Feature-aware Multi-task feature learning for Predicting Cognitive Outcomes in Alzheimer's disease

Peng Cao
College of Computer Science and Engineering，Key Laboratory of Intelligent Computing in Medical Image，Ministry of Education，
Northeastern University
Shenyang, China
caopeng@cse.neu.edu.cn

Shanshan Tang
College of Information Science and Engineering，Northeastern University
Shenyang, China
1700892@stu.neu.edu.cn

Min Huang
College of Information Science and Engineering，Northeastern University
Shenyang, China
mhuang@ise.neu.edu.cn

Jinzhu Yang*
College of Computer Science and Engineering，Northeastern University
Shenyang, China
yangjinzhu@cse.neu.edu.cn

Dazhe Zhao
College of Computer Science and Engineering，Northeastern University
Shenyang, China
zhaodz@neusoft.com

Amine Trabelsi
Department of Computing Science，
University of Alberta
Edmonton, Canada
atrabels@ualberta.ca

Osmar Zaiane
Department of Computing Science
University of Alberta
Edmonton, Canada
zaiane@ualberta.ca

***Abstract*—Machine learning algorithms and multivariate data analysis methods have been widely utilized in the field of Alzheimer's disease (AD) research in recent years. Predicting cognitive performance of subjects from neuroimage measures and identifying relevant imaging biomarkers are important research topics in the study of Alzheimer's disease. Multi-task based feature learning (MTFL) have been widely studied to select a discriminative feature subset from MRI features, and improve the performance by incorporating inherent correlations among multiple clinical cognitive measures. It is known that the brain imaging measures are often correlated with each other, and AD is closely related to the inter-correlation among different brain regions. However, the multi-task based feature learning (MTFL) method neglects the inherent correlation among brain imaging measures. We present a novel regularized multi-task learning approach via a joint sparsity-inducing regularization to effectively incorporate both a relatedness among multiple cognitive score prediction tasks and a useful inherent correlation between brain imaging measures by exploiting correlations among features. It allows the simultaneous selection of a common set of biomarkers for all tasks and the preservation of the inherent structure of imaging measures. The reported experiments on the ADNI dataset show that the proposed method is effective and promising.**

Keywords—Alzheimer's disease, Regression model, Multi-task learning, Magnetic resonance imaging, Biomarkers discovery

I. INTRODUCTION

Alzheimer's disease (AD) is a degenerative brain disease, which mainly affects memory function, ultimately culminating in a dementia state characterized by the progressive loss of memory and cognitive functions[1-2]. The accurate diagnosis of Alzheimer's disease (AD) plays a significant role in patient care, especially at the early stage, because awareness of the severity and risk of progression allows patients to benefit from early intervention, symptomatic treatment for cognitive losses as well as associated behavioral problems before irreversible brain damages are shaped [3].

Many machine learning approaches have been developed to automate detection, diagnosis and quantification of disease [4-6]. The standard diagnosis of AD patients typically begins with a series of neuropsychological tests. Many cognitive measures have been designed to clinically evaluate the cognitive status of the patients and used as important criteria for clinical diagnosis of probable AD, such as Alzheimer's Disease Assessment Scale cognitive total score (ADAS) and Mini Mental State Exam score (MMSE). The regularized multivariate regression model is adopted to associate the imaging markers and the cognitive measures [7-10]. Nowadays, to improve the generalization performance of the predicative regression model, multi-task learning methods are developed to predict the cognitive outcomes by incorporating inherent correlations among multiple clinical cognitive measures [7-10]. Multi-task learning is a learning paradigm which seeks to improve the generalization performance of a learning task with the help of some other related tasks [11]. To overcome the curse of dimensionality in the clinical data from the neuroimaging, the most recent studies [10,12,13] employed sparsity inducing regularized multi-task models with $l_{2,1}$-norm [14] to identify the features that are relevant to all clinical scores.

