previous works [13,14], that has shown great performance in the identification of relevant genetic and phenotypic biomarkers in patients with AD. Our newly proposed method consists of three major components as follows. First, we use the $\ell_{2,1}$-norm regularization [5] to effectively associate input features overtime and generate a sparse solution. Second, we utilize a new group ℓ_1-norm regularization proposed in our previous works [10–12,14] to globally associate the weights of the input imaging and genetic modalities, where a *modality* indicates a single data grouping (*e.g.* brain imaging data, genetic data, diagnostic data, *etc.*). The group ℓ_1-norm regularization is able to determine which input modality is most effective at predicting a particular output. Third, we incorporate the trace norm regularization [2,4,15,16] to determine relationships that occur within modalities.

2 Joint Multi-modal Regression and Classification for Longitudinal Feature Learning

Joint multi-task learning (*e.g.* performing regression and classification at the same time) can help discover more robust patterns than those discovered when the tasks are performed using separate objectives [13,14]. These robust patterns can arise when the learned parameters for the regression task become outliers for the classification task.

In the ADNI data set, a collection of input modalities (*e.g.* VBM, FreeSurfer, SNP) have been collected from patients in every six months. The input imaging features are represented by a set of matrices $\mathcal{X} = \{X_1, X_2, \ldots, X_T\} \in \mathbb{R}^{D \times n \times T}$. The stacked matrices in $\mathcal{X}$ correspond to measurements recorded at T consecutive time points. Each matrix $X_t \in \mathbb{R}^{D \times n}$ is composed of k input modalities where $X_t = [X_{11}, X_{12}, \ldots, X_{1k}]$. Each input modality X_{tj} consists of d_j features such that $D = \sum_{j=1}^{k} d_j$. $\mathcal{X}$ is a tensor with D imaging features, n samples, and T time points.

In addition to the input modalities, the ADNI also collected cognitive information from each patient. The output of our model, a prediction of cognitive diagnoses and scores, is represented by the tensor $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_T\} \in \mathbb{R}^{n \times c \times T}$ where at each time point t from $(1 \leq t \leq T)$ a matrix $Y_t = [Y_{tr}\ Y_{tc}]$ represents the horizontal concatenation of the clinical diagnoses (classification tasks) and cognitive scores (regression tasks) of each patient who participated in the ADNI study.

In order to associate the longitudinal imaging markers and the genetic markers to predict cognitive scores and diagnoses over time, we introduce a tensor implementation of the widely used $\ell_{2,1}$-norm:

$$\left\|\mathcal{W}_{(1)}\right\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} \left\|\mathbf{w}_t^i\right\|_2^2}, \tag{1}$$

where $\mathbf{w}_t^i$ denotes the i-th row of the coefficient matrix W_t at time t. Here we define $\mathcal{W}_{(n)}$ as the unfolding operation of $\mathcal{W}$ along the n-th mode. Given this definition, it follows that $\mathcal{W}_{(1)} = [W_1\ W_2\ \ldots\ W_T] \in \mathbb{R}^{d\times(c\times T)}$. The $\ell_{2,1}$-norm regularization in Eq. (1) will ensure that each feature will either have small, or large values, over the longitudinal dimension.

In heterogeneous feature fusion, the features of a specific input modality can be more discriminative than others for a given task. For example, the features associated with the brain imaging modality may be more useful in determining cognitive scores than the corresponding genetic modality. Conversely, the genetic modality may be more discriminative in predicting a disease diagnosis. To incorporate this global relationship between modalities we use the group ℓ_1-norm (G_1-norm) proposed in our previous works [10,11,14]:

$$\left\|\mathcal{W}_{(1)}\right\|_{G_1} = \sum_{i=1}^{c} \sum_{j=1}^{k} \left\|\mathbf{w}_j^i\right\|_2 , \tag{2}$$

where k is the number of input modalities.

