New Multi-task Learning Model to Predict Alzheimer's Disease Cognitive Assessment

Zhouyuan Huo[1], Dinggang Shen[2], and Heng Huang[1(✉)]

[1] Computer Science and Engineering, University of Texas at Arlington, Arlington, USA
heng@uta.edu

[2] Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, USA

Abstract. As a neurodegenerative disorder, the Alzheimer's disease (AD) status can be characterized by the progressive impairment of memory and other cognitive functions. Thus, it is an important topic to use neuroimaging measures to predict cognitive performance and track the progression of AD. Many existing cognitive performance prediction methods employ the regression models to associate cognitive scores to neuroimaging measures, but these methods do not take into account the interconnected structures within imaging data and those among cognitive scores. To address this problem, we propose a novel multi-task learning model for minimizing the k smallest singular values to uncover the underlying low-rank common subspace and jointly analyze all the imaging and clinical data. The effectiveness of our method is demonstrated by the clearly improved prediction performances in all empirical AD cognitive scores prediction cases.

1 Introduction

Accruing scientific evidences have demonstrated that the neuroimaging techniques, such as magnetic resonance imaging (MRI), are important for the detection of early Alzheimer's Disease (AD) [2,4,7,13]. Current American Academy of Neurology (AAN) guidelines [3] for dementia diagnosis recommend imaging to identify structural brain diseases that can cause cognitive impairment. Because AD is a neurodegenerative disorder characterized by progressive impairment of cognitive functions, it is important to diagnose the degree of brain impairment, and how much it can influence the performance of cognitive tests. As a result, many studies have focused on using regression models to predict cognitive scores and track AD progression [10,11]. In [10], the voxel-based morphometry (VBM) features extracted from the entire brain were jointly analyzed by the relevance

Z. Huo and H. Huang—were supported in part by NSF IIS-1117965, IIS-1302675, IIS-1344152, DBI-1356628, and NIH AG049371. D. Shen was supported in part by NIH AG041721.

S. Ourselin et al. (Eds.): MICCAI 2016, Part I, LNCS 9900, pp. 317–325, 2016.
DOI: 10.1007/978-3-319-46720-7_37

vector regression method to predict different clinical scores individually. However, different neuroimaging features or different cognitive scores are often interrelated. To tackle this problem, several recent studies, such as [11,12], tried to employ the multi-task learning models to uncover the inherent structures among neuroimaging features and cognitive scores. The low-rank regularization is an effective method to extract the common subspace for multiple tasks. Although trace norm is a widely used convex relaxation of low-rank regularization [1], its performance is easily influenced by the large singular values. For example, when the largest singular values of matrix M increase, the rank of M doesn't change but the trace norm of M increases correspondingly.

To address the above problems, in this paper, we propose a novel multi-task learning model to learn the associations between neuroimaging features and cognitive scores and uncover the low-rank common subspace among different tasks by minimizing the k smallest singular values. Our new k minimal singular values minimization regularization is a tighter relaxation than trace norm for rank minimization, such that our new multi-task learning model can have better prediction performance. We derive a new optimization algorithm to solve the proposed objective function and demonstrate the proof of its convergence. The proposed new model is applied to analyze the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort [16] data. In all empirical results, our new multi-task learning method consistently outperforms the widely used multivariate regression method, as well as different state-of-the-art multi-task learning approaches.

2 New Multi-task Learning Model

2.1 New Objective Function

In our new model, we focus on minimizing the k-smallest singular values of W and ignoring the largest singular values, such that our new regularization function is a better relaxation than trace norm. Thus, we propose to solve the following problem for multi-task learning:

$$J_{_{opt}} = \min_{W=[W_1,...,W_T]} \sum_{t=1}^{T} f(W_t^T X_t, Y_t) + \gamma \sum_{i=1}^{k} \sigma_i(W) \tag{1}$$

Suppose there are T learning tasks, the t-th task has n_t training data points $X_t = [x_1^t, x_2^t, ..., x_{n_t}^t] \in \mathbb{R}^{d \times n_t}$. For each data x_i^t, the label y_i^t is given with the label matrix $Y_t = [y_1^t, y_2^t, ..., y_{n_t}^t] \in \mathbb{R}^{c_t \times n_t}$ for each task t. $W_t \in \mathbb{R}^{d \times c_t}$ is the projection matrix to be learned, $W \in R^{d \times c}$ and $c = \sum_{t=1}^{T} c_t$.

It is interesting to see that when γ is large enough, then the k-smallest singular values of the optimal solution W to problem (1) will be zero as all the singular values of a matrix is non-negative. That is, when γ is large enough, it is equal to constrain the rank of W to be $r = m - k$ in the problem (1).