For the existing MTFL methods, a major limitation is that it either selects a feature as relevant to all tasks or excludes it from all models. It has been shown that the brain regions are inter-connected for AD patients, resulting in cognitive decline in AD patients. This is very useful to characterize the task relationship since strong correlated features are likely to have similar model parameters. It motivates us to consider the correlation structure in MRI measures. A feature-aware regularization is proposed by incorporating the inter-feature correlation effects into the MTFL. The feature-aware scheme not only allows to improve the performance of multi-task learning, but also enhances the understanding of the relationships among the features. We can thus employ the ADMM optimization method [15] to efficiently solve the convex optimization problem of the proposed formulation. In summary, the main contributions of this paper are as follow:

- We propose a new regularization for considering the feature structure by flexibly modeling the feature correlation.
- We propose a new feature-aware multi-task learning approach via a joint sparsity-inducing regularization to effectively integrate the correlation between several cognitive score prediction tasks and the useful inherent correlation between brain MRI imaging measurements.
- We conduct extensive experiments using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) with different multi-task learning settings, to demonstrate the effectiveness of our method along various dimensions including prediction performance on the baseline cognitive outcomes and biomarkers identification.

II. Multi-task feature learning

A. Problem formulation

We formulate the prediction of cognitive outcomes problem in this paper as follows. Magnetic resonance imaging (MRI) provides a chance to directly observe brain changes such as cerebral atrophy or ventricular expansion. A number of structural neuroimaging studies have shown that brain atrophy detected by MRI is correlated with cognitive test performance [16-17]. The aim of our work is to predict subjects' cognitive scores (e.g. ADAS, MMSE) using their MRI features (e.g. volume, area and thickness) across the entire brain. Here the prediction problem is captured by a regularized multivariate regression model representing the relationships between structural changes in MRI features (imaging markers) and the cognitive measures. MRI features and cognitive measures (outcomes) are treated as inputs to and outputs of the regression model, respectively. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n \times p}$ be MRI features (e.g. the volume of hippocampus), where n and p are the number of training instances and dimensionality of $\boldsymbol{x}_i$, $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]^T \in \mathbb{R}^{n \times t}$, where $\boldsymbol{y}_i$ is the target cognitive score for $\boldsymbol{x}_i$ and t is the number of tasks, $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_t] \in \mathbb{R}^{p \times t}$, where $\boldsymbol{w}_h$ is the weight vector for the h-th task. In the regression model related to the h-th prediction task, a subject i's cognitive score, under that task, is represented as a linear function of the corresponding MRI features. Analytically, this can be specified by the following regression equation:

$$y_{ih} = \boldsymbol{x}_i^T \boldsymbol{w}_h + \xi_{ih},\ i = 1, \ldots, n;\ h = 1, \ldots, t \quad (1)$$

B. Multi-task learning

Multi-task learning (MTL) [10] is a learning paradigm which seeks to improve the generalization performance of all tasks involved. The fundamental hypothesis of the MTL methods is to assume that if tasks are related then learning of one task can benefit from the learning of other tasks. Learning multiple related tasks simultaneously has been theoretically and empirically shown to often significantly improve the performance. The key of the MTL is how to exploit the correlation among the tasks via an appropriate shared representation. It is known that there exist inherent correlations among different cognitive scores. Analyzing the high-dimensional image measures in the ADNI is a challenging and poses great difficulties to traditional statistical methods. Since not all the brain regions are associated with AD, many of the features are irrelevant and redundant. To better exploit the correlation of tasks and identify the most important biomarkers from high-dimensional image measures, multi-task feature learning with sparsity-inducing norms are applied on the prediction of cognitive score outcomes to produce better performance and to learn a shared subset of features. In the multi-task feature learning, the $l_{2,1}$-norm regularizer imposes the sparsity between all features and non-sparsity between tasks, the features that are discriminative for all tasks will get large weights. The objective function of the $l_{2,1}$-norm regularized MTL (called multi-task feature learning, MTFL) is given by:

$$\min_{\boldsymbol{W}} \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_F^2 + \lambda_1 \|\boldsymbol{W}\|_{2,1} \quad (2)$$

where λ is regularization parameter.