The regularizations defined above in Eqs. (1–2) couple the learning tasks over time and learn the relative significance of each input modality for a given task. We know, as AD develops, that many cognitive measures are related to one another. This kind of correlation, when combined with a multivariate regression model and the hinge loss from a support vector machine (SVM) classifier, can be modeled by minimizing the rank of the unfolded coefficient matrix $\mathcal{W}$ in the following objective:

$$\begin{aligned}\min_{\mathcal{W}} J_2 = &\sum_{t=1}^{T} \left\|X_t^T W_{tr} - Y_{tr}\right\|_F^2 + \sum_{t=1}^{T} h(X_t, Y_{tc}) \\ &+ \gamma_1 \left\|\mathcal{W}_{(1)}\right\|_{2,1} + \gamma_2 \left\|\mathcal{W}_{(1)}\right\|_{G_1} + \gamma_3 \left\|\mathcal{W}_{(1)}\right\|_* ,\end{aligned} \tag{3}$$

where $\|M\|_* = Tr(MM^T)^{1/2}$ denotes the trace norm of the matrix $M \in \mathbb{R}^{n\times m}$, which has been shown as the best convex approximation of the rank-norm [2]. The rank minimization will develop joint correlations across each of the learning tasks at different time points. We call J_2 in Eq. (3) the *Joint High-Order Multi-Modal Multi-Task Feature Learning* model. We will use this newly proposed model to effectively predict the cognitive scores and diagnoses of AD patients.

The algorithm to solve the proposed objective in Eq. (3) is summarized in Algorithm 1. Due to the space limit, the derivation of this algorithm and the rigorous proof of its global convergence will be supplied in an extended journal version of this paper.

Algorithm 1: A new algorithm to minimize J_2 in Equation (3)

Data: $\mathcal{X} = \{X_1, X_2, ..., X_T\} \in \mathbb{R}^{D \times n \times T}$, $\mathcal{Y} = [\mathcal{Y}_r, \mathcal{Y}_c] = \{Y_1, Y_2, ..., Y_T\} \in \mathbb{R}^{n \times c \times T}$.
1. Initialize $\mathcal{W}^{(0)} = [\mathcal{W}_r^{(0)}, \mathcal{W}_c^{(0)}] \in \mathbb{R}^{D \times c \times T}$ where $\mathcal{W}_r^{(0)} \in \mathbb{R}^{D \times c_r \times T}$ is generated using the regression results (Y_{tr}) at each individual time point and $\mathcal{W}_c^{(0)} \in \mathbb{R}^{D \times c_c \times T}$ is derived from T multi-class SVMs fit to Y_{tc}.
while *not converges* **do**
  2. Calculate the diagonal matrices D_r and D_c where the k-th diagonal element is computed as $D_j(i,i) = \frac{1}{2\sqrt{\sum_{t=1}^{T} \|\mathbf{w}_j^k\|_2^2}}$;
  3. Calculate the block-diagonal matrices $\bar{D}_r^i (1 \le i \le c_r)$ and $\bar{D}_c^i (1 \le i \le c_c)$ where the k-th diagonal block of D_j^i is $\frac{1}{2\|(\mathbf{w}_j)_i^k\|_2} I_k$;
  4. Calculate the diagonal matrices $\hat{D}_r$ and $\hat{D}_c$ where $\hat{D}_j = \frac{1}{2}\left(\mathcal{W}_{j(0)}(\mathcal{W}_{j(0)})^T\right)$;
  5. Update each W_{tr} by $W_{tr} = (X_t X_t^T + \gamma_1 D_r + \gamma_2 \bar{D}_r^i + \gamma_3 \hat{D}_r)^{-1} X_t Y_{tr}$;
  6. For each $\mathbf{w}_i (1 \le i \le c_c)$ in each W_{tc}, calculate $(\mathbf{w}_{t+1})_i = \tilde{D}_c^{-\frac{1}{2}} (\tilde{\mathbf{w}}_t)_i$, where $\tilde{\mathbf{w}}_i = arg\min_{\tilde{\mathbf{w}}_i} f_i(\tilde{\mathbf{w}}_i, b_i; \tilde{X}) + \tilde{\mathbf{w}}_i^T \tilde{\mathbf{w}}_i$, $\tilde{X} = \tilde{D}_c^{-\frac{1}{2}} X$ and $\tilde{D}_c = \gamma_1 D_c + \gamma_2 \bar{D}_c^i + \gamma_3 \hat{D}_c$;
  7. Update each W_t by $W_t = [W_r, W_c]$;
  8. $t = t + 1$.
Result: $\mathcal{W} = \{W_1, W_2, ..., W_T\} \in \mathbb{R}^{D \times c \times T}$.