2.2 Optimization Algorithm

As per the definition of $||W||_*$ and singular value decomposition of W, it is known that:

$$\sum_{i=1}^{k} \sigma_i(W) = \|W\|_* - \max_{\substack{F \in R^{d\times r}, F^T F = I, \\ G \in R^{c\times r}, G^T G = I}} Tr(F^T W G)\,, \tag{2}$$

where $\|W\|_*$ is the sum of all the singular values of W, and the optimal solution of right term is sum of r largest singular values, F is the r left singular vectors of W and G is the r right singular vectors of W.

According to Eq. (2), the objective J_{opt} in Eq. (1) is equivalent to:

$$\min_{\substack{W=[W_1,\ldots,W_T], \\ F \in R^{d\times r}, F^T F = I, \\ G \in R^{T\times r}, G^T G = I}} \sum_{t=1}^{T} f(W_t^T X_t, Y_t) + \gamma \|W\|_* - \gamma Tr(F^T W G)\,. \tag{3}$$

When W is fixed, the problem (3) becomes:

$$\max_{\substack{F \in R^{d\times r}, F^T F = I, \\ G \in R^{c\times r}, G^T G = I}} Tr(F^T W G) \tag{4}$$

The optimal solution F to the problem (4) is formed by r left singular vectors of W corresponding to the r largest singular values, and the optimal solution G is formed by r right singular vectors of W corresponding to the r largest singular values.

When F and G are fixed, we define:

$$g(W_t) = f(W_t^T X_t, Y_t) - \gamma Tr(W_t^T F G_t^T), \tag{5}$$

the problem (3) becomes:

$$\min_{W=[W_1,\ldots,W_T]} \sum_{t=1}^{T} g(W_t) + \gamma \|W\|_*. \tag{6}$$

Using the reweighted method [6], we can solve problem (6) by iteratively solving the following problem:

$$\min_{W=[W_1,\ldots,W_T]} \sum_{t=1}^{T} g(W_t) + \gamma \sum_{t=1}^{T} Tr(W_t W_t^T D), \tag{7}$$

where D is computed according to the solution W^* in the last iteration and is defined as:

$$D = \frac{1}{2}(W^* W^{*T})^{-\frac{1}{2}}. \tag{8}$$

We can see that each subproblem of task t is independent of each other in problem (7). Thus, if we use the least square loss function, for each task W_t, the objective function could be written as:

$$\min_{W_t} \left\| W_t^T X_t + b_t \mathbf{1}_t^T - Y_t \right\|_F^2 - \gamma Tr(W_t^T F G_t^T) + \gamma Tr(W_t W_t^T D). \tag{9}$$

We take derivatives of Eq. (9) with respect to b_t and W_t, and set them to zero. The optimal solution to problem (9) is as follows:

$$W_t = (X_t H X_t^T + \gamma D)^{-1}(X_t H Y_t^T + \frac{1}{2}\gamma F G_t^T) \qquad H = I - \frac{1}{n_t}\mathbf{1}_t \mathbf{1}_t^T, \tag{10}$$

$$b_t = \frac{1}{n_t} Y_t \mathbf{1}_t - \frac{1}{n_t} W_t^T X_t \mathbf{1}_t. \tag{11}$$

We summarize the detailed algorithm to solve the objective J_{opt} in Algorithm 1.

Algorithm 1. Algorithm to solve the objective J_{opt} in Eq. (1)

Input: The training data matrix $X_t = [x_1^t, x_2^t, ..., x_{n_t}^t] \in \mathbb{R}^{d \times n_t}$ and the label matrix $Y_t = [y_1^t, y_2^t, ..., y_{n_t}^t] \in \mathbb{R}^{c_t \times n_t}$ for each task t.
Output: $W \in \mathbb{R}^{d \times c}$.
Initialize $W \in \mathbb{R}^{d \times c}$.
repeat
1. Update F and G by the optimal solution to the problem (4).
2. Compute $D = \frac{1}{2}(WW^T)^{-\frac{1}{2}}$.
3. For each t, update W_t by the optimal solution to the problem (7).

until Converges

2.3 Algorithm Analysis

The Algorithm 1 will monotonically decrease the objective of the problem in Eq. (1) in each iteration. To prove it, we need the following lemma:

Lemma 1. *For any positive definite matrices $A, A_t \in R^{m \times m}$, the following inequality holds when $0 < p \leq 2$:*

$$Tr(A^{\frac{p}{2}}) - \frac{p}{2} Tr(A A_t^{\frac{p-2}{2}}) \leq Tr(A_t^{\frac{p}{2}}) - \frac{p}{2} Tr(A_t A_t^{\frac{p-2}{2}}). \tag{12}$$

It is proved in [6] that Lemma 1 holds. Based on the Lemma, we have the following theorem:

Theorem 1. *The Algorithm 1 will monotonically decrease the objective of the problem in Eq. (3) in each iteration till convergence.*

Proof. In each iteration, at first, we fix W and compute $\tilde{F}$ and $\tilde{G}$. According to the solution of Eq. (4), we know:

$$-\gamma Tr(\tilde{F}^T W \tilde{G}) \leq -\gamma Tr(F^T W G). \tag{13}$$

When $\tilde{F}$ and $\tilde{G}$ are fixed, the problem becomes Eq. (7), by assuming that $\tilde{W}$ is the solution in each iteration, we have:

$$\sum_{t=1}^{T} g(\tilde{W}_t) + \frac{\gamma}{2} Tr(\tilde{W}\tilde{W}^T (WW^T)^{-\frac{1}{2}}) \leq \sum_{t=1}^{T} g(W_t) + \frac{\gamma}{2} Tr(WW^T (WW^T)^{-\frac{1}{2}}). \tag{14}$$

On the other hand, according to Lemma 1, when $p = 1$, we have:

$$Tr((\tilde{W}\tilde{W}^T)^{\frac{1}{2}}) - \frac{1}{2} Tr(\tilde{W}\tilde{W}^T (WW^T)^{-\frac{1}{2}}) \leq Tr((WW^T)^{\frac{1}{2}}) - \frac{1}{2} Tr((WW^T)(WW^T)^{-\frac{1}{2}}). \tag{15}$$

Combining (13), (14), and (15), we arrive at:

$$\sum_{t=1}^{T} f(\tilde{W}_t^T X_t, Y_t) + \gamma ||\tilde{W}||_* - \gamma Tr(\tilde{F}^T W \tilde{G}) \leq \sum_{t=1}^{T} f(W_t^T X_t, Y_t) + \gamma \|W\|_* - \gamma Tr(F^T W G). \tag{16}$$

Thus the Algorithm 1 will not increase the objective function in (3) at each iteration. Note that the equalities in above questions hold only when the algorithm converges. Therefore, the Algorithm 1 monotonically decreases the objective value in each iteration till the convergence.

Because we alternatively solve F, G, and W, the Algorithm 1 will converge to the local optimum of the problem (3), which is equivalent to the proposed objective function.

3 Experimental Results and Discussions

3.1 Data Set Description

Data used in this paper were obtained from the ADNI database (adni.loni.usc.edu). One goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, we refer interested readers to visit www.adni-info.org.

The data processing steps are as follows. Each MRI T1-weighted image was first anterior commissure (AC)'s posterior commissure (PC) corrected using MIPAV2, intensity inhomogeneity corrected using the N3 algorithm [9], skull stripped [15] with manual editing, and cerebellum-removed [14]. We then used FAST [17] in the FSL package3 to segment the image into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), and further used HAMMER [8] to register the images to a common space. GM volumes obtained from 93 ROIs defined in [5], normalized by the total intracranial volume, were extracted as features. Nine cognitive scores from five independent cognitive assessments

were downloaded, including three scores from RAVLT cognitive assessment; two scores from Fluency cognitive assessment (FLU); two scores from Trail making test (TRAIL). A total of 525 subjects are involved in our study, including 78 AD, 260 MCI, and 187 HC participants.

3.2 Improved Cognitive Status Prediction for Individual Assessment Tests

First, we apply the proposed method to the ADNI cohort, and separately predict each of the following three sets of cognitive scores: RAVLT, TRAILS and FLUENCY. The morphometric variables $\{x_i\}_{i=1}^n \in \mathbb{R}^d$, and $d = 93$ in this experiment.

We compare the proposed multi-task learning method to three most related methods: multivariate regression (MRV), multi-task learning model with $\ell_{2,1}$-norm regularization ($\ell_{2,1}$) [11], and multi-task learning model with trace norm (LS_TRACE) [1], in cognitive performance prediction. For each test case, we use 5-fold cross validation and the prediction performance is assessed by the root mean square error (RMSE). All experimental results are reported in Table 1. The proposed method consistently outperforms other methods in nearly all the test cases for all the cognitive tasks.