III. Feature aware Multi-task feature learning (FAS-MTFL)

In this section, we introduce the proposed feature-aware multi-task learning model to account for both the correlations of tasks and the features of the cognitive outcomes. First, we introduce feature correlation matrix indicating the inter-correlation among the features. Then we show how to integrate the estimated prior knowledge of inter-correlation into the multi-task feature learning process. Finally, we present the optimization of the proposed feature-aware multi-task feature learning algorithm by ADMM.

A. Feature correlation matrix and graph

Existing multi-task learning, such as MTFL, exploit only the correlation in tasks, neglecting the potentially grouping information among multiple neuroimaging measures. Motivated by the above observations, we propose a feature aware multi-task learning algorithm. Firstly, we construct a feature correlation matrix based on each pairwise feature link on the training data. We propose to construct a correlation matrix to capture the similarity among features. The symmetric correlation matrix $\boldsymbol{C} \in \mathbb{R}^{p \times p}$ encodes the inter-correlation among the features. The correlations correspond to the off-diagonal entries of the correlation matrix of the data. The estimation of correlation value c_{ml} of the m-th feature and l-th feature is calculated as:

$$c_{ml} = \frac{cov(X_m, X_l)}{\sigma_{X_m} \sigma_{X_l}} = \frac{E[(X_m - \mu_{X_m})(X_l - \mu_{X_l})]}{\sigma_{X_m} \sigma_{X_l}} \quad (3)$$

Fig. 2 shows the correlation distrubition in all features. It shows how different regions of the brain are inter-connected.

We find that the strength of pairwise feature correlation strength is unstable. To build a more stable correlation matrix, a threshold technique is applied to connect only highly correlated features. Two features are considered as highly correlated if their correlation coefficient exceeds a given threshold τ. We chose a τ of 0.6 and set reserved correlation coefficients to 1. The calculated correlation matrix is shown in Fig. 3. The useful connectivity among different brain regions can be identified from the estimated feature correlation matrix [18-19].

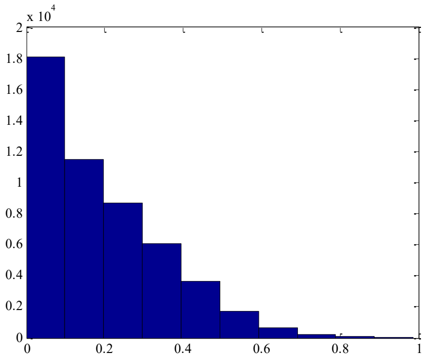

Fig. 2. Distribution of pairwise feature correlation. (*The horizontal axis denotes the correlation value, and the vertical axis denotes the amount of the feature pairs corresponding to the correlation value.*)

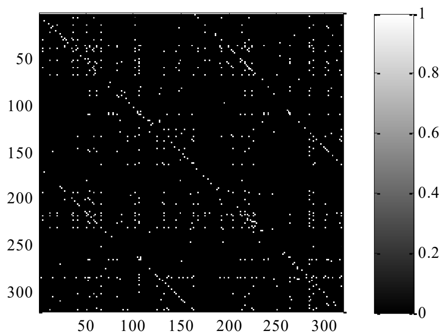

Fig. 3. The estimated feature correlation matrix

B. *Feature-aware sparsity-inducing regularization*

Given the estimated correlation matrix, we aim to utilize the correlation information among the features to help improve the performance of MTFL. To achieve this, a feature-aware sparsity-inducing regularization (FAS) is developed. The regularization of FAS encourages the correlated features to take similar values by shrinking the difference between them toward zero. If we regard features as points and connect feature pairs whose correlation coefficients are greater than τ with undirected edges, then we can get an undirected graph. The edge connecting a feature pair, such as feature m and l, is represented by $e(m,l) \in E$. Then the feature-aware sparsity-inducing norm is defined as:

$$\|\boldsymbol{SW}\|_1 = \sum_{e(m,l)\in E} |s_{m,l}| \left\| \boldsymbol{w}^m - sign(s_{m,l})\boldsymbol{w}^l \right\|_1 \quad (4)$$

where $\boldsymbol{w}^i$ is the i-th row of matrix W corresponding to i-th feature coefficients in all tasks. $\boldsymbol{S}$ is a normlaized version of correlation matrix $\boldsymbol{C}$. $\boldsymbol{C}$ is normalized by the number of correlation edges, $k = |E|$, and the matrix $\boldsymbol{S}$ is defined as:

$$s_{m,l} = \begin{cases} -\frac{c_{m,l}}{k} & (m,l) \in E, m \neq l \\ \frac{\sum_{m=1, m\neq l}^{p} |c_{m,l}|}{k} & (m,l) \in E, m = l \\ 0 & otherwise \end{cases} \quad (5)$$

C. *Formulation of Feature aware Multi-task feature learning*

To preserve the strength of the correlations among the features during modeling the multi-task learning, the penalty expressed in Eq. (4) is inserted to the formulation in Eq. (2). The objective function of feature aware Multi-task feature learning is given in the following optimization problem:

$$\min_{\boldsymbol{W}} \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{XW}\|_F^2 + \lambda_1 \|\boldsymbol{W}\|_{2,1} + \lambda_2 \|\boldsymbol{SW}\|_1 \quad (6)$$

We consider prior feature correlation modeling and multi-task learning simultaneously in an unified framework. The objective contains two regularization processes: (1) all tasks are regularized by an $l_{2,1}$-norm; (2) a feature-aware sparsity-inducing norm is enforced on the features. Incorporating the prior knowledge of the correlation can constrain the hypothesis space by a joint sparsity-inducing regularization.

D. *Optimization*

In this section, we present a novel solver for problem (6) based on the ADMM. It is easy to show that the objective function is convex. However, the proposed optimization problem is difficult to solve. This is due to the non-smooth of the two sparse norms including $l_{2,1}$-norm and feature-aware norm. In order to effectively deal with the non-smoothness nature of the two constraints, we propose an optimization method built on the Alternating Direction Method of Multipliers (ADMM). An effective ADMM-based algorithm is proposed for that purpose.

$$\min_{\boldsymbol{W},\boldsymbol{Q},\boldsymbol{R}} \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{XW}\|_F^2 + \lambda_1 \|\boldsymbol{Q}\|_{2,1} + \lambda_2 \|\boldsymbol{R}\|_1$$
$$s.t.\ \boldsymbol{W} - \boldsymbol{Q} = 0\ ,\ \boldsymbol{SW} - \boldsymbol{R} = 0 \quad (7)$$

where Q, S are slack variables.

The augmented Lagrangian of Eq. (7) is:

$$L_\rho(\boldsymbol{W},\boldsymbol{Q},\boldsymbol{R},\boldsymbol{U}_{(1)},\boldsymbol{U}_{(2)}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{XW}\|_F^2 + \lambda_1\|\boldsymbol{Q}\|_{2,1}$$
$$+\lambda_2\|\boldsymbol{R}\|_1 + \langle \boldsymbol{U}_{(1)}, \boldsymbol{W} - \boldsymbol{Q} \rangle + \frac{\rho}{2}\|\boldsymbol{W} - \boldsymbol{Q}\|^2$$
$$+\langle \boldsymbol{U}_{(2)}, \boldsymbol{SW} - \boldsymbol{R} \rangle + \frac{\rho}{2}\|\boldsymbol{SW} - \boldsymbol{R}\|^2 \quad (8)$$

where $\boldsymbol{U}_{(1)}$ and $\boldsymbol{U}_{(2)}$ are augmented Lagrangian multipliers.

