Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning

Hua Wang[1], Feiping Nie[1], Heng Huang[1,*], Shannon L. Risacher[2], Andrew J. Saykin[2] and Li Shen[2,*]; For the Alzheimer's Disease Neuroimaging Initiative[†]

[1]Department of Computer Science and Engineering, University of Texas at Arlington, TX 76019 and [2]Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN 46202, USA

ABSTRACT

Motivation: Recent advances in brain imaging and high-throughput genotyping techniques enable new approaches to study the influence of genetic and anatomical variations on brain functions and disorders. Traditional association studies typically perform independent and pairwise analysis among neuroimaging measures, cognitive scores and disease status, and ignore the important underlying interacting relationships between these units.

Results: To overcome this limitation, in this article, we propose a new sparse multimodal multitask learning method to reveal complex relationships from gene to brain to symptom. Our main contributions are three-fold: (i) introducing combined structured sparsity regularizations into multimodal multitask learning to integrate multidimensional heterogeneous imaging genetics data and identify multimodal biomarkers; (ii) utilizing a joint classification and regression learning model to identify disease-sensitive and cognition-relevant biomarkers; (iii) deriving a new efficient optimization algorithm to solve our non-smooth objective function and providing rigorous theoretical analysis on the global optimum convergency. Using the imaging genetics data from the Alzheimer's Disease Neuroimaging Initiative database, the effectiveness of the proposed method is demonstrated by clearly improved performance on predicting both cognitive scores and disease status. The identified multimodal biomarkers could predict not only disease status but also cognitive function to help elucidate the biological pathway from gene to brain structure and function, and to cognition and disease.

Availability: Software is publicly available at: http://ranger.uta.edu/%7eheng/multimodal/

Contact: heng@uta.edu; shenli@iupui.edu

1 INTRODUCTION

Recent advances in acquiring multimodal brain imaging and genome-wide array data provide exciting new opportunities to study the influence of genetic variation on brain structure and function. Research in this emerging field, known as *imaging genetics*, holds great promise for a system biology of the brain to better understand complex neurobiological systems, from genetic determinants to cellular processes to the complex interplay of brain structure, function, behavior and cognition. Analysis of these multimodal

datasets will facilitate early diagnosis, deepen mechanistic understanding and improved treatment of brain disorders.

Machine learning methods have been widely employed to predict Alzheimer's disease (AD) status using imaging genetics measures (Batmanghelich *et al.*, 2009; Fan *et al.*, 2008; Hinrichs *et al.*, 2009b; Shen *et al.*, 2010a). Since AD is a neurodegenerative disorder characterized by progressive impairment of memory and other cognitive functions, regression models have also been investigated to predict clinical scores from structural, such as magnetic resonance imaging (MRI), and/or molecular, such as fluorodeoxyglucose positron emission tomography (FDG-PET), neuroimaging data (Stonnington *et al.*, 2010; Walhovd *et al.*, 2010). For example, Walhovd *et al.* (2010) performed stepwise regression in a pairwise fashion to relate each of MRI and FDG-PET measures of eight candidate regions to each of four Rey's Auditory Verbal Learning Test (RAVLT) memory scores. This univariate approach, however, did not consider either interrelated structures within imaging data or those within cognitive data. Using relevance vector regression, Stonnington *et al.* (2010) jointly analyzed the voxel-based morphometry (VBM) features extracted from the entire brain to predict each selected clinical score, while the investigations of different clinical scores are independent from each other.

One goal of imaging genetics is to identify genetic risk factors and/or imaging biomarkers via intermediate quantitative traits (QTs, e.g. cognitive memory scores used in this article) on the chain from gene to brain to symptom. Thus, both disease classification and QT prediction are important machine learning tasks. Prior imaging genetics research typically employs a two-step procedure for identifying risk factors and biomarkers: one first determines disease-relevant QTs, and then detects the biomarkers associated with these QTs. Since a QT could be related to many genetic or imaging markers on different pathways that are not all disease specific (e.g. QT 2 and Gene 3 in Fig. 1), an ideal scenario would be to discover only those markers associated with both QT and disease status for a better understanding of the underlying biological pathway specific to the disease.

On the other hand, identifying genetic and phenotypic biomarkers from large-scale multidimensional heterogeneous data is an important biomedical and biological research topic. Unlike simple feature selection working on a single data source, multimodal learning describes the setting of learning from data where observations are represented by multiple types of feature sets. Many multimodal methods have been developed for classification and clustering purposes, such as co-training (Abney, 2002; Brefeld and Scheffer, 2004; Ghani, 2002; Nigam

---

*To whom correspondence should be addressed.
†Data for the Alzheimer's Disease Neuroimaging Initiative are provided in the Acknowledgement section.
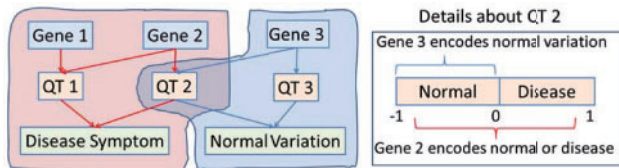
Fig. 1. A simplified schematic example of two pathways from gene to QTs to phenotypic endpoints: the red one is disease relevant while the blue one yields only normal variation. Traditional two-stage imaging genetic strategy identifies QT 1 and QT 2 first and then Genes 1, 2, 3. Our new method will identify only disease relevant genes (i.e. Gene 1 and Gene 2); and Gene 3 would not be identified because it cannot be used to classify disease status

Fig. 2. The proposed sparse multimodal multitask feature selection method will identify biomarkers from multimodal heterogeneous data resources. The identified biomarkers could predict not only disease status, but also cognitive functions to help researchers better understand the underlying mechanism from gene to brain structure and function, and to cognition and disease

et al., 2000) and multiview clustering (Bickel and Scheffer, 2004; Dhillon *et al.*, 2003). However, they typically assume that the multimodal feature sets are conditionally independent, which does not hold in many real-world applications such as imaging genetics. Considering different representations give rise to different kernel functions, several Multiple Kernel Learning (MKL) approaches (Bach *et al.*, 2004; Hinrichs *et al.*, 2009a; Kloft *et al.*, 2008; Lanckriet *et al.*, 2004; Rakotomamonjy *et al.*, 2007; Sonnenburg *et al.*, 2006; Suykens *et al.*, 2002; Ye *et al.*, 2008; Yu *et al.*, 2010; Zien and Ong, 2007) have been recently studied and employed to integrate heterogeneous data and select multitype features. However, such models train a single weight for all features from the same modality, i.e. all features from the same data source are weighted equally, when they are combined with the features from other sources. This limitation often yields inadequate performance.

To address the above challenges, we propose a new sparse multimodal multitask learning algorithm that integrates heterogeneous genetic and phenotypic data effectively and efficiently to identify disease-sensitive and cognition-relevant biomarkers from multiple data sources. Different to LASSO (Tibshirani, 1996), group LASSO (Yuan and Lin, 2006) and other related methods that mainly find the biomarkers correlated to each individual QT (memory score), we consider predicting each memory score as a regression task and select biomarkers that tend to play an important role in influencing multiple tasks. A joint classification and regression multitask learning model is utilized to select the biomarkers correlated to memory scores and disease categories simultaneously.