The heat maps of parameter weights are shown in Fig. 1. Visualizing the parameter weights can help us locate the features which play important roles in the corresponding cognitive prediction tasks. In this way, there is much potential to identify the relevant imaging predictors and explain the effects of morphometric changes in relation to cognitive performance. As we can see, different coefficient values are represented in different colors in heat map. The blue polar

Table 1. Prediction performance measured by RMSE (mean ± std)

Test cases	Algorithm	Score1	Score2	Score3
FLUENCY	MVR	6.2292 ± 0.4191	4.1210 ± 0.4733	-
	LS_TRACE	5.9792 ± 0.6339	4.0492 ± 0.4294	-
	$\ell_{2,1}$	5.7431 ± 0.2796	3.9567 ± 0.2143	-
	Our method	**5.4377 ± 0.3125**	**3.9498 ± 0.3505**	-
RAVLT	MVR	10.8194 ± 0.9530	4.0606 ± 0.3071	4.0616 ± 0.3928
	LS_TRACE	10.6359 ± 1.1303	4.0252 ± 0.2896	4.0399 ± 0.2250
	$\ell_{2,1}$	10.4451 ± 0.8905	3.9618 ± 0.2484	3.7906 ± 0.1444
	Our method	**9.7834 ± 0.4867**	**3.7261 ± 0.1368**	**3.6984 ± 0.1603**
TRAILS	MVR	22.3629 ± 1.0656	78.1796 ± 7.3501	70.9399 ± 7.2238
	LS_TRACE	20.7686 ± 1.1213	75.0121 ± 6.4147	65.3007 ± 6.0726
	$\ell_{2,1}$	19.5400 ± 2.8240	72.7200 ± 8.6480	63.4796 ± 7.3528
	Our method	**18.1809 ± 2.0390**	**66.9982 ± 5.1144**	**58.0915 ± 4.0492**

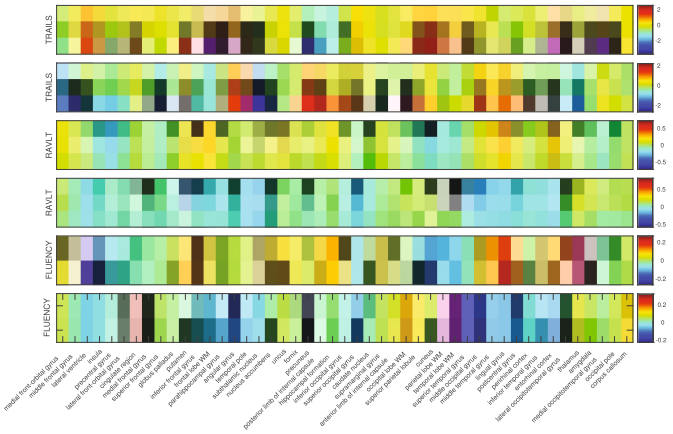

Fig. 1. Heat map of corresponding features for cognitive score prediction.

and red polar mean a significant effect of corresponding features on cognitive score performance.

3.3 Improved Cognitive Performance Prediction for Joint Assessment Tests

To further evaluate the multi-task joint analysis power, we apply the proposed method to predict all five types of cognitive scores (RAVLT, TRAILS, FLUENCY) jointly. Such experiments will demonstrate how the interrelations among cognitive assessment tests are utilized to enhance the prediction performance.

Table 2. Prediction performance measured by RMSE (mean ± std) for joint assessment tests.

Algorithm	Score name	Score1	Score2	Score3
MVR	FLUENCY	6.0282 ± 0.2255	4.1852 ± 0.4346	-
	RAVLT	11.0376 ± 0.4489	4.0608 ± 0.2554	4.0561 ± 0.1547
	TRAILS	21.7435 ± 1.3936	77.0161 ± 5.2578	68.1576 ± 4.837
LS_TRACE	FLUENCY	5.7778 ± 0.1130	3.9681 ± 0.2965	-
	RAVLT	10.8519 ± 0.8808	3.8674 ± 0.4112	3.8772 ± 0.1943
	TRAILS	20.5224 ± 1.1906	74.4795 ± 4.5967	64.3386 ± 4.2974
$\ell_{2,1}$	FLUENCY	5.8100 ± 0.9274	3.9139 ± 0.3538	-
	RAVLT	10.4500 ± 0.3846	3.9806 ± 0.2158	3.8797 ± 0.2050
	TRAILS	19.7753 ± 1.5802	70.9585 ± 5.5396	62.3717 ± 4.9592
Our method	FLUENCY	**5.4644 ± 0.3515**	**3.8724 ± 0.1908**	-
	RAVLT	**10.4492 ± 0.8235**	**3.6522 ± 0.2542**	**3.7086 ± 0.1814**
	TRAILS	**17.8778 ± 1.8126**	**66.3821 ± 5.6292**	**57.7588 ± 5.3360**

Similar to the previous experiment, we also compare our method to three other related models. For each test case, we use 5-fold cross validation to evaluate the average performance of each algorithm. The prediction results are evaluated by RMSE and reported in Table 2. In all prediction cases, our method outperforms other methods.

