Sparse multi-kernel based multi-task learning for joint prediction of clinical scores and biomarker identification in Alzheimer's Disease

Peng Cao[1], Xiaoli Liu[1], Jinzhu Yang[1],
Dazhe Zhao[1], and Osmar Zaiane[2]

[1] Key Laboratory of Medical Image Computing of Ministry of Education, College of
Computer Science and Engineering, Northeastern University, China
[2] University of Alberta, Canada

Abstract. Machine learning methods have been used to predict the clinical scores and identify the image biomarkers from individual MRI scans. Recently, the multi-task learning (MTL) with sparsity-inducing norm have been widely studied to investigate the prediction power of neuroimaging measures by incorporating inherent correlations among multiple clinical cognitive measures. However, most of the existing MTL algorithms are formulated linear sparse models, in which the response (e.g., cognitive score) is a linear function of predictors (e.g., neuroimaging measures). To exploit the nonlinear relationship between the neuroimaging measures and cognitive measures, we consider that tasks to be learned share a common subset of features in the kernel space as well as the kernel functions. Specifically, we propose a multi-kernel based multi-task learning with a mixed sparsity-inducing norm to better capture the complex relationship between the cognitive scores and the neuroimaging measures. The formation can be efficiently solved by mirror-descent optimization. Experiments on the Alzheimers Disease Neuroimaging Initiative (ADNI) database showed that the proposed algorithm achieved better prediction performance than state-of-the-art linear based methods both on single MRI and multiple modalities.

1 Introduction

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. Multivariate regression models have been studied in AD for revealing relationships between neuroimaging measures and cognitive scores to understand how structural changes in brain can influence cognitive status, and predict the cognitive performance with neuroimaging measures based on the estimated relationships. Many clinical/cognitive measures have been designed to evaluate the cognitive status of the patients and used as important criteria for clinical diagnosis of probable AD[13, 5, 15]. There exists a correlation among multiple cognitive tests, and many multi-task learning (MTL) seeks to improve the performance of a task by exploiting the intrinsic relationships among the related cognitive tasks.

The assumption of the commonly used MTL methods is that all tasks share the same data representation with $\ell_{2,1}$-norm regularization, since a given imaging marker can affect multiple cognitive scores and only a subset of the imaging features (brain region) are relevant[13, 15]. However, they assumed linear relationship between the MRI features and the cognitive outcomes and the $\ell_{2,1}$-norm regularization only consider the shared representation from the features in the original space. Unfortunately, this assumption usually does not hold due to the inherently complex structure in the dataset[11]. Kernel methods have the ability to capture the nonlinear relationships by mapping data to higher dimensions where it exhibits linear patterns. However, the choice of the types and parameters of the kernels for a particular task is critical, which determines the mapping between the input space and the feature space. To address the above issues, we propose a sparse multi-kernel based multi-task Learning (SMKMTL) with a mixed sparsity-inducing norm to better capture the complex relationship between the cognitive scores and the neuroimaging measures. The multiple kernel learning (MKL) [4] not only learns a optimal combination of given base kernels, but also exploits the nonlinear relationship between MRI measures and cognitive performance. The assumption of SMKMTL is that the not only the kernel functions but also the features in the high dimensional space induced by the combination of only few kernels are shared for the multiple cognitive measure tasks. Specifically, SMKMTL explicitly incorporates the task correlation structure with $\ell_{2,1}$-norm regularization on the high dimensional features in the RKHS space, which builds the relationship between the MRI features and cognitive score prediction tasks in a nonlinear manner, and ensures that a small subset of features will be selected for the regression models of all the cognitive outcomes prediction tasks; and an ℓ_q-norm on the kernel functions, which ensures various schemes of sparsely combining the base kernels by varying q. Moreover, MKL framework has advantage of fusing multiple modalities, we apply our SMKMTL on multi-modality data (MRI, PET and demographic information) in our study.

