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

Permalink
https://escholarship.org/uc/item/60g6t7q6

Journal
Bioinformatics (Oxford, England), 28(12)

ISSN
1367-4803

Authors
Wang, Hua
Nie, Feiping
Huang, Heng
et al.

Publication Date
2012-06-01

DOI
10.1093/bioinformatics/bts228

License
https://creativecommons.org/licenses/by/4.0/ 4.0

Peer reviewed
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 in 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/ h/wang/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) 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 markers 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 (Shawe, 2003; Brefeld and Schaeffer, 2004; Shani, 2006; Nigum, 2005),...
The structured sparsity is usually obtained through different sparse problems. From the view of sparsity organization, we have two by imposing non-smooth norms as regularizers in the optimization Kim and Xing, 2010; Micchelli and applied to multitask learning models (Argyriou related methods that mainly find the biomarkers correlated to each (Tibshirani, 1996), group LASSO (Yuan and Lin, 2006) and other biomarkers from multiple data sources. Different to LASSO 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 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 $\ell_1$-norm regularization (blue circles in Fig. 2) 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. 2) 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. Such multitask learning can discover robust patterns (because such 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 $w^i$ and $w^j$, respectively. The $\ell_{2,1}$-norm of the matrix $W$ is defined as $\|W\|_{2,1} = \sum_{j=1}^n \|w^j\|_2$ (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 $\{(x_i, y_i)\}_{i=1}^n$, where $x_i = (x_i^1, \ldots, x_i^q)^T \in \mathbb{R}^d$ is the input vector including all features from a total of $k$ different modalities and each modality $j$ has $d_j$ features ($d = \sum_{j=1}^k d_j$). $y_i \in \mathbb{R}$ is the class label vector of data point $x_i$ (only one element in $y_i$ is 1, and others are zeros), where $c$ is the number of classes (tasks). Let $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$ and $Y = [y_1, \ldots, y_n] \in \mathbb{R}^{c \times n}$. Different to MKL, we directly learn a $d \times c$ parameter matrix as:

$$W = \begin{bmatrix} w_1^1 & \cdots & w_1^k \\ \vdots & \ddots & \vdots \\ w_n^1 & \cdots & w_n^k \end{bmatrix} \in \mathbb{R}^{d \times c},$$

(1)

where $w^i_j \in \mathbb{R}^{d_j}$ 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:

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

(2)

where $\gamma$ 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 $\ell_1$-norm ($G_1$-norm) as a regularization term in Equation (4), which is defined over $W$ as following:

$$\|W\|_{G_1} = \sum_{i=1}^c \sum_{j=1}^k \|w^j_i\|_2,$$

(3)

which is illustrated by the blue circles in Figure 3. Then the Equation (4) becomes:

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

(4)

Since the group $\ell_1$-norm uses $\ell_2$-norm within each modality and $\ell_1$-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 $\ell_1$-norm regularizer captures the global relationships between data modalities.

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., 2014). A similar model was used in Wang et al. (2009) for...
We take the derivatives of Equation (8) with respect to random solution variables to be optimized. In this article, we provide an iterative algorithm to solve Equation (8). First, we guess a regression model to predict memory scores. Under the regression model, three diagnostic groups respectively (c1 = 3): AD, mild cognitive impairment (MCI), and control health (HC).

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

$$L_2(P) = \|X^T P - Z\|^2_F,$$

where $X$ is the data points matrix, $P$ is the coefficient matrix of regression with $c_2$ tasks, the label matrix $Z = \begin{bmatrix} (x^1)^T, (x^2)^T, \ldots, (x^n)^T \end{bmatrix} \in \mathbb{R}^{n \times c_2}$.

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

$$\min_{V} \sum_{i=1}^{n} \sum_{k=1}^{c_1} \left( \gamma_k \log \sum_{l=1}^{c_2} e^{w_k^T X_l} - \gamma_k w_k^T x_i \right)$$

+ $\|X^T P - Z\|^2_F$ \quad (8)

where $V = [W] \in \mathbb{R}^{d \times (c_1 + c_2)}$. 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 non-smooth 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 zeroes, we have

$$\frac{\partial C_1(W)}{\partial W} + 2\gamma_1 \sum_{i=1}^{n} D_i w_i + 2\gamma_2 D P W = 0,$$