4 Conclusion

In this paper, we proposed a new multi-task learning model for minimizing k smallest singular values to predict the cognitive scores for complex brain disorders. This proposed new low-rank regularization is a better approximation of rank minimization regularization problem than the standard trace norm regularization, thus our new multi-task learning method can uncover the shared common subspace efficiently and sufficiently. As a result, cognitive score prediction results are enhanced by the learned hidden structures among tasks and features. We also introduced an efficient optimization algorithm to solve our proposed objective function with rigorous theoretical analysis. Our experiments were conducted on the MRI and multiple cognitive scores data of the ADNI cohort and yield promising results: (1) Prediction performance of the proposed multi-task learning model is better than all related methods in all cases; (2) Our method can predict multiple cognitive scores at the same time and has a potential to play an important role in determining cognitive functions and characterizing AD progression.

References

1. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Mach. Learn. **73**(3), 243–272 (2008)
2. Batmanghelich, N., Taskar, B., Davatzikos, C.: A general and unifying framework for feature construction, in image-based pattern classification. In: Prince, J.L., Pham, D.L., Myers, K.J. (eds.) IPMI 2009. LNCS, vol. 5636, pp. 423–434. Springer, Heidelberg (2009)
3. De Leon, M., George, A., Stylopoulos, L., Smith, G., Miller, D.: Early marker for Alzheimer's disease: the atrophic hippocampus. Lancet **334**(8664), 672–673 (1989)
4. Hassabis, D., Maguire, E.A.: Deconstructing episodic memory with construction. Trends Cogn. Sci. **11**(7), 299–306 (2007)
5. Kabani, N.J.: 3D anatomical atlas of the human brain. Neuroimage **7**, P-0717 (1998)
6. Nie, F., Huang, H., Ding, C.H.: Low-rank matrix recovery via efficient schatten p-Norm minimization. In: AAAI (2012)
7. Rosen, H.J., Gorno-Tempini, M.L., Goldman, W., Perry, R., Schuff, N., Weiner, M., Feiwell, R., Kramer, J., Miller, B.L.: Patterns of brain atrophy in frontotemporal dementia and semantic dementia. Neurology **58**(2), 198–208 (2002)
8. Shen, D., Davatzikos, C.: Hammer: hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imaging **21**(11), 1421–1439 (2002)

9. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging **17**(1), 87–97 (1998)
10. Stonnington, C.M., Chu, C., Klöppel, S., Jack Jr., C.R., Ashburner, J., Frackowiak, R.S.: Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage **51**(4), 1405–1413 (2010)
11. Wang, H., Nie, F., Huang, H., Risacher, S., Ding, C., Saykin, A.J., Shen, L.: Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 557–562. IEEE (2011)
12. Wang, H., Nie, F., Huang, H., Risacher, S., Saykin, A.J., Shen, L., ADNI: joint classification and regression for identifying ad-sensitive and cognition-relevant imaging biomarkers. In: 14th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 115–123 (2011)
13. Wang, H., Nie, F., Huang, H., Risacher, S.L., Saykin, A.J., Shen, L.: ADNI: identifying disease sensitive and quantitative trait relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multi-modal multitask learning. Bioinformatics **28**(12), i127–i136 (2012)
14. Wang, Y., Nie, J., Yap, P.T., Li, G., Shi, F., Geng, X., Guo, L., Shen, D., Initiative, A.D.N., et al.: Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. PloS One **9**(1), e77810 (2014)
15. Wang, Y., Nie, J., Yap, P.-T., Shi, F., Guo, L., Shen, D.: Robust deformable-surface-based skull-stripping for large-scale studies. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011. LNCS, vol. 6893, pp. 635–642. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23626-6_78
16. Weiner, M.W., Aisen, P.S., Jack Jr., C.R., Jagust, W.J., Trojanowski, J.Q., Shaw, L., Saykin, A.J., Morris, J.C., Cairns, N., Beckett, L.A., et al.: The Alzheimer's disease neuroimaging initiative: progress report and future plans. Alzheimer's Dement. **6**(3), 202–211 (2010)
17. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging **20**(1), 45–57 (2001)