Sparsity regularizations have recently been widely investigated and applied to multitask learning models (Argyriou *et al.*, 2007; Kim and Xing, 2010; Micchelli *et al.*, 2010; Obozinski *et al.*, 2006, 2010; Sun *et al.*, 2009). Sparse representations are typically achieved by imposing non-smooth norms as regularizers in the optimization problems. From the view of sparsity organization, we have two types: (i) The flat sparsity is often achieved by ℓ_0-norm or ℓ_1-norm regularizer or trace norm in matrix/tensor completion. Optimization techniques include LARS (Efron *et al.*, 2004), linear gradient search (Liu *et al.*, 2009), proximal methods (Beck and Teboulle, 2009). (ii) The structured sparsity is usually obtained through different sparse regularizers such as $\ell_{2,1}$-norm (Kim and Xing, 2010; Obozinski *et al.*, 2010; Sun *et al.*, 2009), $\ell_{2,0}$-norm (Luo *et al.*, 2010), $\ell_{\infty,1}$-norm (Quattoni *et al.*, 2009) (also denoted as $\ell_{1,2}$-norm, $\ell_{1,\infty}$-norm in different papers) and group ℓ_1-norm (Yuan and Lin, 2006) which can be solved by methods in Micchelli *et al.* (2010) and Argyriou *et al.* (2008). We propose a new combined structured sparse
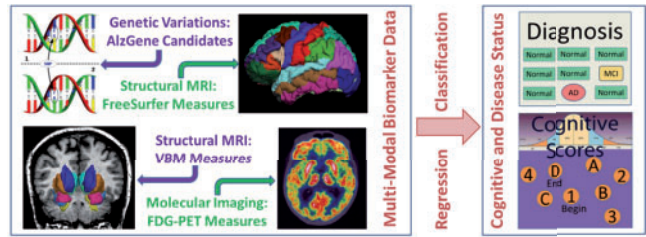
regularization to integrate features from different modalities and to learn a weight for each feature leading to a more flexible scheme for feature selection in data integration, which is illustrated in Figure 3. In our combined structured sparse regularization, the group ℓ_1-norm regularization (blue circles in Fig. 3) learns the feature global importance, i.e. the modal-wise feature importance of every data modality on each class (task), and the $\ell_{2,1}$-norm regularization (red circles in Fig. 3) explores the feature local importance, i.e. the importance of each feature for multiple classes/tasks. The proposed method is applied to identify AD-sensitive biomarkers associated with the cognitive scores by integrating heterogeneous genetic and phenotypic data (as shown in Fig. 2). Our empirical results yield clearly improved performance on predicting both cognitive scores and disease status.

2 IDENTIFYING DISEASE SENSITIVE AND QT-RELEVANT BIOMARKERS FROM HETEROGENEOUS IMAGING GENETICS DATA

Pairwise univariate correlation analysis can quickly provide important association information between genetic/phenotypic data and QTs. However, it treats the features and the QTs as independent and isolated units, therefore the underlying interacting relationships between the units might be lost. We propose a new sparse multimodal multitask learning model to reveal genetic and phenotypic biomarkers, which are disease sensitive and QT-relevant, by simultaneously and systematically taking into account an ensemble of SNPs (single nucleotide polymorphism) and phenotypic signatures and jointly performing two heterogeneous tasks, i.e. biomarker-to-QT regression and biomarker-to-disease classification. The QTs studied in this article are the cognitive scores.

In multitask learning, given a set of input variables (i.e. features such as SNPs and MRI/PET measures), we are interested in learning a set of related models (e.g. relations between genetic/imaging markers and cognitive scores) to predict multiple outcomes (i.e. tasks such as predicting cognitive scores and disease status). Because these tasks are relevant, they share a common input space. As a result, it is desirable to learn all the models jointly rather than treating each task as independent and fitting each model separately, such as Lasso (Tibshirani, 1996) and group Lasso (Yuan and Lin, 2006). Such multitask learning can discover robust patterns (because significant patterns in a single task could be outliers for other tasks) and potentially increase the predictive power.

In this article, we write matrices as uppercase letters and vectors as boldface lowercase letters. Given a matrix $W = [w_{ij}]$, its i-th row and j-th column are denoted as $\mathbf{w}^i$ and $\mathbf{w}_j$, respectively. The $\ell_{2,1}$-norm of the matrix W is defined as $||W||_{2,1} = \sum_{i=1} ||\mathbf{w}^i||_2$ (also denoted as $\ell_{1,2}$-norm by other researchers).

2.1 Heterogeneous data integration via combined structured sparse regularizations

First, we will systematically propose our new multimodal learning method to integrate and select the genetic and phenotypic biomarkers from large-scale heterogeneous data. In the supervised learning setting, we are given n training samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, where $\mathbf{x}_i = (\mathbf{x}_i^1, \cdots, \mathbf{x}_i^k)^T \in \Re^d$ is the input vector including all features from a total of k different modalities and each modality j has d_j features ($d = \sum_{j=1}^k d_j$). $\mathbf{y}_i \in \Re^c$ is the class label vector of data point $\mathbf{x}_i$ (only one element in $\mathbf{y}_i$ is 1, and others are zeros), where c is the number of classes (tasks). Let $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \Re^{d \times n}$ and $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_c] \in \Re^{c \times n}$. Different to MKL, we directly learn a $d \times c$ parameter matrix as:

$$W = \begin{bmatrix} \mathbf{w}_1^1 & \cdots & \mathbf{w}_c^1 \\ \cdots & \cdots & \cdots \\ \mathbf{w}_1^k & \cdots & \mathbf{w}_c^k \end{bmatrix} \in \Re^{d \times c}, \quad (1)$$

where $\mathbf{w}_p^q \in \Re^{d_q}$ indicates the weights of all features in the q-th modality with respect to the p-th task (class). Typically, we can use a convex loss function $\mathcal{L}(X, W)$ to measure the loss incurred by W on the training samples. Compared with MKL approaches that learn one weight for one kernel matrix representing one modality, our method will learn the weight for each feature to capture the local feature importance. Since the features come from heterogeneous data sources, we impose the regularizer $\mathcal{R}(W)$ to capture the interrelationships of modalities and features as:

$$\min_W \mathcal{L}(X, W) + \gamma \mathcal{R}(W), \quad (2)$$

where γ is a trade-off parameter. In heterogeneous data fusion, from multiview perspective of view, the features of a specific view (modality) can be more or less discriminative for different tasks (classes). Thus, we propose a new group ℓ_1-norm (G_1-norm) as a regularization term in Equation (2), which is defined over W as following:

$$\|W\|_{G_1} = \sum_{i=1}^c \sum_{j=1}^k ||\mathbf{w}_i^j||_2, \quad (3)$$

which is illustrated by the blue circles in Figure 3. Then the Equation (2) becomes:

$$\min_W \mathcal{L}(X, W) + \gamma_1 \|W\|_{G_1}. \quad (4)$$

Since the group ℓ_1-norm uses ℓ_2-norm within each modality and ℓ_1-norm between modalities, it enforces the sparsity between different modalities, i.e. if one modality of features are not discriminative for certain tasks, the objective in Equation (4) will assign zeros (in ideal case, usually they are very small values) to them for corresponding tasks; otherwise, their weights are large. This new group ℓ_1-norm regularizer captures the global relationships between data modalities.
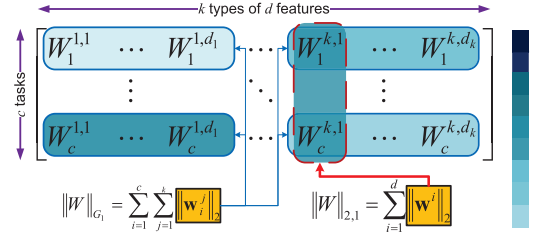
Fig. 3. Illustration of the feature weight matrix W^T. The elements in matrix with deep blue color have large values. The group ℓ_1-norm (G_1-norm) emphasizes the learning of the *group-wise* weights for a type of features (e.g. all the SNPs features, or all the MRI imaging features, or all the FDG-PET imaging features) corresponding to each task (e.g. the prediction for a disease status or a memory score) and the $\ell_{2,1}$-norm accentuates the *individual* weight learning cross multiple tasks

However, in certain cases, even if most features in one modality are not discriminative for the classification or regression tasks, a small number of features in the same modality can still be highly discriminative. From the multitask learning point of view, such important features should be shared by all/most tasks. Thus, we add an additional $\ell_{2,1}$-norm regularizer into Equation (4) as:

$$\min_W \mathcal{L}(X, W) + \gamma_1 \|W\|_{G_1} + \gamma_2 \|W\|_{2,1}. \quad (5)$$

The $\ell_{2,1}$-norm was popularly used in multitask feature selection (Argyriou *et al.*, 2008; Obozinski *et al.*, 2010). Since the $\ell_{2,1}$-norm regularizer impose the sparsity between all features and non-sparsity between tasks, the features that are discriminative for all tasks will get large weights.

Our regularization items consider the heterogeneous features from both group-wise and individual viewpoints. Figure 3 visualizes the matrix W^T as a demonstration. In Figure 3, the elements with deep blue color have large values. The group ℓ_1-norm emphasizes the group-wise weights learning corresponding to each task and the $\ell_{2,1}$-norm accentuates the individual weight learning cross multiple tasks. Through the combined regularizations, for each task (class), many features (not all of them) in the discriminative modalities and a small number of features (may not be none) in the non-discriminative modalities will learn large weights as the important and discriminative features.

The multidimensional data integration has been increasingly important to many biological and biomedical studies. So far, the MKL methods are most widely used. Due to the learning model deficiency, the MKL methods cannot explore both modality-wise importance and individual importance of features simultaneously. Our new structured sparse multimodal learning method integrates the multidimensional data in a more efficient and effective way. The loss function $\mathcal{L}(X, W)$ in Equation (8) can be replace by either least square loss function or logistic regression loss function to perform regression/classification tasks.

2.2 Joint disease classification and QT regression

Since we are interested in identifying the disease-sensitive and QT-relevant biomarkers, we consider performing both logistic regression for classifying disease status and multivariate regression for predicting cognitive memory scores simultaneously (Wang *et al.*, 2011). A similar model was used in Yang *et al.* (2009) for

heterogeneous multitask learning. Regular multitask learning only considers homogeneous tasks such as regression or classification individually. Joint classification and regression can be regarded as a learning paradigm for handling heterogeneous tasks.

First, logistic regression is used for disease classification, which minimizes the following loss function:

$$\mathcal{L}_1(W) = \sum_{i=1}^{n} \sum_{k=1}^{c_1} \left(y_{ik} \log \sum_{l=1}^{c_1} e^{\mathbf{w}_l^T \mathbf{x}_i} - y_{ik} \mathbf{w}_k^T \mathbf{x}_i \right). \quad (6)$$

Here, we perform three binary classification tasks for the following three diagnostic groups respectively ($c_1 = 3$): AD, mild cognitive impairment (MCI), and health control (HC).

Second, we use the traditional multivariate least squares regression model to predict memory scores. Under the regression matrix $P \in \Re^{d \times c_2}$, the least squares loss is defined by

$$\mathcal{L}_2(P) = \left\| X^T P - Z \right\|_F^2, \quad (7)$$

where X is the data points matrix, P is the coefficient matrix of regression with c_2 tasks, the label matrix $Z = \left[\left(\mathbf{z}^1\right)^T, \left(\mathbf{z}^2\right)^T, \cdots, \left(\mathbf{z}^n\right)^T \right]^T \in \Re^{n \times c_2}$.

We perform the joint classification and regression tasks, the disease-sensitive and QT-relevant biomarker identification task can be formulated as the following objective:

$$\min_{V} \sum_{i=1}^{n} \sum_{k=1}^{c_1} \left(y_{ik} \log \sum_{l=1}^{c_1} e^{\mathbf{w}_l^T \mathbf{x}_i} - y_{ik} \mathbf{w}_k^T \mathbf{x}_i \right) \\ + \left\| X^T P - Z \right\|_F^2 + \gamma_1 \|V\|_{G_1} + \gamma_2 \|V\|_{2,1}, \quad (8)$$

where $V = [W \; P] \in \Re^{d \times (c_1 + c_2)}$. As a result, the identified biomarkers will be correlated to memory scores and also be discriminative to disease categories.

Since the objective in Equation (8) is a non-smooth problem and cannot be easily solved in general, we derive a new efficient algorithm to solve this problem in the next subsection.

2.3 Optimization algorithm

We take the derivatives of Equation (8) with respect to W and P respectively, and set them to zeros, we have

$$\frac{\partial \mathcal{L}_1(W)}{\partial W} + 2\gamma_1 \sum_{i=1}^{c_1} D_i \mathbf{w}_i + 2\gamma_2 DW = 0, \quad (9)$$

$$2XX^T P - 2XZ + 2\gamma_1 \sum_{i=c_1+1}^{c_2} D_i \mathbf{p}_i + 2\gamma_2 DP = 0, \quad (10)$$

where $D_i (1 \le i \le c_1 + c_2)$ is a block diagonal matrix with the k-th diagonal block as $\frac{1}{2\|\mathbf{v}_i^k\|_2} I_k$ (I_k is a d_k by d_k identity matrix), D is a diagonal matrix with the k-th diagonal element as $\frac{1}{2\|\mathbf{v}^k\|_2}$. Since $D_i (1 \le i \le c_1 + c_2)$ and D depend on $V = [W \; P]$, they are also unknown variables to be optimized. In this article, we provide an iterative algorithm to solve Equation (8). First, we guess a random solution $V \in \Re^{d \times (c_1 + c_2)}$, then we calculate the matrices $D_i (1 \le i \le c_1 + c_2)$ and D according to the current solution V. After

obtaining the $D_i (1 \le i \le c_1 + c_2)$ and D, we can update the solution $V = [W \; P]$ based on Equation (9). Specifically, the i-th column of P is updated by $\mathbf{p}_i = (XX^T + \gamma_1 D_i + \gamma_2 D)^{-1} X \mathbf{z}_i$. We cannot update W with a closed form solution based on Equation (9), but we can obtained the updated W by the Newton's method. According to Equation (9), we need to solve the following problem:

$$\min_{W} \mathcal{L}_1(W) + \gamma_1 \sum_{i=1}^{c_1} \mathbf{w}_i^T D_i \mathbf{w}_i + \gamma_2 Tr(W^T DW). \quad (11)$$

Similar to the traditional method in the logistic regression (Krishnapuram *et al.*, 2005; Lee *et al.*, 2006), we can use the Newton's method to obtain the solution W.