We presented mirror descent-type algorithm to efficiently solve the proposed optimization problems, and conducted extensive experiments using data from the Alzheimers Disease Neuroimaging Initiative (ADNI) to demonstrate our methods with respect to the prediction performance and multi-modality fusion.

2　Sparse Multi kernel Multi-task learning, SMKMTL

Consider a multi-task learning (MTL) setting with m tasks. Let p be the number of covariates, shared across all the tasks, n be the number of samples. Let $X \in \mathbb{R}^{n \times p}$ denote the matrix of covariates, $Y \in \mathbb{R}^{n \times m}$ be the matrix of responses with each row corresponding to a sample, and $\Theta \in \mathbb{R}^{p \times m}$ denote the parameter matrix, with column $\theta_{.t} \in \mathbb{R}^p$ corresponding to task t, $t = 1, \ldots, m$, and row $\theta_{i.} \in \mathbb{R}^m$ corresponding to feature i, $i = 1, \ldots, p$.

$$\min_{\Theta \in \mathbb{R}^{p \times t}} \quad L(Y, X, \Theta) + \lambda R(\Theta) , \tag{1}$$

where $L(\cdot)$ denotes the loss function and $R(\cdot)$ is the regularizer.

The commonly used MTL is MT-GL model with $\ell_{2,1}$-norm regularization, which considers $R(\Theta) = \|\Theta\|_{2,1} = \sum_{l=1}^{p} \|\theta_{l\cdot}\|_2$ and is suitable for simultaneously enforcing sparsity over features for all tasks. Moreover, Argyriou proposed a Multi-Task Feature Learning (MTFL) with $\ell_{2,1}$-norm[1], the formulation of which is: $\|Y - \mathbf{U}^T X\Theta\|_F^2 + \|\Theta\|_{2,1}$, where $\mathbf{U}$ is an orthogonal matrix which is to be learnt.

In these learning methods, each task is traditionally performed by formulating a linear regression problem, in which the cognitive score is a linear function of the neuroimaging measures. However, the assumption of these existing linear models usually does not hold due to the inherently complex patterns between brain images and the corresponding cognitive outcomes. Modeling cognitive scores as nonlinear functions of neuroimaging measures may provide enhanced flexibility and the potential to better capture the complex relationship between the two quantities. In this paper, we consider the case that the features are associated to a kernel and hence they are in general nonlinear functions of the features. With the advantage of MKL, we assume that $\boldsymbol{x}_i$ can be mapped to k different Hilbert spaces, $\boldsymbol{x}_i \to \phi_j(\boldsymbol{x}_i), j = 1, \ldots, k$, implicitly with k nonlinear mapping functions, and the objective of MKL is to seek the optimal kernel combination.

In order to capture the intrinsic relationships among multiple related tasks in the RKHS space, we proposed a multi-kernel based multi-task learning with mixed sparsity-inducing norm. With the ε-insensitive loss function, the formulation can be expressed as:

$$
\min_{\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{\xi},\mathbf{U}} \quad \frac{1}{2} \left(\sum_{j=1}^{k} \left(\sum_{l=1}^{\hat{p}_j} \|\theta_{\cdot jl}\|_2 \right)^q \right)^{\frac{2}{q}} + C \sum_{t=1}^{m} \sum_{i=1}^{n_t} (\xi_{ti} + \xi_{ti}^*)
$$

$$
\text{s.t.} \quad
\begin{cases}
y_{ti} - \sum_{j=1}^{k} \boldsymbol{\theta}_{tj}^{\mathrm{T}} \mathbf{U}_j^{\mathrm{T}} \phi_j(x_{ti}) - b_t \leq \varepsilon + \xi_{ti} \\[2mm]
\sum_{j=1}^{k} \boldsymbol{\theta}_{tj}^{\mathrm{T}} \mathbf{U}_j^{\mathrm{T}} \phi_j(x_{ti}) + b_t - y_{ti} \leq \varepsilon + \xi_{ti}^*, \qquad \forall t, i \\[2mm]
\xi_{ti}, \xi_{ti}^* \geq 0, \mathbf{U}_j \in O^{\hat{p}_j}
\end{cases}
\tag{2}
$$