3 Experiments

In this section, we will evaluate the proposed method on the data set provided by the ADNI. The goal of our experiments is to determine the relationships between the brain imaging data (FreeSurfer and VBM), genotypes encoded by SNPs, and the corresponding cognitive scores and AD diagnoses.

We downloaded 1.5 T MRI scans, SNP genotypes, and demographic information for 821 ADNI-1 participants. We performed voxel-based morphometry (VBM) and FreeSurfer automated parcellation on the MRI data by following [6], and extracted mean modulated gray matter (GM) measures for 90 target regions of interest (ROIs). We followed the SNP quality control steps discussed in [8]. We also downloaded the longitudinal scores of the participants' Rey Auditory Verbal Learning Test (RAVLT) and their clinical diagnoses in three categories: healthy control (HC), mild cognitive impairment (MCI), and AD. The details of these cognitive assessments can be found in the ADNI procedure manuals. The time points examined in this study for both imaging markers and cognitive assessments included baseline (BL), Month 6 (M6), Month 12 (M12) and Month 24 (M24). All the participants with no missing BL/M6/M12/M24 MRI measurements, SNP genotypes, and cognitive measures were included in this study; this resulted in a set of 412 subjects with 155 HC, 110 MCI, and 147 AD.

3.1 Joint Regression and Classification Performance

In order to evaluate the effectiveness of our new *Joint High-Order Multi-Modal Multi-Task Feature Learning* method, we tested its regression and classification performance against an array of popular machine learning models. In each experiment, we fine tune the parameters of our model (γ_1, γ_2 and γ_3) by searching a grid of powers of 10 between 10^{-5} to 10^5. The experiments are performed using a classical 5-fold cross-validation strategy for each of the chosen algorithms.

Table 1. Regression: Root mean squared error (RMSE) results of the proposed algorithm compared to linear regression, ridge regression, Lasso regression, K-nearest neighbors (KNN), and a multi-layer perceptron (MLP) classifier. **Classification:** F_1 scores of classifying HC, MCI, and AD patients of the proposed algorithm compared to logistic regression, random forest, support vector machine (SVM) (with *RBF* kernel), K-nearest-neighbors (KNN), and a multi-layer perceptron (MLP) regressor.

Regression Performance (RAVLT)

	Linear	**Ridge**	**Lasso**
RMSE	1.41e+13±1.19e+12	0.333±0.016	0.333±0.016
	KNN	**MLP**	**Ours**
RMSE	0.344±0.009	0.318±0.026	**0.284±0.011**

Classification Performance (Diagnosis)

	Logistic	**RandomForest**	**SVM**
F_1 *(HC)*	0.472±0.054	0.434±0.048	0.310±0.073
F_1 *(MCI)*	0.420±0.065	0.448±0.045	0.460±0.071
F_1 *(AD)*	0.456±0.044	0.494±0.098	0.450±0.088
	KNN	**MLP**	**Ours**
F_1 *(HC)*	0.340±0.069	0.424±0.089	**0.560±0.034**
F_1 *(MCI)*	0.396±0.054	0.386±0.092	**0.508±0.039**
F_1 *(AD)*	0.354±0.093	0.444±0.039	**0.644±0.120**

Results. In Table 1 we can see that our proposed algorithm performs significantly better than a collection of "out-of-the-box" machine learning methods. The significant performance improvements in both regression and classification are due to the fact that our algorithm is the only one capable of incorporating the important longitudinal information into its prediction. The various regularizations ($\ell_{2,1}$-, group ℓ_1- and trace norms) that we apply to the unfolded matrix $\mathcal{W}$ ensure that our proposed algorithm is able to incorporate the longitudinal patterns that are intrinsic to many clinical studies (including the ADNI).