For the first term, the traditional logistic regression derivatives can be applied to get the first-and second-order derivatives (Lee *et al.*, 2006).

For the second term, the first-and second-order derivatives are

$$\frac{\partial \sum_{i=1}^{c_1} \mathbf{w}_i^T D_i \mathbf{w}_i}{\partial W_{up}} = 2D_p(u,u)W_{up},$$

$$\frac{\partial \sum_{i=1}^{c_1} \mathbf{w}_i^T D_i \mathbf{w}_i}{\partial W_{up} \partial W_{vq}} = 2D_p(u,u)\delta_{uv}\delta_{pq}, \quad (12)$$

where $D_p(u,u)$ is the u-th diagonal element of D_p.

For the third term, the first-and second-order derivatives are

$$\frac{\partial Tr(W^T DW)}{\partial W_{up}} = 2D(u,u)W_{up},$$

$$\frac{\partial Tr(W^T DW)}{\partial W_{up} \partial W_{vq}} = 2D(u,u)\delta_{uv}\delta_{pq}. \quad (13)$$

After obtaining the updated solution $V = [W \; P]$, we can calculate the new matrices $D_i (1 \le i \le c_1 + c_2)$ and D. This procedure is repeated until the algorithm converges. The detailed algorithm is listed in Algorithm 1. We will prove that the above algorithm will converge to the global optimum.

2.4 Algorithm analysis

To prove the convergence of the proposed algorithm, we need a lemma as follows.

LEMMA 1. *For any vectors* $\mathbf{v}$ *and* $\mathbf{v}_0$, *we have the following inequality:* $\|\mathbf{v}\|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}_0\|_2} \le \|\mathbf{v}_0\|_2 - \frac{\|\mathbf{v}_0\|_2^2}{2\|\mathbf{v}_0\|_2}$.

PROOF. Obviously, $-(\|\mathbf{v}\|_2 - \|\mathbf{v}_0\|_2)^2 \le 0$, so we have

$$-(\|\mathbf{v}\|_2 - \|\mathbf{v}_0\|_2)^2 \le 0 \Rightarrow 2\|\mathbf{v}\|_2 \|\mathbf{v}_0\|_2 - \|\mathbf{v}\|_2^2 \le \|\mathbf{v}_0\|_2^2$$

$$\Rightarrow \|\mathbf{v}\|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}_0\|_2} \le \|\mathbf{v}_0\|_2 - \frac{\|\mathbf{v}_0\|_2^2}{2\|\mathbf{v}_0\|_2}, \quad (14)$$

which completes the proof. □

Then we prove the convergence of the algorithm, which is described in the following theorem.

THEOREM 1. *The algorithm decreases the objective value of problem (8) in each iteration.*

PROOF. In each iteration, suppose the updated W is $\breve{W}$, and the updated P is $\breve{P}$, then the updated V is $\tilde{V} = [\,\breve{W}\ \breve{P}\,]$. From Step 3 in the Algorithm 1, we know that:

$$\mathcal{L}_1(\breve{W}) + \gamma_1 \sum_{i=1}^{c_1} \tilde{\mathbf{w}}_i^T D_i \tilde{\mathbf{w}}_i + \gamma_2 Tr(\breve{W}^T D \breve{W})$$
$$\leq \mathcal{L}_1(W) + \gamma_1 \sum_{i=1}^{c_1} \mathbf{w}_i^T D_i \mathbf{w}_i + \gamma_2 Tr(W^T DW). \tag{15}$$

According to Step 4, we have:

$$\left\| X^T \breve{P} - Y \right\|_F^2 + \gamma_1 \sum_{i=1}^{c_2} \tilde{\mathbf{p}}_i^T D_i \tilde{\mathbf{p}}_i + \gamma_2 Tr(\breve{P}^T D \breve{P})$$
$$\leq \left\| X^T P - Y \right\|_F^2 + \gamma_1 \sum_{i=1}^{c_2} \mathbf{p}_i^T D_i \mathbf{p}_i + \gamma_2 Tr(P^T DP). \tag{16}$$

Based on the definitions of $D_i (1 \leq i \leq c_1 + c_2)$ and D, and Lemma 1, we have two following inequalities:

$$\sum_{k=1}^{K} \left\| \tilde{\mathbf{v}}_i^k \right\|_2 - \sum_{k=1}^{K} \frac{\left\| \tilde{\mathbf{v}}_i^k \right\|_2^2}{2 \left\| \mathbf{v}_i^k \right\|_2} \leq \sum_{k=1}^{K} \left\| \mathbf{v}_i^k \right\|_2 - \sum_{k=1}^{K} \frac{\left\| \mathbf{v}_i^k \right\|_2^2}{2 \left\| \mathbf{v}_i^k \right\|_2}$$
$$\Rightarrow \sum_{k=1}^{K} \left\| \tilde{\mathbf{v}}_i^k \right\|_2 - \tilde{\mathbf{v}}_i^T D_i \tilde{\mathbf{v}}_i \leq \sum_{k=1}^{K} \left\| \mathbf{v}_i^k \right\|_2 - \mathbf{v}_i^T D_i \mathbf{v}_i$$
$$\Rightarrow \gamma_1 \sum_{i=1}^{c_1+c_2} \sum_{k=1}^{K} \left\| \tilde{\mathbf{v}}_i^k \right\|_2 - \gamma_1 \sum_{i=1}^{c_1+c_2} \tilde{\mathbf{v}}_i^T D_i \tilde{\mathbf{v}}_i$$
$$\leq \gamma_1 \sum_{i=1}^{c_1+c_2} \sum_{k=1}^{K} \left\| \mathbf{v}_i^k \right\|_2 - \gamma_1 \sum_{i=1}^{c_1+c_2} \mathbf{v}_i^T D_i \mathbf{v}_i, \tag{17}$$

and

$$\sum_{k=1}^{d} \left\| \tilde{\mathbf{v}}^k \right\|_2 - \sum_{k=1}^{d} \frac{\left\| \tilde{\mathbf{v}}^k \right\|_2^2}{2 \left\| \mathbf{v}^k \right\|_2} \leq \sum_{k=1}^{d} \left\| \mathbf{v}^k \right\|_2 - \sum_{k=1}^{d} \frac{\left\| \mathbf{v}^k \right\|_2^2}{2 \left\| \mathbf{v}^k \right\|_2}$$
$$\Rightarrow \gamma_2 \sum_{k=1}^{d} \left\| \tilde{\mathbf{v}}^k \right\|_2 - \gamma_2 Tr(\tilde{V}^T D \tilde{V})$$
$$\leq \gamma_2 \sum_{k=1}^{d} \left\| \mathbf{v}^k \right\|_2 - \gamma_2 Tr(V^T DV). \tag{18}$$

Note that the following two equalities:

$$\sum_{i=1}^{c_1+c_2} \mathbf{v}_i^T D_i \mathbf{v}_i = \sum_{i=1}^{c_1} \mathbf{w}_i^T D_i \mathbf{w}_i + \sum_{i=1}^{c_2} \mathbf{p}_i^T D_i \mathbf{p}_i, \tag{19}$$
$$Tr(V^T DV) = Tr(W^T DW) + Tr(P^T DP),$$

Algorithm 1 An efficient iterative algorithm to solve the optimization problem in Equation (8).