where $\boldsymbol{\theta}_j$ is the weight matrix for the j-th kernel, $\theta_{tjl}(l = 1, \ldots, \hat{p}_j)$ is the entries of $\boldsymbol{\theta}_{tj}$, n_t is the number of samples in the t-th task, $\hat{p}_j$ is the dimensionality of the feature space induced by the j-th kernel, ε is the parameter in the ε-insensitive loss, ξ_{ti} and ξ_{ti}^* are slack variables, and C is the regularization parameter.

In the formulation of Eq.(2), the use of $\ell_{2,1}$-norm for $\boldsymbol{\theta}_j$, which forces the weights corresponding to the i-th feature across multiple tasks to be grouped together and tends to select features based on the k tasks jointly in the kernel space. Moreover, an ℓ_q norm ($q \in [1, 2]$) over kernels is used over kernels instead of ℓ_1 norm to obtain various schemes of sparsely combining the base kernels by varying q.

Lemma 1 *Let $a_i \geq 0, i = 1 \ldots d$ and $1 < r < \infty$. Then,*

$$\min_{\eta \in \Delta_{d,r}} \sum_i \frac{a_i}{\eta_i} = \left(\sum_{i=1}^{d} a_i^{\frac{r}{r+1}} \right)^{\frac{r+1}{r}} \tag{3}$$

where $\Delta_{d,r} = \left\{ \mathbf{z} \equiv [z_1 \ldots z_d]^{\mathrm{T}} \mid \sum_{i=1}^{d} z_i^r \leq 1, z_i \geq 0, i = 1 \ldots d \right\}$. According to the **Lemma 1** introduced in [8], we introduces new variables $\lambda = [\lambda_1 \ldots \lambda_k]^T$, and $\gamma_j = [\gamma_{j1} \ldots \gamma_{j\hat{p}_j}]^{\mathrm{T}}, j = 1, \ldots, k$. Thus, the regularizer in (2) can be written as: $\min_{\lambda \in \Delta_{k,\bar{q}}} \min_{\gamma_j \in \Delta_{\hat{p}_j,1}} \sum_{t=1}^{m} \sum_{j=1}^{k} \sum_{l=1}^{\hat{p}_j} \frac{\theta_{tjl}^2}{\gamma_{jk}\lambda_j}$, where $\bar{q} = \frac{q}{2-q}$. Now we perform a change of variables: $\frac{\theta_{tjl}}{\sqrt{\gamma_{jk}\lambda_j}} = \bar{\theta}_{tjl}, l = 1, \ldots, \hat{p}_j$, and construct the Lagrangian for our optimization problem in (2) as:

$$\min_{\lambda, \gamma_j, \mathbf{U}_j} \sum_{t=1}^{m} \max_{\boldsymbol{\alpha}_t} \mathbf{y}^{\mathrm{T}}(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_t^*) - \frac{1}{2}(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_t^*)^{\mathrm{T}} \left(\sum_{j=1}^{k} \Phi_{tj}^{\mathrm{T}} \mathbf{U}_j^{\mathrm{T}} \Lambda_j \mathbf{U}_j \Phi_{tj} \right) (\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_t^*)$$

$$\text{s.t.} \qquad \lambda \in \Delta_{k,\bar{q}}, \gamma_j \in \Delta_{\hat{p}_j,1} \mathbf{U}_j \in O^{\hat{p}_j}, \boldsymbol{\alpha}_t \in S_{n_t}(C) \tag{4}$$