3.2 Identification of Longitudinal Imaging Biomarkers

FreeSurfer. The coefficients associated with the FreeSurfer modality in $\mathcal{X}$ are extracted from $\mathcal{W}$ at each time point (BL, M6, M12, M24). Each corresponding coefficient is mapped onto Automated Anatomical Labeling (AAL) [9] regions of

the brain (Fig. 1). When we look at the FreeSurfer brain heatmap we can draw a few interesting conclusions. First, the images show the same sparse image representation over time. This observation shows us that that the $\ell_{2,1}$-norm is working as expected and is successfully associating features across time, which illustrates the longitudinal predictive potential (a clinically important distinction) of our method. Second, we see that multiple parts of the brain related to the frontal gyrus have high weights compared to other parts of the brain not connected to AD, which is nicely consistent with existing clinical findings [3].

Voxel-Based Morphometry. The coefficients associated with the VBM modality in $\mathcal{X}$ are extracted from the coefficient matrix $\mathcal{W}$ at each time point. Each coefficient weight is mapped onto AAL regions of the brain (Fig. 2). The images associated with the VBM modality share the same longitudinal sparsity that we observed in the FreeSurfer coefficient matrix. Although, in this case, a completely different set of brain imaging features was discovered: features associated with the hippocampus. Here we see the remarkable effect of the G_1-norm regularization combined with the trace norm. Hippocampus atrophy has been shown to be highly predictive of AD.

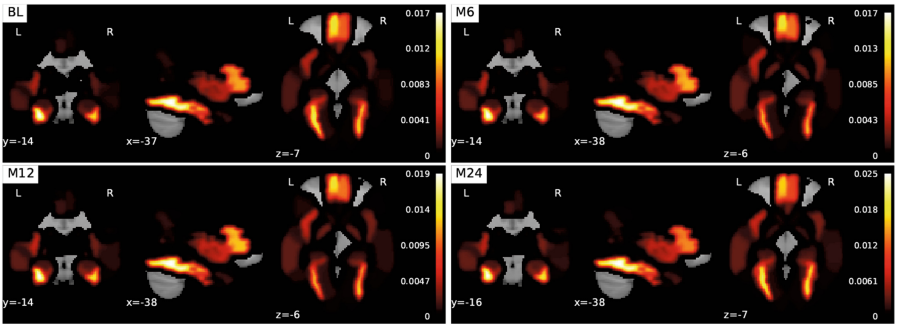

Fig. 1. Visualization of the FreeSurfer modality coefficients derived from $\mathcal{W}$ at various times (BL/M6/M12/M24). The top ten AAL regions are as follows (largest to smallest): *Fusiform_L, Fusiform_R, Frontal_Med_Orb_L, Frontal_Inf_Tri_L, Frontal_Med_Orb_R, Frontal_Inf_Tri_R, ParaHippocampal_L, Insula_L, Pallidum_R,* and *Pallidum_L.*

Single Nucleotide Polymorphism. The coefficients associated with the SNP modality in $\mathcal{X}$ are extracted from the coefficient matrix $\mathcal{W}$. Similar to two previous modalities, there was little difference between the coefficient matrices at each time point. The only orange bar in Fig. 3 is the coefficient that is associated with the *rs429358* SNP: the apolipoprotein E (ApoE) gene. *Schuff et al.* [7] and many others have discovered that the ApoE gene is related to increased rates of hippocampus atrophy. It is surprising that no other SNPs show up given that SNPs on the same gene are frequently associated with one another. One reason for this could be that the tuning coefficient on the $\ell_{2,1}$-norm is too large.

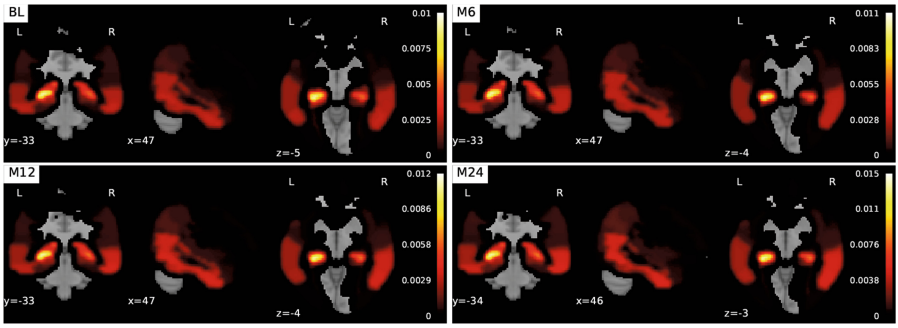

Fig. 2. Visualization of the voxel-based morphometry modality coefficients derived from $\mathcal{W}$ at various times (BL/M6/M12/M24). The top ten AAL regions are as follows (largest to smallest): *Hippocampus_L, Amygdala_L, Hippocampus_R Temporal_Inf_R Temporal_Mid_R, Temporal_Inf_L, ParaHippocampal_L, Amygdala_R, Temporal_Mid_L, ParaHippocampal_R, Angular_R,* and *Temporal_Pole_Sup_L.*