Input: $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \Re^{d \times n}$, $Y = \left[\left(\mathbf{y}^1\right)^T, \left(\mathbf{y}^2\right)^T, \cdots, \left(\mathbf{y}^n\right)^T \right]^T \in \{0,1\}^{n \times c_1}$ and $Z = \left[\left(\mathbf{z}^1\right)^T, \left(\mathbf{z}^2\right)^T, \cdots, \left(\mathbf{z}^n\right)^T \right]^T \in \Re^{n \times c_2}$.

Output: $W \in \Re^{d \times c_1}$ and $P \in \Re^{d \times c_2}$.

1. Initialize $W \in \Re^{d \times c_1}$, $P \in \Re^{d \times c_2}$. Let $V = [\,W\ P\,] \in \Re^{d \times (c_1+c_2)}$.

repeat

2. Calculate the block diagonal matrices $D_i (1 \leq i \leq c_1 + c_2)$, where the k-th diagonal block of D_i is $\frac{1}{2\|\mathbf{v}_i^k\|_2} I_k$. Calculate the diagonal matrix D, where the k-th diagonal element is $\frac{1}{2\|\mathbf{v}^k\|_2}$.

3. Update $\mathbf{w}$ by $\mathbf{w} - B^{-1}a$, where the $d*(p-1) + u(1 \leq u \leq d, 1 \leq p \leq c_1)$-th element of $a \in \Re^{dc_1 \times 1}$ is $\frac{\partial\left(\mathcal{L}_1(W) + \gamma_1 \sum_{i=1}^{c_1} \mathbf{w}_i^T D_i \mathbf{w}_i + \gamma_2 Tr(W^T DW)\right)}{\partial W_{up}}$, the $(d*(p-1) + u, d*(q-1)+v)(1 \leq u, v \leq d, 1 \leq p, q \leq c_1)$-th element of $B \in \Re^{dc_1 \times dc_1}$ is $\frac{\partial\left(\mathcal{L}_1(W) + \gamma_1 \sum_{i=1}^{c_1} \mathbf{w}_i^T D_i \mathbf{w}_i + \gamma_2 Tr(W^T DW)\right)}{\partial W_{up} \partial W_{vq}}$.

Construct the updated $W \in \Re^{d \times c_1}$ by the updated vector $\mathbf{w} \in \Re^{dc_1}$, where the (u,p)-th element of W is the $(d*(p-1)+u)$-th element of $\mathbf{w}$.

4. Update the i-th column of P by $\mathbf{p}_i = (XX^T + \gamma_1 D_i + \gamma_2 D)^{-1} X \mathbf{z}_i$.

5. Update the V by $V = [\,W\ P\,]$.

until Converges

---

then by adding Equations (15–18) in the both sides, we arrive at

$$\mathcal{L}_1(\breve{W}) + \mathcal{L}_2(\breve{P}) + \gamma_1 \sum_{i=1}^{c_1+c_2} \sum_{k=1}^{K} \left\| \tilde{\mathbf{v}}_i^k \right\|_2 + \gamma_2 \sum_{k=1}^{d} \left\| \tilde{\mathbf{v}}^k \right\|_2$$
$$\leq \mathcal{L}_1(W) + \mathcal{L}_2(P) + \gamma_1 \sum_{i=1}^{c_1+c_2} \sum_{k=1}^{K} \left\| \mathbf{v}_i^k \right\|_2 + \gamma_2 \sum_{k=1}^{d} \left\| \mathbf{v}^k \right\|_2.$$

Therefore, the algorithm decreases the objective value of problem (8) in each iteration. □

In the convergence, W, P, $D_i (1 \leq i \leq c_1 + c_2)$ and D satisfy the Equation (9). As the Equation (8) is a convex problem, satisfying the Equation (9) indicates that $V = [\,W\ P\,]$ is a global optimum solution to the Equation (8). Therefore, the Algorithm 1 will converge to the global optimum of the Equation (8). Since our algorithm has the closed form solution in each iteration, the convergency is very fast.

3 EMPIRICAL STUDIES AND DISCUSSIONS

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). One goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (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, see www.adni-info.org. Following a prior imaging genetics study (Shen *et al.*, 2010b), 733 non-Hispanic Caucasian participants were included in this study. We empirically evaluate the proposed method by applying it to the ADNI cohort, where a wide range of multimodal biomarkers are examined and selected to predict memory performance measured by five RAVLT scores and classify participants into HC, MCI and AD.

3.1 Experimental design

Overall setting: our primary goal is to identify relevant genetic and imaging biomarkers that can classify disease status and predict memory scores (Fig. 2). We describe our genotyping, imaging and memory data in Section 3.1; present the identified biomarkers in Section 3.2; discuss the disease classification in Section 3.3; and demonstrate the memory score prediction in Section 3.4.

Genotyping data: the single-nucleotide polymorphism (SNP) data (Saykin *et al.*, 2010) were genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA, USA). Among all SNPs, only SNPs, belonging to the top 40 AD candidate genes listed on the AlzGene database (www.alzgene.org) as of June 10, 2010, were selected after the standard quality control (QC) and imputation steps. The QC criteria for the SNP data include (i) call rate check per subject and per SNP marker, (ii) gender check, (iii) sibling pair identification, (iv) the Hardy–Weinberg equilibrium test, (v) marker removal by the minor allele frequency and (vi) population stratification. The quality-controlled SNPs were then imputed using the MaCH software to estimate the missing genotypes. After that, the Illumina annotation information based on the Genome build 36.2 was used to select a subset of SNPs, belonging or proximal to the top 40 AD candidate genes. This procedure yielded 1224 SNPs, which were annotated with 37 genes (Wang *et al.*, 2012). For the remaining 3 genes, no SNPs were available on the genotyping chip.