where Λ_j is a diagonal matrix with entries as $\lambda_j \gamma_{jl}, l = 1, \ldots, \hat{p}_j, \lambda \in \Delta_{k,\bar{q}}, \gamma_j \in \Delta_{\hat{p}_j,1}, \Phi_{tj}$ is the data matrix with columns as $\phi_j(x_{ti}), i = 1, \ldots, n_t, \boldsymbol{\alpha}_t, \boldsymbol{\alpha}_t^*$ are vectors of Lagrange multipliers corresponding to the t-th task in the SMKMTL formulation, $S_{n_t}(C) \equiv \{\boldsymbol{\alpha}_t \mid 0 \leq \alpha_{ti}, \alpha_{ti}^* \leq C, \ i = 1, \ldots, n_t, \sum_{i=1}^{n_t}(\alpha_{ti} - \alpha_{ti}^*) = 0\}$. Denoting $\mathbf{U}_j^{\mathrm{T}} \Lambda_j \mathbf{U}_j$ by $\bar{\mathbf{Q}}_j$ and eliminating variables $\lambda, \gamma, \mathbf{U}$ leads to:

$$\min_{\bar{\mathbf{Q}}} \sum_{t=1}^{m} \max_{\boldsymbol{\alpha}_t \in S_{n_t}(C)} \mathbf{y}^{\mathrm{T}}(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_t^*) - \frac{1}{2}(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_t^*)^{\mathrm{T}} \left(\sum_{j=1}^{k} \Phi_{tj}^{\mathrm{T}} \bar{\mathbf{Q}}_j \Phi_{tj} \right) (\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_t^*)$$

$$\text{s.t.} \qquad \bar{\mathbf{Q}}_j \succeq 0, \sum_{j=1}^{k} (\mathrm{Tr}(\bar{\mathbf{Q}}_j))^{\bar{q}} \leq 1 \tag{5}$$

Then, we use the method described in [6] to kernelize the formulation. Let $\Phi_j \equiv [\Phi_{1j} \ldots \Phi_{Tj}]$ and the compact SVD of Φ_j be $\mathbf{U}_j \Sigma_j \mathbf{V}_j^{\mathrm{T}}$. Now, introduce new variables $\mathbf{Q}_j$ such that $\bar{\mathbf{Q}}_j = \mathbf{U}_j \mathbf{Q}_j \mathbf{U}_j^{\mathrm{T}}$. Here, $\mathbf{Q}_j$ is a symmetric positive semidefinite matrix of size same as rank of Φ_j. Eliminating variables $\bar{\mathbf{Q}}_j$, we can re-write the above problem using $\mathbf{Q}_j$ as:

$$\min_{\mathbf{Q}} \sum_{t=1}^{m} \max_{\boldsymbol{\alpha}_t \in S_{n_t}(C)} \mathbf{y}^{\mathrm{T}}(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_t^*) - \frac{1}{2}(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_t^*)^{\mathrm{T}} \left(\sum_{j=1}^{k} \mathbf{M}_{tj}^{\mathrm{T}} \mathbf{Q}_j \mathbf{M}_{tj} \right) (\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_t^*)$$

$$\text{s.t.} \qquad \mathbf{Q}_j \succeq 0, \sum_{j=1}^{k} (\mathrm{Tr}(\mathbf{Q}_j))^{\bar{q}} \leq 1 \tag{6}$$

where $\mathbf{M}_{tj} = \Sigma_j^{-1} \mathbf{V}_j^{\mathrm{T}} \Phi_j^{\mathrm{T}} \Phi_{tj}$. Given $\mathbf{Q}_j$s, the problem is equivalent to solving m SVM problems individually. The $\mathbf{Q}_j$s are learnt using training examples of all the tasks and are shared across the tasks, and this formulation with trace norm as constraint can be solved by a mirror-descent based algorithm proposed in [2, 6].