$$2XX^T P - 2XXZ + 2\gamma_1 \sum_{i=1}^{n} D_i P_i + 2\gamma_2 D P P = 0,$$

where $D_i(1 \leq i \leq c_1 + c_2)$ is a block diagonal matrix with the $k$-th diagonal block as $\frac{1}{2} \mathbf{I}_{k}$ (if is a $d_k$ by $d_k$ identity matrix), $D$ is a diagonal matrix with the $k$-th diagonal element as $\frac{1}{2}$. Since $D_i(1 \leq i \leq c_1 + c_2)$ 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 $V = [W P]$; then we calculate the matrices $D_i(1 \leq i \leq c_1 + c_2)$ and $D$ according to the current solution $V$. After obtaining the $D_i(1 \leq i \leq c_1 + c_2)$ and $D$, we can update the solution $V = [W P]$ based on Equation (8). Specifically, the $j$-th column of $P$ is updated by $p_j = (X^T + \gamma_1 D_1 + \gamma_2 D)^{-1} X z_j$. We cannot update $W$ with a closed form solution based on Equation (8), but we can obtain the updated $W$ by the Newton’s method. According to Equation (8), we need to solve the following problem:

$$\min_{W} \|L_1(W) + \sum_{i=1}^{n} w_i^T D_i w_i + 2\gamma_2 D (W^T D W)\|_F,$$ 

where $L_1(W) = \sum_{i=1}^{n} \sum_{k=1}^{c_1} \left( \gamma_k \log \sum_{l=1}^{c_2} e^{w_k^T X_l} - \gamma_k w_k^T x_i \right)$.

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. 2004).

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

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

$$\frac{\partial}{\partial W_{up}} 2\|W^T D W\|_F = 2D_{\alpha(u, w)}W_{up},$$

where $D_{\alpha(u, w)}$ is the $u$-th diagonal element of $D_{\alpha}$.

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

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

$$\frac{\partial}{\partial W_{up}} 2\|W^T D W\|_F = 2D_{\alpha(u, w)}W_{up},$$

After obtaining the updated solution $V = [W P]$, we can calculate the new matrices $D_i(1 \leq i \leq c_1 + c_2)$ 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 $v$ and $w$, we have the following inequality:

$$\|v\|^2_2 - \frac{\|v\|^2_2}{\|w\|^2_1} \leq \|v\|^2_2 - \frac{\|v\|^2_2}{\|w\|^2_1}.$$  

**Proof.** Obviously, $-(\|v\|^2_2 - \|w\|^2_1)^2 \leq 0$, so we have

$$-(\|v\|^2_2 - \|w\|^2_1)^2 \leq 0 \Rightarrow 2\|v\|^2_2 \|w\|^2_1 - \|v\|^2_2 \leq \|w\|^2_1$$

$$\Rightarrow \|v\|^2_2 - \frac{\|v\|^2_2}{\|w\|^2_1} \leq \|w\|^2_1 - \frac{\|w\|^2_1}{\|w\|^2_1},$$

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 \( W_i \), and the updated \( P \) is \( P_i \), then the updated \( V \) is \( \bar{V} \). From Step 3 in the Algorithm 1, we know that:

\[
\mathcal{L}_1(W) + \gamma_2 \mathcal{T}(\bar{W}^T D \bar{W}) \leq \mathcal{L}_1(W) + \gamma_2 \mathcal{T}(P^T D P). 
\]

According to Step 4, we have:

\[
\left| X^T P - Y \right|^2_p + \gamma_1 \sum_{i=1}^{c_1} \mathcal{H}^i \mathcal{D} \mathcal{P} + \gamma_2 \mathcal{T}(P^T D P) \leq \left| X^T P - Y \right|^2_p + \gamma_1 \sum_{i=1}^{c_1} \mathcal{H}^i \mathcal{D} \mathcal{P} + \gamma_2 \mathcal{T}(P^T D P). 
\]

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

\[
\sum_{k=1}^{c_1} \sum_{i=1}^{c_1} \mathcal{H}^i \mathcal{D} \mathcal{P} \leq \sum_{k=1}^{c_1} \sum_{i=1}^{c_1} \mathcal{H}^i \mathcal{D} \mathcal{P} 
\]

\[
\sum_{i=1}^{c_1} \mathcal{P}^2 \mathcal{D}_{i} \mathcal{P} \leq \sum_{i=1}^{c_1} \mathcal{P}^2 \mathcal{D}_{i} \mathcal{P} 
\]

\[
\mathcal{D}_{i} \mathcal{P} \mathcal{D}_{i} \leq \mathcal{D}_{i} \mathcal{P} \mathcal{D}_{i} 
\]

\[
\left| X^T P - Y \right|^2_p + \gamma_1 \sum_{i=1}^{c_1} \mathcal{P}^2 \mathcal{D}_{i} \mathcal{P} + \gamma_2 \mathcal{T}(P^T D P) \leq \left| X^T P - Y \right|^2_p + \gamma_1 \sum_{i=1}^{c_1} \mathcal{P}^2 \mathcal{D}_{i} \mathcal{P} + \gamma_2 \mathcal{T}(P^T D P). 
\]

Algorithm 1 An efficient iterative algorithm to solve the optimization problem in Equation (9).