Imaging biomarkers: in this study, we use the baseline structural MRI and molecular FDG-PET scans, from which we extract imaging biomarkers. Two widely employed automated MRI analysis techniques were used to process and extract imaging genotypes across the brain from all baseline scans of ADNI participants as previously described (Shen *et al.*, 2010b). First, voxel-based morphometry (VBM) (Ashburner and Friston, 2000) was performed to define global gray matter (GM) density maps and extract local GM density values for 86 target regions (Fig. 4a). Second, automated parcellation via freeSurfer V4 (Fischl *et al.*, 2002) was conducted to define 56 volumetric and cortical thickness values (Fig. 4b) and to extract total intracranial volume (ICV). Further information about these measures is available in Shen *et al.* (2010b). All these measures were adjusted for the baseline age, gender, education, handedness and baseline ICV using the regression weights derived from the healthy control participants. For PET images, following Landau *et al.* (2009), mean glucose metabolism (CMglu) measures of 26 regions of interest (ROIs) in the Montreal Neurological Institute (MNI) brain atlas space were employed in this study (Fig. 4c).

Memory data: The cognitive measures we use to test the proposed method are the baseline RAVLT memory scores from all ADNI participants. The standard RAVLT format starts with a list of 15 unrelated words (List A) repeated over five different trials and
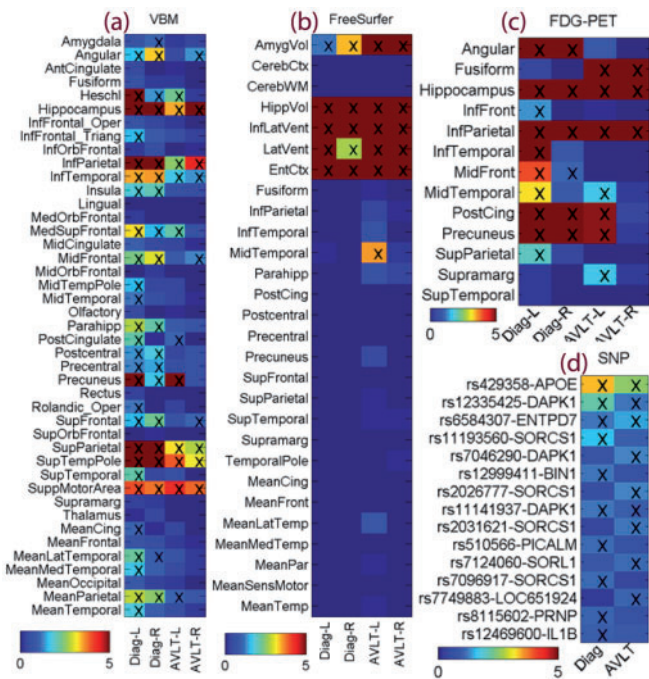
Fig. 4. Weight maps for multimodal data: **(a)** VBM measures from MRI, **(b)** FreeSurfer measures from MRI, **(c)** glucose metabolism from FDG-PET, and **(d)** top SNP findings. Weights for disease classification were labeled as Diag-L (left side), Diag-R (right side) or Diag; and weights for RAVLT regression were labeled as AVLT-L, AVLT-R or AVLT. In (a–c), weights were normalized by dividing the corresponding threshold used for feature selection, and thus all selected features had normalized weights ≥1 and were marked with 'x'. In (d), only top SNPs were shown, weights were normalized by dividing the weight of the 10th top SNP, and the top 10 SNPs for either classification or regression task had normalized weights ≥1 and were marked with 'x'

Table 1. RAVLT cognitive measures as responses in multitask learning

Task ID	Description of RAVLT scores
TOTAL	Total score of the first 5 learning trials
TOT6	Trial 6 total number of words recalled
TOTB	List B total number of words recalled
T30	30 minute delay total number of words recalled
RECOG	30 minute delay recognition score

participants are asked to repeat. Then the examiner presents a second list of 15 words (List B), and the participant is asked to remember as many words as possible from List A. Trial 6, termed as 5 min recall, requests the participant again to recall as many words as possible from List A, without reading it again. Trial 7, termed as 30 min recall, is administrated in the same way as Trial 6, but after a 30 min delay. Finally, a recognition test with 30 words read aloud, requesting the participant to indicate whether or not each word is on List A. The RAVLT has proven useful in evaluating verbal learning and memory. Table 1 summarizes five RAVLT scores used in our experiments.

Table 2. Multimodal feature sets as predictors in multiview learning

View ID (feature set ID)	Modality	No. of features
VBM	MRI	86
FreeSurfer	MRI	56
FDG-PET	FDG-PET	26
SNP	Genetics	1244

Participant selection: In this study, we included only participants with no missing data for all above four types (views) of features and cognitive scores, which resulted in a set of 345 subjects (83 HC, 174 MCI and 88 AD). The feature sets extracted from baseline multimodal data of these subjects are summarized in Table 2.

3.2 Biomarker identifications

The proposed heterogeneous multitask learning scheme aims to identify genetic and phenotypic biomarkers that are associated with both cognition (e.g. RAVLT in this study) and disease status in a joint regression and classification framework. Here we first examine the identified biomarkers. Shown in Figure 4 is a summarization of selected features for all four data types, where the regression/classification weights are color-mapped for each feature and each task.

In Figure 4a, many VBM measures are selected to be associated with disease status, which is in accordance with known global brain atrophy pattern in AD. The VBM measures associated with RAVLT scores seem to be a subset of those disease-sensitive markers, showing a specific memory circuitry contributing to the disease, as well as suggesting that the disease is implicated by not only this memory function but also other complicated factors. Evidently, the proposed method could have a potential to offer deep mechanistic understandings. Shown in Figure 5 is a comparison between RAVLT-relevant markers and AD-relevant markers and their associated weights mapped onto a standard brain space.

Figure 4b shows the identified markers from the FreeSurfer data. In this case, a small set of markers are discovered. These markers, such as hippocampal volume, amygdala volume and entorhinal cortex thickness, are all well-known AD-relevant markers, showing the effectiveness of the proposed method. These markers are also shown to be associated with both AD and RAVLT. The FDG-PET findings (Fig. 4c) are also interesting and promising. The AD-relevant biomarkers include angular, hippocampus, middle temporal and post cingulate regions, which agrees with prior findings e.g. Landau *et al.* (2009). Again, a subset of these markers are also relevant to RAVTL scores.

As to the genetics, only top findings are shown in Figure 4d. The APOE E4 SNP (rs429358), the best known AD risk factor, shows the strongest link to both disease status and RAVLT scores. A few other important AD genes, including recently discovered and replicated PICALM and BIN1, are also included in the results. For those newly identified SNPs, further investigation in independent cohorts should be warranted.