3 Experimental Results

3.1 Data and Experimental Setting

In this work, only ADNI-1 subjects with no missing features or cognitive scores are included. This yields a total of $n = 816$ subjects, who are categorized into 3 baseline diagnostic groups: Cognitively Normal (CN, $n_1 = 228$), Mild Cognitive Impairment (MCI, $n_2 = 399$), and Alzheimer's Disease (AD, $n_3 = 189$). The dataset has been processed by a team from UCSF (University of California at San Francisco), who performed cortical reconstruction and volumetric segmentations with the FreeSurfer image analysis suite. There were $p = 319$ MRI features in total, including the cortical thickness average (TA), standard deviation of thickness (TS), surface area (SA), cortical volume (CV) and subcortical volume (SV) for a variety of ROIs. In order to sufficiently investigate the comparison, we further evaluate the performance on all the widely used cognitive assessments (e.g. ADAS, MMSE, RAVLT, FLU and TRAILS, totally $m = 10$ tasks)[14, 12, 11]. We use 10-fold cross valuation to evaluate our model and conduct the comparison. In each of twenty trials, a 5-fold nested cross validation procedure for all the comparable methods in our experiments is employed to tune the regularization parameters. Data was z-scored before applying regression methods. The candidate kernels are: six different kernel bandwidths ($2^{-2}, 2^{-1}, \ldots, 2^3$), polynomial kernels of degree 1 to 3, and a linear kernel, which totally yields 10 kernels. The kernel matrices were pre-computed and normalized to have unit trace. To have a fair comparison, we validate the regularization parameters of all the methods in the same search space C (from 10^{-1} to 10^3) and q (1,1.2,1.4,..,2) in our method on a subset of the training set, and use the optimal parameters to train the final models. Moreover, a warm-start technique is used for successive SVM retrainings.

In this section, we conduct empirical evaluation for the proposed methods by comparing with three single task learning methods: Lasso, ridge and simpleMKL, all of which are applied independently on each task. Moreover, we compare our method with two baseline multi-task learning methods: MTL with $\ell_{2,1}$-norm (MT-GL) and MTFL. We also compare our proposed method with several popular state-of-the-art related methods: Clustered Multi-Task Learning (CMTL)[16]: CMTL($\min_{\Theta: F^T F = I_k} L(X, Y, \Theta) + \lambda_1(\text{tr}(\Theta^T \Theta) - \text{tr}(F^T \Theta^T \Theta F)) + \lambda_2 \text{tr}(\Theta^T \Theta)$, where $F \in \mathbb{R}^{m \times k}$ is an orthogonal cluster indicator matrix) incorporates a regularization term to induce clustering between tasks and then share information only to tasks belonging to the same cluster. In the CMTL, the number of clusters is set to 5 since the 7 tasks belong to 5 sets of cognitive functions. Trace-Norm Regularized Multi-Task Learning (Trace)[7]: The assumption that all models share a common low-dimensional subspace ($\min_\Theta L(X, Y, \Theta) + \lambda \|\Theta\|_*$, where $\| \cdot \|_*$ denotes the trace norm defined as the sum of the singular values). Table 1 shows the results of the comparable MTL methods in term of root mean squared error (rMSE). Experimental results show that the proposed methods significantly outperform the most recent state-of-the-art algorithms proposed in terms of rMSE for most of the scores. Moreover, compared with the other

multi-task learning with different assumption, MT-GL, MTFL and our proposed methods belonging to the multi-task feature learning methods with the idea of sparsity, have a advantage over the other comparative multi-task learning methods. Since not all the brain regions are associated with AD, many of the features are irrelevant and redundant. Sparse based MTL methods are appropriate for the task of prediction cognitive measures and better than the non sparse based MTL methods. Furthermore, CMTL and Trace are worse than the Ridge, which demonstrates that the model assumption in them may be incorrect for modeling the correlation among the cognitive tasks.

Table 1: Performance comparison of various methods in terms of rMSE.