Input: \( X = [x_{1}, x_{2}, \ldots, x_{n}] \in \mathbb{R}^{d \times c_1} \), \( Y = \left[ (y_{1})^T, (y_{2})^T, \ldots, (y_{n})^T \right] \in \mathbb{R}^{(1 \times c_1)} \), and \( Z = \left[ (z_{1})^T, (z_{2})^T, \ldots, (z_{n})^T \right] \in \mathbb{R}^{(1 \times c_2)}. \)

Output: \( W \in \mathbb{R}^{d \times c_1} \) and \( P \in \mathbb{R}^{d \times c_2}. \) Let \( V = \left[ W \right. \left. P \right] \in \mathbb{R}^{d \times (c_1+c_2)}. \)

1. Initialize \( W \in \mathbb{R}^{d \times c_1}, P \in \mathbb{R}^{d \times c_2}. \) Let \( V = \left[ W \right. \left. P \right] \in \mathbb{R}^{d \times (c_1+c_2)}. \)

2. Calculate the block diagonal matrices \( D_j (1 \leq j \leq c_1 + c_2) \), where the \( k \)-th diagonal block of \( D_j \) is \( \mathcal{P}^2 \mathcal{D}_{i} \mathcal{P} \). Calculate the diagonal matrix \( D_i \), where the \( k \)-th diagonal element is \( \frac{1}{\mathcal{P}} \).

3. Update \( w \) by \( w = \mathcal{B}^{-1} u \), where the \( d \cdot (p-1) + u \) (1 \( \leq u \leq d, 1 \leq p \leq c_1 \)) element of \( a \in \mathbb{R}^{d \times 1} \) is

\[
\frac{\mathcal{L}_1(W) + \gamma_1 \sum_{i=1}^{c_1} \mathcal{P}^2 \mathcal{D}_{i} \mathcal{P}}{\mathcal{L}_1(W) + \gamma_2 \mathcal{T}(P^T D P)}. 
\]

4. Update the \( i \)-th column of \( P \) by \( p_i = (XX^T + \gamma_1 D_i + \gamma_2 D_i^{-1} X \bar{X}) \).

5. Update the \( V \) by \( V = \left[ W \right. \left. P \right]. \)

until Converges

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

\[
\mathcal{L}_1(W) + \mathcal{L}_2(P) + \gamma_2 \sum_{i=1}^{c_1+c_2} \mathcal{H}^i \mathcal{D} \mathcal{P} \leq \mathcal{L}_1(W) + \mathcal{L}_2(P) + \gamma_2 \sum_{i=1}^{c_1+c_2} \mathcal{H}^i \mathcal{D} \mathcal{P} \]

Therefore, the algorithm decreases the objective value of problem (9) in each iteration. \( \square \)

In the convergence, \( W, P, D_j (1 \leq j \leq c_1 + c_2) \) and \( D \) satisfy the Equation (9). As the Equation (9) is a convex problem, satisfying the Equation (9) indicates that \( V = \left[ W \right. \left. P \right] \) is a global optimum solution to the Equation (9). Therefore, the Algorithm 1 will converge to the global optimum of the Equation (9). Since our algorithm has the closed form solution in each iteration, the convergence 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 (ADNI1, ADNI2, ADNI-Late). 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](http://www.adni-info.org). Following a prior imaging genetics study [Shen et al., 2010a], 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. 4). We describe our genotyping, imaging and memory data in Section 3.1.1 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., 2010a]. First, voxel-based morphometry (VBM) [Ashburner and Friston, 2000] was performed to define global gray matter (GM) density maps and extract local GM morphometry (VBM) (Ashburner and Friston, 2000) was performed as previously described (Shen et al., 2010). All these measures are available in [Shen et al., 2010a]. For the remaining 3 genes, no SNPs were available on the genotyping chip.

**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 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, without reading it again. Trial 7, termed as 30 min recall, requests the participant again to recall as many words as possible from List A, without reading it again. Following Landau et al. [2000], mean glucose metabolism (CMGlu) measures of 26 regions of interest (ROIs) in the Montreal Neurological Institute (MNI) brain space were employed in this study (Fig. 4).

### Table 1. RAVLT cognitive measures as responses in multitask learning

<table>
<thead>
<tr>
<th>Task ID</th>
<th>Description of RAVLT scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>Total score of the first 5 learning trials</td>
</tr>
<tr>
<td>TOT6</td>
<td>Trial 6 total number of words recalled</td>
</tr>
<tr>
<td>TOTB</td>
<td>List B total number of words recalled</td>
</tr>
<tr>
<td>T30</td>
<td>30 minute delay total number of words recalled</td>
</tr>
<tr>
<td>RECOG</td>
<td>30 minute delay recognition score</td>
</tr>
</tbody>
</table>

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’.