3.3 Improved disease classification

We classify the selected participants of ADNI cohort using the proposed methods by integrating the four different types of data.
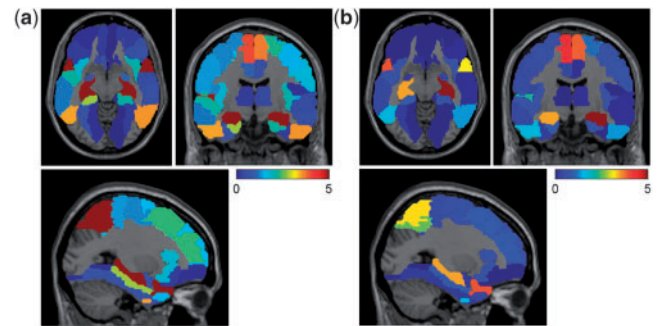
Fig. 5. VBM weights of joint regression of AVLT scores and classification of disease status were mapped onto brain **(a)** Overall weights for disease classification; **(b)** Overall weights for AVLT regression

We report the classification performances of our method. We compare our methods against several most recent MKL methods that are able to make use of multiple types of data including SVM ℓ_∞ MKL method (Sonnenburg *et al.*, 2006), SVM ℓ_1 MKL (Lanckriet *et al.*, 2004), SVM ℓ_2 MKL method (Kloft *et al.*, 2008), least square (LSSVM) ℓ_∞ MKL method (Ye *et al.*, 2008), LSSVM ℓ_1 MKL method (Suykens *et al.*, 2002) and LSSVM ℓ_2 MKL method (Yu *et al.*, 2010). We also compare a related method, Heterogeneous Multitask Learning (HML) method (Yang *et al.*, 2009), which simultaneously conducts classification and regression like our method. However, because this method is designed for homogenous input data and is not able to deal with multiple types of data at the same time, we concatenate the four types of features as its input. In addition, we report the classification performances by our method and SVM on each individual types of data as baselines. SVM on a simple concatenation of all four types of features are also reported. In our experiments, we conduct three-class classification, which is more desirable and more challenging than binary classifications using each pair of three categories.

We conduct standard 5-fold cross-validation and report the average results. For each of the five trials, within the training data, an internal 5-fold cross-validation is performed to fine tune the parameters. The parameters of our methods [γ_1 and γ_2 in Equation (8)] are optimized in the range of $\left\{10^{-5}, 10^{-4}, \ldots, 10^4, 10^5\right\}$. For SVM method and MKL methods, one Gaussian kernel is constructed for each type of features $\left[i.e.\mathcal{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2\right)\right]$, where the parameters γ are fine tuned in the same range used as our method. We implement the MKL methods using the codes published by Yu *et al.* (2010). Following Yu *et al.* (2010), in LSSVM ℓ_∞ and ℓ_2 methods, the regularization parameter λ is estimated jointly as the kernel coefficient of an identity matrix; in LSSVM ℓ_1 method, λ is set to 1; in all other SVM approaches, the *C* parameter of the box constraint is set to 1. We use LIBSVM (http://www.csie.ntu.edu.tw/ cjlin/libsvm/) software package to implement SVM. We implement HML method following the details in its original work, and set the parameters to be optimal. The classification performances measured by classification accuracy of all compared methods in AD detection are reported in Table 3.

A first glance at the results shows that our methods consistently outperform all other compared methods, which demonstrates the effectiveness of our methods in early AD detection. In addition,

Table 3. Classification performance comparison between the proposed method and related methods for distinguishing HC, MCI and AD

Methods	Accuracy (mean + SD)
SVM (SNP)	$0.561 \pm .026$
SVM (FreeSurfer)	$0.573 \pm .012$
SVM (VBM)	$0.541 \pm .032$
SVM (PET)	$0.535 \pm .026$
SVM (all)	$0.575 \pm .019$
HML (all)	$0.638 \pm .019$
SVM ℓ_∞ MKL method	0.624 ± 0.031
SVM ℓ_1 MKL method	0.593 ± 0.042
SVM ℓ_2 MKL method	0.561 ± 0.037
LSSVM ℓ_∞ MKL method	0.614 ± 0.031
LSSVM ℓ_1 MKL method	0.585 ± 0.018
LSSVM ℓ_2 MKL method	0.577 ± 0.033
Our method (SNP)	0.673 ± 0.021
Our method (FreeSurfer)	0.689 ± 0.029
Our method (VBM)	0.669 ± 0.031
Our method (PET)	0.621 ± 0.028
Our method	0.726 ± 0.032

Table 4. Comparison of memory prediction performance measured by average RMSEs (smaller is better)

Test case	TOTAL	TOT6	TOTB	T30	RECOG
MRV (SNP)	6.153	2.476	2.168	2.201	3.483
MRV (FreeSurfer)	5.928	2.235	2.039	2.088	3.339
MRV (VBM)	6.093	2.289	2.142	2.137	3.394
MRV (PET)	6.246	2.514	2.237	2.215	3.615
MRV (all)	5.909	2.232	1.992	2.032	3.306
Ridge (SNP)	6.076	2.416	2.147	2.117	3.368
Ridge (FreeSurfer)	5.757	2.203	2.004	2.017	3.237
Ridge (VBM)	5.976	2.147	2.038	2.129	3.249
Ridge (PET)	6.153	2.443	2.186	2.107	3.515
Ridge (all)	5.704	2.143	1.989	1.994	3.193
Our method (SNP)	5.991	2.201	2.008	2.001	3.107
Our method (FreeSurfer)	5.601	2.106	1.947	1.886	3.015
Our method (VBM)	5.715	2.011	1.899	1.974	3.041
Our method (PET)	6.013	2.241	2.017	2.017	3.331
Our method (all)	5.506	1.984	1.886	1.841	2.989

the methods using multiple data sources are generally better than their counterparts using one single type of data. This confirms the usefulness of data integration in AD diagnosis. Moreover, our methods always outperform the MKL methods in these experiments, although both take advantage of multiple data sources. This observation is consistent with our theoretical analysis. That is, our methods not only assign proper weight to each type of data, but also consider the relevance of the features inside each individual type of data. In contrast, the MKL methods address the former while not taking into account the latter.

3.4 Improved memory performance prediction

Now we evaluate the memory performance prediction capability of the proposed method. Since the cognitive scores are continuous, we evaluate the proposed method via regression and compare it to two baseline methods, i.e. multivariate linear regression (MRV) and ridge regression. Since both MRV and ridge regression are for single-type input data, we conduct regression on each of the four types of features and a simple concatenation of them. Similarly, we also predict memory performance by our method on the same test conditions. When multiple-type input data are used, as demonstrated in Section 3.2, our method automatically and adaptively select the prominent biomarkers for regression. For each test case, we conduct standard 5-fold cross-validation and report the average results. For each of the five trials, within the training data, an internal 5-fold cross-validation is performed to fine tune the parameters in the range of $\left\{10^{-5}, 10^{-4}, \ldots, 10^4, 10^5\right\}$ for both ridge regression and our method. For our method, in each trial, from the learned coefficient matrix we sum the absolute values of the coefficients of a single feature over all the tasks as the overall weight, from which we pick up the features with non-zero weights (i.e. $w > 10^{-3}$) to predict regression responses for test data. The performance assessed by root mean square error (RMSE), a widely used measurement for statistical regression analysis, are reported in Table 4.

From Table 4 we can see that the proposed method always has better memory prediction performance. Among the test cases, the FreeSurfer imaging measures and VBM imaging measure have similar predictive power, which are better than those of PET imaging measures and SNP features. In general, combining the four types of features are better than only using one type of data. Since our method adaptively weight each type of data and each feature inside a type of data, it has the least regression error when using all available input data. These results, again, demonstrated the usefulness of our method and data integration in early AD diagnosis.