	Ridge	Lasso	MT-GL	MTFL	CMTL	Trace	simpleMKL	SMKMTL
ADAS	7.89±0.55	6.84±0.36	6.77±0.31	6.82±0.41	7.64±0.37	8.18 ±0.61	6.70±0.31	**6.61±0.45**
MMSE	2.76±0.14	2.21±0.07	2.21±0.09	2.26±0.09	3.08±0.46	6.11±2.04	2.21±0.08	**2.09±0.12**
RAVLT-TOTAL	11.6±0.52	10.0±0.54	**9.61±0.45**	9.44±0.53	11.5±0.51	13.1±3.12	9.65±0.47	9.63±0.51
RAVLT-TOT6	3.70±0.30	**3.32±0.20**	3.34±0.15	3.38±0.18	3.91±0.26	3.78±0.49	3.41±0.22	3.33±0.20
RAVLT-T30	3.79±0.27	**3.44±0.17**	**3.44±0.15**	3.46±4.04	3.24±0.25	3.91±0.43	3.41±0.23	3.46±0.19
RAVLT-RECOG	4.43±0.25	3.64±0.21	3.64±0.25	3.63 ±0.19	4.38±0.23	4.52±0.86	3.64± 0.25	**3.39±0.20**
FLU-ANIM	6.69±0.42	5.35±0.45	5.29±0.44	5.25± 0.49	6.61±0.56	6.74±1.42	5.30±0.44	**5.23±0.43**
FLU-VEG	4.47±0.21	3.75±0.10	3.70±0.10	3.71±0.11	4.39±0.29	4.67±0.79	4.82±0.22	**3.47±0.16**
TRAILS-A	26.7±1.80	23.8±1.40	23.4±1.11	23.4±1.12	27.5±1.98	28.8±3.28	24.1±1.81	**21.1±1.47**
TRAILS-B	81.3±2.52	71.2±2.81	71.3±2.95	70.9±2.52	83.6±5.44	89.7±7.83	72.8±2.74	**69.8±1.23**

3.2 Fusion of multi-modality

To estimate the effect of combining multi-modality image data with our SMKMTL methods and provide a more comprehensive comparison of the result from the proposed model, we further perform some experiments, that are (1) using only MRI modality, (2) using only PET modality, (3) combining two modalities: PET and MRI (MP), and (4) combining three modalities: PET, MRI and demographic information including age, years of education and ApoE genotyping (MPD). Different the above experiments, the samples from ADNI-2 are used instead of ADNI-1, since the amount of the patients with PET is sufficient. From the ADNI-2, we obtained all the patients with both MRI and PET, totally 756 samples. The PET imaging data are from the ADNI database processed by the UC Berkeley team, who use a native-space MRI scan for each subject that is segmented and parcellated with Freesurfer to generate a summary cortical and subcortical ROI, and coregister each florbetapir scan to the corresponding MRI and calculate the mean florbetapir uptake within the cortical and reference regions. The procedure of image processing is described in http://adni.loni.usc.edu/updated-florbetapir-av-45-pet-analysis-results/. In our SMKMTL, ten different kennel function described in the first experiment are used for each modality. To show the advantage of SMKMTL, we compare our SMKMTL with MTFL, which concatenated the multiple modalities features into a long vector features. The prediction performance results are shown in Table

2. From the results, it is clear that the method with multi-modality outperforms the methods using one single modality of data. This validates our assumption that the complementary information among different modalities is helpful for cognitive function prediction. Regardless of two or three modalities, the proposed SMKMTL achieved better performances than the linear based multi-task learning for the most cases, same as for the single modality learning task above.

Table 2: Performance comparison with multi-modality data in terms of rMSE.