Update ***W***: According to the augmented Lagrangian in Eq. (8), the update of W can be carried out by:

$$\boldsymbol{W}^{t+1} = arg\min_{\boldsymbol{W}} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{XW}\|_F^2 + \langle \boldsymbol{U}_{(1)}^t, \boldsymbol{W} - \boldsymbol{Q}^t \rangle$$
$$+\frac{\rho}{2}\|\boldsymbol{W} - \boldsymbol{Q}^t\|^2 + \langle \boldsymbol{U}_{(2)}^t, \boldsymbol{SW} - \boldsymbol{R}^t \rangle + \frac{\rho}{2}\|\boldsymbol{SW} - \boldsymbol{R}^t\|^2 \quad (9)$$

$\boldsymbol{W}^{t+1}$ can be updated efficiently using Cholesky factorization. The optimal solution is given by $\boldsymbol{W}^{t+1} = \boldsymbol{F}^{-1}\boldsymbol{B}^t$, where

$$\boldsymbol{F} = \boldsymbol{X}^T\boldsymbol{X} + \rho\boldsymbol{I} + \rho\boldsymbol{SS} \quad (10)$$

$$\boldsymbol{B}^t = \boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{U}_{(1)}^t + \rho\boldsymbol{Q}^t - \boldsymbol{U}_{(2)}^t S + \rho\boldsymbol{S}\boldsymbol{R}^t \quad (11)$$

Update $\boldsymbol{Q}$: According to the augmented Lagrangian in Eq. (8), the update of **Q** can be solved as follow

$$\boldsymbol{Q}^{t+1} = arg\,\min_{\boldsymbol{Q}} \frac{1}{2}\left\|\boldsymbol{Q} - \boldsymbol{\Lambda}_{(1)}\right\|^2 + \frac{\lambda_1}{\rho}|\boldsymbol{Q}|_{2,1} \quad (12)$$

where $\boldsymbol{\Lambda}_{(1)} = \boldsymbol{W}^{t+1} + \frac{\boldsymbol{U}_{(1)}^t}{\rho}$.

The solution to Eq. (12) is computed using the following formula:

$$\boldsymbol{Q}^{t+1} = \frac{max\left\{\left\|\Lambda_{(1)}\right\|_2 - \frac{\lambda_1}{\rho}, 0\right\}}{\left\|\Lambda_{(1)}\right\|_2}\boldsymbol{\Lambda}_{(1)} \quad (13)$$

Update $\boldsymbol{R}$: According to the augmented Lagrangian in Eq. (8), the update of **R** can be carried out by:

$$\boldsymbol{R}^{t+1} = arg\,\min_{\boldsymbol{R}} \frac{1}{2}\left\|\boldsymbol{R} - \boldsymbol{\Lambda}_{(2)}\right\|^2 + \frac{\lambda_2}{\rho}\|\boldsymbol{R}\|_1 \quad (14)$$

where $\boldsymbol{\Lambda}_{(2)} = \boldsymbol{S}\boldsymbol{W}^{t+1} + \frac{\boldsymbol{U}_{(2)}^t}{\rho}$.

The solution to Eq. (14) is computed using the following formula:

$$\boldsymbol{R}^{t+1} = sign(\boldsymbol{\Lambda}_{(2)})\,max\left\{\left|\boldsymbol{\Lambda}_{(2)}\right| - \frac{\lambda_2}{2\rho}, 0\right\} \quad (15)$$

Update $\boldsymbol{U}_{(1)}$ and $\boldsymbol{U}_{(2)}$: According to the standard ADMM, we can get the following updated formula:

$$\boldsymbol{U}_{(1)}^{t+1} = \boldsymbol{U}_{(1)}^t + \rho(\boldsymbol{W}^{t+1} - \boldsymbol{Q}^{t+1}) \quad (16)$$
$$\boldsymbol{U}_{(2)}^{t+1} = \boldsymbol{U}_{(2)}^t + \rho(\boldsymbol{S}\boldsymbol{W}^{t+1} - \boldsymbol{R}^{t+1}) \quad (17)$$

The complete algorithm is described in Algorithm 1.