4 CONCLUSIONS

We proposed a novel sparse multimodal multitask learning method to identify the disease-sensitive biomarkers via integrating heterogeneous imaging genetics data. We utilized the joint classification and regression learning model to identify the disease-sensitive and QT-relevant biomarkers. We introduced a novel combined structured sparsity regularization to integrate heterogeneous imaging genetics data, and derived a new efficient optimization algorithm to solve our non-smooth objective function and followed with the rigorous theoretical analysis on the global convergency. The empirical results showed our method improved both memory scores prediction and disease classification accuracy.

Conflict of Interest: none declared.

REFERENCES

Abney,S. (2002) Bootstrapping. *Annual Meeting of the Association for Computational Linguistics*, pp. 360–367.

Argyriou,A. *et al*. (2007) Multi-task feature learning. In *Advances in Neural Information Processing System (NIPS)*, The MIT Press, pp. 41–48.

Argyriou,A. *et al*. (2008) Convex multitask feature learning. *Machine Learning*, **73**, 243–272.

Ashburner,J. and Friston,K. (2000) Voxel-based morphometry–the methods. *Neuroimage*, **11**, 805–821.

Bach,F. *et al*. (2004) Multiple Kernel Learning, Conic Duality, and the SMOAlgorithm. In *International Conference on Machine Learning (ICML)*, ACM, pp. 6.

Batmanghelich,N. *et al*. (2009) A general and unifying framework for feature construction, in image-based pattern classification. *Inf Process Med Imaging*, **21**, 423–434.

Beck,A. and Teboulle.,M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, **2**, 183–202.

Bickel,S. and Scheffer,T. (2004) Multi-view clustering. In *IEEE International Conference on Data Mining (ICDM)*, pp. 36.

Brefeld,U. and Scheffer,T. (2004) Co-em support vector learning. In *International Conference on Machine Learning (ICML)*, ACM, pp. 16.

Dhillon,I.S. *et al*. (2003) Information-theoretic co-clustering. In *ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) International Conference on Knowledge Discovery and Data Mining*, pp. 89–98.

Efron,B. *et al*. (2004) Least angle regression. *Ann. Stat.*, **32**, 407–499.

Fan,Y. *et al*. (2008) Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage*, **39**, 1731–1743.

Fischl,B. *et al*. (2002) Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, **33**, 341–355.

Ghani,R. (2002) Combining labeled and unlabeled data for multi-class text categorization. In *International Conference on Machine Learning*, ACM, pp. 187–194.

Hinrichs,C. *et al*. (2009a) MKL for robust multi-modality ad classification. In *Proceedings of the 12th International Conference on Medical Image Computing and Computer-Assisted Intervention: Part II*, pp. 786–794.

Hinrichs,C. *et al*. (2009b) Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. *Neuroimage*, **48**, 138–149.

Kim,S. and Xing,E. (2010) Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity. In *International Conference on Machine Learning (ICML)*. pp. 352–359.

Kloft,M. *et al*. (2008) Non-sparse multiple kernel learning. In *Proceedings of the NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*.

Krishnapuram,B. *et al*. (2005) Sparse multinomial logistic regression: fast algorithms and generalization bounds. In *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 957–968.

Lanckriet,G. *et al*. (2004) Learning the kernel matrix with semidefinite programming. In *JMLR*, **5**, 27–72.

Landau,S. *et al*. (2009) Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI. *Neurobiol. Aging*, **32**, 1207–1218.

Lee,S.-I. *et al*. (2006) Efficient l1 regularized logistic regression. In *The 21st National Conference on Artificial Intelligence (AAAI)*, pp. 401.

Liu,J. *et al*. (2009) Large-scale sparse logistic regression. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 547–556.

Luo,D. *et al*. (2010) Towards structural sparsity: an explicit l_2/l_0 approach. In *IEEE International Conference on Data Mining (ICDM)*, pp. 344–353.

Micchelli,C. *et al*. (2010) A family of penalty functions for structured sparsity. In *Advances in Neural Information Processing System (NIPS)*, The MIT Press, pp. 1612–1623.

Nigam,K. *et al*. (2000) Text classification from labeled and unlabeled documents using em. *Machine Learning*, **39**, 103–134.

Obozinski,G. *et al*. (2006) Multi-task feature selection. *Technical report*, Department of Statistics, University of California, Berkeley.

Obozinski,G. *et al*. (2010) Joint covariate selection and joint subspace selection for multiple classification problems. *Stat. Comput.*, **20**, 231–252.

Quattoni,A. *et al*. (2009) An efficient projection for $l_{1,\infty}$ regularization. In *International Conference on Machine Learning (ICML)*, ACM, pp. 857–864.

Rakotomamonjy,A. *et al*. (2007) More efficiency in multiple kernel learning. In *International Conference on Machine Learning (ICML)*, ACM, pp. 775–782.

Saykin,A.J. *et al*. (2010) Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans. *Alzheimers Dement*, **6**, 265–273.

Shen,L. *et al*. (2010a) Sparse bayesian learning for identifying imaging biomarkers in AD prediction. *Med. Image Comput. Comput. Assist. Interv.*, **13**(Pt 3), 611–618.

Shen,L. *et al*. (2010b) 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**, 1051–1063.

Sonnenburg,S. *et al*. (2006) Large scale multiple kernel learning. In *JMLR*, **7**, 1531–1565.

Stonnington,C.M. *et al*. (2010) Predicting clinical scores from magnetic resonance scans in alzheimer's disease. *Neuroimage*, **51**, 1405–1413.

Sun,L. *et al*. (2009) Efficient recovery of jointly sparse vectors. In *Advances in Neural Information Processing Systems (NIPS) 22*, The MIT Press, pp. 1812–1820.

Suykens,J. *et al*. (2002) *Least Squares Support Vector Machines*, World Scientific, Singapore. (ISBN 981-238-151-1)

Tibshirani,R. (1996) Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc B.*, **58**, 267–288.

Walhovd,K. *et al*. (2010) Multi-modal imaging predicts memory performance in normal aging and cognitive decline. *Neurobiol. Aging*, **31**, 1107–1121.

Wang,H. *et al*. (2011) Identifying AD-Sensitive and Cognition-Relevant Imaging Biomarkers via Joint Classification and Regression. In *The Proceedings of The 14th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2011), Lecture Notes in Computer Science (LNCS) 6893*, Springer, pp. 115–123.

Wang,H. *et al*. (2012) Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics*, **28**, 229–237.

Yang,X. *et al*. (2009) Heterogeneous multitask learning with joint sparsity constraints. In *Advances in Neural Information Processing System (NIPS)*, The MIT Press, pp. 2151–2159.

Ye,J. *et al*. (2008) Multi-class discriminant kernel learning via convex programming. In *JMLR*, **9**, 719–758.

Yu,S. *et al*. (2010). L 2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics*, **11**, 309.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B*, **68**, 49C–67.

Zien,A. and Ong,C. (2007) Multiclass multiple kernel learning. In *International Conference on Machine Learning (ICML)*, ACM, pp. 1191–1198.