	MTFL				SMKMTL			
	MRI	PET	MP	MPD	MRI	PET	MP	MPD
ADAS	6.28±0.33	6.09±0.27	6.05±0.29	5.83±0.33	6.19±0.53	5.95±0.17	5.87±0.22	**5.79±0.27**
MMSE	1.96±0.12	1.92±0.23	1.91±0.18	1.85±0.19	1.87±0.21	1.83±0.19	1.82±0.11	**1.77±0.15**
RAVLT-TOTAL	9.82±0.44	9.69±0.43	9.55±0.51	**9.51±0.41**	9.80±0.41	9.71±0.35	9.56±0.42	**9.51±0.33**
RAVLT-TOT6	3.24±0.15	3.19±0.16	3.09±0.11	2.95±0.11	3.11±0.12	3.03±0.21	2.97±0.09	**2.84±0.13**
RAVLT-T30	3.16±0.20	3.21±0.12	3.14±0.10	**3.05±0.15**	3.18±0.25	3.18±0.22	3.11±0.18	3.07±0.12
RAVLT-RECOG	3.70±0.25	3.52±0.12	3.44±0.28	3.30±0.18	3.55±0.21	3.54±0.15	3.34±0.11	**3.17±0.10**
FLU-ANIM	4.95±0.27	4.46±0.32	4.51±0.28	4.29±0.22	4.71±0.19	4.54±0.18	4.45±0.22	**4.21±0.16**
FLU-VEG	3.65±0.20	3.55±0.15	3.47±0.25	3.38±0.21	3.49±0.18	3.32±0.21	3.15±0.17	**3.09±0.10**
TRAILS-A	16.2±2.72	15.8±1.56	14.7±1.43	13.8±1.25	15.3±1.33	13.9±1.18	13.1±0.84	**12.5±1.08**
TRAILS-B	54.9±1.78	52.8±1.43	50.5±2.02	48.7±2.22	51.8±1.84	50.9±1.66	48.9±1.52	**46.0±1.47**

4 Conclusions

In this paper, we propose to multi-kernel based multi-task learning with $\ell_{2,1}$-norm on the correlation of tasks in the kernel space combined with ℓ_q-norm on the kernels in a joint framework. Extensive experiments illustrate that the proposed method not only yields superior performance on cognitive outcomes prediction, but also is a powerful tool for fusing different modalities.

5 Acknowledgment

References

1. A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.
2. J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.
3. T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(Apr):615–637, 2005.
4. M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12(Jul):2211–2268, 2011.

5. Z. Huo, D. Shen, and H. Huang. New multi-task learning model to predict alzheimers disease cognitive assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 317–325. Springer, 2016.

6. P. Jawanpuria and J. S. Nath. Multi-task multiple kernel learning. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 828–838. SIAM, 2011.

7. S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th annual international conference on machine learning*, pages 457–464. ACM, 2009.

8. C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of machine learning research*, 6(Jul):1099–1125, 2005.

9. A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9(Nov):2491–2521, 2008.

10. A. Rakotomamonjy, R. Flamary, G. Gasso, and S. Canu. lp lq penalty for sparse linear and sparse multiple kernel multitask learning. *IEEE Transactions on Neural Networks*, 22(8):1307–1320, 2011.

11. J. Wan, Z. Zhang, B. D. Rao, S. Fang, J. Yan, A. J. Saykin, and L. Shen. Identifying the neuroanatomical basis of cognitive impairment in alzheimer's disease by correlation-and nonlinearity-aware sparse bayesian learning. *IEEE transactions on medical imaging*, 33(7):1475–1487, 2014.

12. J. Wan, Z. Zhang, J. Yan, T. Li, B. D. Rao, S. Fang, S. Kim, S. L. Risacher, A. J. Saykin, and L. Shen. Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer's disease. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 940–947, 2012.

13. H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A. J. Saykin, L. Shen, et al. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 557–562. IEEE, 2011.

14. J. Yan, T. Li, H. Wang, H. Huang, J. Wan, K. Nho, S. Kim, S. L. Risacher, A. J. Saykin, L. Shen, et al. Cortical surface biomarkers for predicting cognitive outcomes using group l 2, 1 norm. *Neurobiology of aging*, 36:S185–S193, 2015.

15. D. Zhang, D. Shen, A. D. N. Initiative, et al. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *NeuroImage*, 59(2):895–907, 2012.

16. J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *Advances in neural information processing systems*, pages 702–710, 2011.