Algorithm 1 ADMM optimization of FAS-MTFL

Require: $\boldsymbol{X}, \boldsymbol{Y}, \lambda_1, \lambda_2, \rho, \boldsymbol{S}$

Initialization: $\boldsymbol{W}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{U}_{(1)}, \boldsymbol{U}_{(2)}$

Compute the Cholesky factorization of $\boldsymbol{F}$.

Repeat

Update $\boldsymbol{W}^{t+1}$ according to Eq. (9-11).

Update $\boldsymbol{Q}^{t+1}$ according to Eq. (13).

Update $\boldsymbol{R}^{t+1}$ according to Eq. (15).

Update $\boldsymbol{U}_{(1)}^{t+1}$ according to Eq. (16).

Update $\boldsymbol{U}_{(2)}^{t+1}$ according to Eq. (17).

Until Convergence.

IV. Experiments

A. Data and experimental setting

Totally, 48 cortical regions and 44 subcortical regions are generated. For each cortical region, the cortical thickness average (TA), standard deviation of thickness (TS), surface area (SA) and cortical volume (CV) were calculated as features. For each subcortical region, subcortical volume was calculated as features. The SA of left and right hemisphere and total intracranial volume (ICV) were also included. This yielded a total of p = 319 MRI features extracted from cortical/subcortical brain regions (region-of-interest, ROI) in each hemisphere. (including 275 cortical and 44 subcortical features from 115 brain ROI totally). In this work, the number samples n = 788 subjects.

We randomly split the data into training and testing sets using a ratio 9:1. In each of 10 trials, a 5-fold nested cross validation procedure is employed to tune the regularization parameters. Data was z-scored before applying regression methods. The range of each parameter varied from 0.1 to 1000. The reported results were the best results of each method with the optimal parameter. The threshold τ is empirically chosen as 0.5.

For the quantitative performance evaluation, we employed the metrics of Correlation Coefficient ($\mathrm{CC}(y,\hat{y}) = \frac{cov(y,\hat{y})}{\sigma(y)\sigma(\hat{y})}$) and Root Mean Squared Error ($\mathrm{rMSE}(y,\hat{y}) = \frac{\|y-\hat{y}\|_2^2}{n}$) between the predicted clinical scores and the target clinical scores for each regression task. Moreover, to evaluate the overall performance on all the tasks, the normalized mean squared error ($\mathrm{nMSE}(Y,\hat{Y}) = \frac{\sum_{i=1}^{t}\frac{\|Y_i-\hat{Y}_i\|_2^2}{\sigma(Y_i)}}{\sum_{i=1}^{t} n_i}$) and weighted R-value ($\mathrm{wR}(Y,\hat{Y}) = \frac{\sum_{i=1}^{t}\mathrm{CC}(Y_i,\hat{Y}_i)n_i}{\sum_{i=1}^{t} n_i}$) are used.

B. Comparision

In this section, we conduct empirical evaluation for the proposed methods by comparing with Lasso, Ridge and MTFL. Five cognitive scores are examined including ADAS, MMSE, RAVLT-TOTAL, RAVLT-T30 and RAVLT-RECOG, which are commonly used in modeling the relationship between MRI and cognitive performance. The average and standard deviation of performance measures are calculated by 10 fold cross validation. It is worth noting that we use the same training and testing data across the experiments for all the methods for fair comparison. The experimental results are reported in Table II and III where the best results are boldfaced. A first glance at the results shows that FAS-MTFL generally outperforms all other compared methods on both metrics of nMSE and CC. Compared with the single task learning (Lasso and Ridge), both the multi-task feature learning methods improve the prediction performance by utilizing different intrinsic relationships among multiple related tasks. Moreover, FAS-MTFL obtains a better performance compared with MTFL, which verifies the benefits of utilizing the knowledge from feature correlation to assist the traditional multi-task learning.