[www.adni-info.org](http://www.adni-info.org)
Table 2. Multimodal feature sets as predictors in multiview learning

<table>
<thead>
<tr>
<th>View ID (feature set ID)</th>
<th>Modality</th>
<th>No. of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>VBM</td>
<td>MRI</td>
<td>86</td>
</tr>
<tr>
<td>FreeSurfer</td>
<td>MRI</td>
<td>56</td>
</tr>
<tr>
<td>FDG-PET</td>
<td>FDG-PET</td>
<td>26</td>
</tr>
<tr>
<td>SNP</td>
<td>Genetics</td>
<td>1244</td>
</tr>
</tbody>
</table>

**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-coded 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 4b is a comparison between RAVLT-relevant markers and AD-relevant markers and their associated weights mapped onto a standard brain space.

Figure 4c 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 findings (Fig 2c) 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 RAVLT 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.](image)

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, Pencina et al. (2010), SVM $ℓ_1$ MKL method, De Vito et al. (2009), least square (LSSVM) $ℓ_∞$ MKL method, Yang et al. (2008), LSSVM $ℓ_1$ MKL method, Sykora et al. (2009) 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 $γ_i$ and $γ_j$ in Equation (8) are optimized in the range of $10^{-7}, 10^{-4}, 10^{-1}, 10^1, 10^5$. For SVM method and MKL methods, one Gaussian kernel is constructed for each type of features $e.g.\mathcal{K}(x, x_i) = \exp(-γ \|x - x_i\|^2_2)$, 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,
Copyedited by: TRJ

MANUSCRIPT CATEGORY:

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy (mean ± SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (SNP)</td>
<td>0.561 ± 0.026</td>
</tr>
<tr>
<td>SVM (FreeSurfer)</td>
<td>0.573 ± 0.012</td>
</tr>
<tr>
<td>SVM (VBM)</td>
<td>0.541 ± 0.032</td>
</tr>
<tr>
<td>SVM (PET)</td>
<td>0.535 ± 0.026</td>
</tr>
<tr>
<td>SVM (all)</td>
<td>0.575 ± 0.019</td>
</tr>
<tr>
<td>HML (all)</td>
<td>0.638 ± 0.019</td>
</tr>
<tr>
<td>SVM, MKL method</td>
<td>0.624 ± 0.031</td>
</tr>
<tr>
<td>SVM, MKL method</td>
<td>0.593 ± 0.042</td>
</tr>
<tr>
<td>SVM, MKL method</td>
<td>0.561 ± 0.037</td>
</tr>
<tr>
<td>LSSVM, MKL method</td>
<td>0.614 ± 0.031</td>
</tr>
<tr>
<td>LSSVM, MKL method</td>
<td>0.585 ± 0.018</td>
</tr>
<tr>
<td>LSSVM, MKL method</td>
<td>0.577 ± 0.033</td>
</tr>
<tr>
<td>Our method (SNP)</td>
<td>0.673 ± 0.021</td>
</tr>
<tr>
<td>Our method (FreeSurfer)</td>
<td>0.689 ± 0.029</td>
</tr>
<tr>
<td>Our method (VBM)</td>
<td>0.669 ± 0.031</td>
</tr>
<tr>
<td>Our method (PET)</td>
<td>0.621 ± 0.028</td>
</tr>
<tr>
<td>Our method</td>
<td>0.726 ± 0.032</td>
</tr>
</tbody>
</table>

The methods using multiple data sources are generally better than their counterparts using only 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 they 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 prediction performance

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 3.3, 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 $10^{-5}, 10^{-4}, \ldots, 10^{5}, 10^{6}$ 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 is 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 convergence. The empirical results showed our method improved both memory scores prediction and disease classification accuracy.

ACKNOWLEDGEMENT

Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.
Funding: [This research was supported by National Science Foundation Grants CCF-0830780, CCF-0917274, DMS-0915228, and IIS-1117965] at UTA; and by [National Science Foundation Grant IIS-1117735, National Institutes of Health Grants UL1 RR025761, U01 AG024904, NIA RC2 AG036353, NIA R01 AG19771, and NIA P30 AG10133-18S1] at IU.

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Amorflex Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Biotronik AG; Biogen Idec Inc.; Eli Lilly and Company; F Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Ingenomics; N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Medstat Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH [P30 AG010129, K01 AG030514] and the Dana Foundation.

Conflict of Interest: none declared.

REFERENCES


Dhillon,I.S. et al. (2001) Information-theoretic clustering. In ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) International Conference on Knowledge Discovery and Data Mining, pp. 89–98.


H. Wang et al.