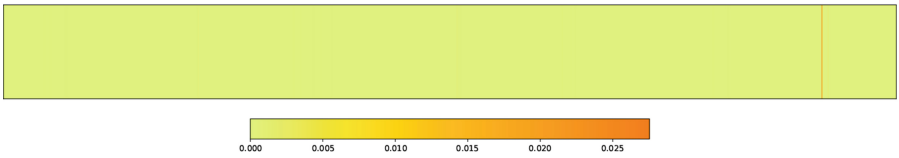

Fig. 3. Heatmap visualization of the SNPs along the x-axis against the corresponding coefficients derived from $\mathcal{W}$. The single orange line on the right-hand side is the weight associated with *rs429358*.

4 Conclusion

Learning effective mappings between different input and output modalities is an important research task in AD research. In the proposed *Joint High-Order Multi-Modal Multi-Task Feature Learning* model, we use various regularizations to learn the relationships between modalities over time. Our proposed method shows superior performance compared to traditional machine learning models.

Acknowledgement. This research was partially supported by NSF-IIS 1423591 and NSF-IIS 1652943; NSF-IIS 1302675, NSF-IIS 1344152, NSF-DBI 1356628, NSF-IIS 1619308, NSF-IIS 1633753, and NIH R01 AG049371; NIH R01 EB022574, NIH R01 LM011360, NIH R01 AG19771, NIH U19 AG024904, and NIH P30 AG10133.

References

1. Alzheimer, Association, Sciencestaff, Alzorg: 2017 Alzheimer's disease facts and figures (2017). https://doi.org/10.1016/j.jalz.2017.02.001
2. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717 (2009). https://doi.org/10.1007/s10208-009-9045-5

3. Galton, C.J., et al.: Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia. Neurology **57**(2), 216–225 (2001)
4. Lu, L., Wang, H., Yao, X., Risacher, S., Saykin, A., Shen, L.: Predicting progressions of cognitive outcomes via high-order multi-modal multi-task feature learning. In: IEEE ISBI 2018, pp. 545–548 (2018)
5. Nie, F., Huang, H., Cai, X., Ding, C.H.: Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In: NIPS 2010, pp. 1813–1821 (2010)
6. Risacher, S.L., et al.: Longitudinal MRI atrophy biomarkers: relationship to conversion in the ADNI cohort. Neurobiol. Aging **31**(8), 1401–1418 (2010)
7. Schuff, N., et al.: MRI of hippocampal volume loss in early Alzheimers disease in relation to ApoE genotype and biomarkers. Brain **132**(4), 1067–1077 (2009)
8. Shen, L.: Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort. NeuroImage **53**(3), 1051–1063 (2010). imaging Genetics
9. Tzourio-Mazoyer, N., et al.: Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage **15**(1), 273–289 (2002)
10. Wang, H., Nie, F., Huang, H., Ding, C.: Heterogeneous visual features fusion via sparse multimodal machine. In: IEEE CVPR 2013, pp. 3097–3102 (2013)
11. Wang, H., Nie, F., Huang, H.: Multi-view clustering and feature learning via structured sparsity. In: International Conference on Machine Learning (ICML 2013), pp. 352–360 (2013)
12. Wang, H.: Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. Bioinformatics **28**(2), 229–237 (2011)
13. Wang, H., et al.: Identifying AD-sensitive and cognition-relevant imaging biomarkers via joint classification and regression. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011. LNCS, vol. 6893, pp. 115–123. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23626-6_15
14. Wang, H., et al.: Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. Bioinformatics **28**(12), i127–i136 (2012)
15. Wang, H., et al.: From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant SNPs. Bioinformatics **28**(18), i619–i625 (2012)
16. Wang, H., et al.: High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction. In: NIPS 2012, pp. 1277–1285 (2012)