TABLE. II. Performance comparison of various methods in terms of CC and wR on five cognitive prediction tasks. (*avg*(*std*))

Method	ADAS	MMSE	TOTAL	T30	RECOG	wR
Lasso	0.640 (0.056)	0.536 (0.059)	0.447 (0.059)	0.488 (0.116)	0.366 (0.109)	0.495 (0.077)
Ridge	0.529 (0.061)	0.331 (0.059)	0.347 (0.133)	0.330 (0.140)	0.240 (0.125)	0.355 (0.080)
MTFL	0.636 (0.087)	0.542 (0.074)	**0.507** **(0.114)**	0.496 (0.114)	**0.425** **(0.127)**	0.521 (0.088)
FAS-MTFL	**0.664** **(0.068)**	**0.549** **(0.063)**	0.504 (0.100)	**0.514** **(0.118)**	0.416 (0.125)	**0.529** **(0.081)**

TABLE. III. Performance comparison of various methods in terms of rMSE and nMSE on five cognitive prediction tasks. (*avg*(*std*))

Method	ADAS	MMSE	TOTAL	T30	RECOG	nMSE
Lasso	6.919 (0.365)	2.190 (0.090)	10.465 (0.670)	3.489 (0.229)	3.741 (0.316)	4.773 (0.442)
Ridge	8.416 (0.530)	2.949 (0.247)	12.531 (1.060)	4.497 (0.313)	4.761 (0.560)	7.342 (1.073)
MTFL	6.987 (0.573)	2.343 (0.123)	9.851 (0.817)	3.484 (0.298)	3.615 (0.177)	4.561 (0.296)
FAS-MTFL	**6.675** **(0.472)**	**2.173** **(0.098)**	**9.801** **(0.654)**	**3.426** **(0.243)**	**3.612** **(0.253)**	**4.372** **(0.328)**

The patterns of correlated features will provide useful imaging-based biomarkers. We investigate the effect of feature correlation in the multi-task feature learning. Through the experiments in the biomarker discovery, the amount of selected features with nonzero weight in FAS-MTFL is more than MTFL on the five tasks since the feature aware scheme takes the correlation in features into account. Some features with zero weight become nonzero due to its strong correlation with other important features. For example, the feature of TA-R.Inferiorparietal and TA-L.MiddleTemporal is correlated (the correlation value is 0.66). In the experiment of multi-task learning with five tasks, the weight coefficient of TA-R.Inferiorparietal is zero but the one of L.MiddleTemporal is not zero obtained from MTFL. We find the corresponding weight coefficients of both features are not zeros in FAS-MTFL, which is the influence of the Feature-aware sparsity-inducing norm regularization in FAS-MTFL. In addition, Inferiorparietal is associated with progression from healthy aging to Alzheimer's disease [21], and it is also identified by our FAS-MTFL method as the one of the top 10 features. The results demonstrate that feature-aware sparsity-inducing regularization can amend the major drawback of MTFL.

V. Conlusion

In this paper, we tackle multi-task feature learning for predicting cognitive outcomes in Alzheimer's disease by exploiting the correlation structure of brain imaging measurements. To achieve this, we propose feature aware multi-task feature learning, which employs a joint sparsity-inducing regularization with generalized fused lasso for features and $l_{2,1}$-norm for the tasks. Experiments on benchmark datasets show the effectiveness of the proposed feature aware multi-task feature learning. The correlation matrix plays an important role in feature-aware MTFL. For future work, we plan to investigate other types of correlation calculation (such as inverse covariance matrix). Moreover, we are interested in optimizing the feature correlation during the multi-task learning rather than doing prior calculation.

References

[1] Alzheimer's Association, et al., "Alzheimer's disease facts and figures, Alzheimer Dementia", 2016, 12 (4): 459–509.

[2] N. Batsch, M. Mittelman, "World Alzheimer Report 2012: Overcoming the Stigma of Dementia", 2015.

[3] E. Pellegrini, L. Ballerini, M.D.C.V. Hernandez, F.M.Chappell, V. González-Castro, D. Anblagan, S. Danso, et al, "Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review", Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 2018, 10, pp.519-535.

[4] G. Fiscon, E.Weitschek, C. De, G. Felici and P. Bertolazzi, "An integrated approach based on EEG signals processing combined with supervised methods to classify Alzheimer's disease patients", Proceedings of 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 2750-2752.

[5] J. Pu, J. Wang, Y. Zheng, H. Ye, W. Shen, Y. Li and H. Zha , "Boosting Alzheimer diagnosis accuracy with the help of incomplete privileged information", 2017, Proceedings of 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017, pp. 595-599.

[6] M. Hon and N.M. Khan, "Towards Alzheimer's disease classification through transfer learning", Proceedings of 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) , 2017, pp. 1166-1169.

[7] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A.D.N. Initiative, et al., "Multimodal classification of Alzheimer's disease and mild cognitive impairment", Neuroimage, 2011, 55 (3): 856–867.

[8] H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A.J. Saykin, L. Shen, ADNI, "Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance", Proceedings of the International Conference on Computer Vision, 2011, pp. 6–13.

[9] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, L. Shen, "High-order multi-task feature learning to identify longitudinal phenotypic markers for Alzheimer's disease progression prediction", Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1277-1285.

[10] J. Zhou, L.Yuan, J.Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression", Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 814-822.

[11] Y. Zhang and Q. Yang, "A survey on multi-task learning", 2017, arXiv preprint arXiv:1707.08114.

[12] H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A.J. Saykin, L. Shen, "Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance", Proceedings of the International Conference on Computer Vision, 2011, pp. 6–13

[13] D. Zhang and D. Shen, "Multi-modal multi-task learning for joint prediction of clinical scores in Alzheimer's disease", International Workshop on Multimodal Brain Image Analysis, 2011, pp. 60-67.

[14] J. Liu, S. Ji and J. Ye, "Multi-task feature learning via efficient $l_{2,1}$-norm minimization", Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence, 2009, pp. 339-348.

[15] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers", Foundations and Trends in Machine learning, 2011, 3(1), pp.1-122.

[16] J. Wan, Z. Zhang, J. Yan, T. Li, B.D. Rao, S. Fang, S. Kim, S.L. Risacher, A.J. Saykin, L. Shen, "Sparse Bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer's disease", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 940–947.

[17] J. Wan, Z. Zhang, B.D. Rao, S. Fang, J. Yan, A.J. Saykin, L. Shen, "Identifying the neuroanatomical basis of cognitive impairment in Alzheimer's disease by correlation-and nonlinearity-aware sparse bayesian learning", IEEE Trans. Med. Imaging, 2014, 33 (7): 1475–1487.

[18] L. Sun, R. Patel, J. Liu, K. Chen, T. Wu, J. Li, E. Reiman and J. Ye, "Mining brain region connectivity for alzheimer's disease study via sparse inverse covariance estimation", In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 1335-1344.

[19] I. O. Korolev, L. S. Laura ymonds, C. B. Andrea, and Alzheimer's Disease Neuroimaging Initiative, "Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification", PloS one, 2016, 11(2): e0138866.

[20] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression", In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 814-822.

[21] S.J. Greene, R.J. Killiany and Alzheimer's Disease Neuroimaging Initiative, "Subregions of the inferior parietal lobule are affected in the progression to Alzheimer's disease", Neurobiology of aging, 2010, 31(8):1304-131