Multi-Source Learning for Joint Analysis of Incomplete Multi-Modality Neuroimaging Data

Lei Yuan[1,2], Yalin Wang[1], Paul M. Thompson[3], Vaibhav A. Narayan[4], Jieping Ye[1,2]
[1]Center for Evolutionary Medicine and Informatics, The Biodesign Institute, ASU, Tempe, AZ
[2]Department of Computer Science and Engineering, ASU, Tempe, AZ
[3]Department of Neurology, UCLA, Los Angeles, CA
[4]Johnson & Johnson Pharmaceutical Research & Development, LLC, Titusville, NJ

ABSTRACT

Incomplete data present serious problems when integrating large-scale brain imaging data sets from different imaging modalities. In the Alzheimer's Disease Neuroimaging Initiative (ADNI), for example, over half of the subjects lack cerebrospinal fluid (CSF) measurements; an independent half of the subjects do not have fluorodeoxyglucose positron emission tomography (FDG-PET) scans; many lack proteomics measurements. Traditionally, subjects with missing measures are discarded, resulting in a severe loss of available information. We address this problem by proposing two novel learning methods where all the samples (with at least one available data source) can be used. In the first method, we divide our samples according to the availability of data sources, and we learn shared sets of features with state-of-the-art sparse learning methods. Our second method learns a base classifier for each data source independently, based on which we represent each source using a single column of prediction scores; we then estimate the missing prediction scores, which, combined with the existing prediction scores, are used to build a multi-source fusion model. To illustrate the proposed approaches, we classify patients from the ADNI study into groups with Alzheimer's disease (AD), mild cognitive impairment (MCI) and normal controls, based on the multi-modality data. At baseline, ADNI's 780 participants (172 AD, 397 MCI, 211 Normal), have at least one of four data types: magnetic resonance imaging (MRI), FDG-PET, CSF and proteomics. These data are used to test our algorithms. Comprehensive experiments show that our proposed methods yield stable and promising results.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - Data Mining

General Terms

Algorithms

Keywords

Multi-source feature learning, multi-task learning, incomplete data

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
KDD'12, August 12–16, 2012, Beijing, China.
Copyright 2012 ACM 978-1-4503-1462-6 /12/08 ...$10.00.

1. INTRODUCTION

Alzheimer's disease (AD) is a highly prevalent neurodegenerative disease, and is widely recognized as a major, escalating epidemic and a world-wide challenge to global health care systems [22]. AD is the the most common type of dementia, accounting for 60-80% of age-related dementia cases. The direct cost of care for AD patients by family members and healthcare professionals is more than $100 billion per year; this figure is expected to rise dramatically as the population ages over the next several decades [33]. In AD patients, neurons and their connections are progressively destroyed, leading to loss of cognitive function and ultimately death. The underlying pathology is thought to precede the onset of cognitive symptoms by many years [3, 19]. Efforts are underway to find early diagnostic markers to evaluate AD risk pre-symptomatically in a rapid and rigorous way. Such findings will help establish early interventions to prevent or at least postpone the onset of AD, or reduce the risk of developing the disease.

Neuroimaging is a powerful tool to measure disease progression and therapeutic efficacy in AD and mild cognitive impairment (MCI). Neuroimaging research offers great potential to discover features that can identify individuals early in the course of dementing illness; several candidate neuroimaging biomarkers have been examined in recent cross-sectional and longitudinal neuroimaging studies [9, 12]. Reduced fluorodeoxyglucose (FDG) PET measurements of the cerebral metabolic rate for glucose in brain regions preferentially affected by AD, structural MRI measures of brain shrinkage, and cerebrospinal fluid (CSF) measurements are among the best established biomarkers of AD progression and pathology [33]. Realizing the importance of combining neuroimaging and genetics, NIH in 2003 funded the Alzheimer's Disease Neuroimaging Initiative (ADNI [29, 20], PI: Michael W. Weiner). The initiative is facilitating the scientific evaluation of neuroimaging data including magnetic resonance imaging (MRI), positron emission tomography (PET), other biomarkers, and clinical and neuropsychological assessments for predicting the onset and progression of MCI and AD. By identifying more sensitive and specific markers of very early AD progression, these efforts should make it easier to diagnose AD earlier as well as develop, assess, and monitor new treatments.

Clinical and research studies commonly acquire complementary brain images, neuropsychological and genetic data for each participant for a more accurate and rigorous assessment of the disease status and likelihood of progression. Advances in image analysis make it possible to use one image modality to support the analysis of a complementary image modality [17, 6, 18, 23]. However, only a few systems, e.g., [40, 28, 11, 41, 38, 21, 39, 24, 43], applied data mining and machine learning techniques such as the multivariate linear model and partial least squares to characterize the

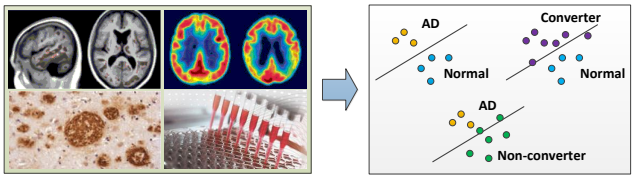

Figure 1: Illustration of integrating multiple heterogeneous data sources for disease status prediction tasks. More details on the different data sources and prediction tasks used in this study may be found in Section 3.

linkage between the patterns of information from the same individual's brain images and other biological measures. Instead, most researchers perform statistical analysis by analyzing different images separately. In general, these "unimodal" analysis could be improved by considering other sources of relevant information from multiple imaging modalities, e.g., PET and MRI, and non-imaging data sets from genomics and proteomics. It is a common belief that by integrating multiple heterogeneous sources (as illustrated in Figure 1), one may not only provide more accurate information on AD progression and pathology, but also better predict cognitive decline before the onset of illness, or at least in the earliest stages of disease.

One common problem that hampers the use of multi-modality imaging approach is the problem of *missing data*. Missing data present a special challenge when integrating large-scale biomedical data. Incomplete data is ubiquitous in real-world biomedical applications. In ADNI, over half of the subjects lack CSF measurements; an independent half of the subjects do not have FDG-PET; many lack proteomics measurements. Missing data may be due to the high cost of certain measures (e.g., PET scans), poor data quality, dropout of the patients from the study, etc. Some measures, such as CSF biomarkers, require more invasive procedures (such as lumbar puncture) which not all study participants are willing to consent to. Some subjects in a longitudinal study may miss at least one of the regular assessments, or their data quality may be insufficient for accurate analysis at some time points. The simplest approach removes all samples with missing values, but this throws away a vast amount of useful information and dramatically reduces the number of samples in the analysis. As a result, a subject with incomplete data cannot be studied for classification and prognosis. Moreover, with this approach, the resource and time devoted to those subjects with incomplete data are totally wasted. A number of previous works acknowledged the challenge of missing data and discussed general strategies [37, 14, 32]. An alternative and popular approach is to estimate missing entries based on the observed values. Many algorithms have been proposed for this [15, 34, 13, 35]. While these methods work well when missing data are rare, they are less effective when a significant amount of data is missing, e.g., when all PET features from half of the subjects are missing. Recently, trace norm minimization has been proposed for missing data estimation [4, 5]. This can be effective even when a large amount of data is missing. However, it does assume that the missing locations are random; it is less effective when a complete block of the data is missing, e.g., the complete block of all PET features from half of the subjects. Therefore, computational methods are needed to integrate heterogeneous data with a block-wise missing pattern ("block-wise missing" means a large chunk of data is missing for one or more data sources - an example is shown in Figure 2). Without such a method, it is quite challenging to build a highly accurate classifier to process any real multi-modality imaging data sets.

To achieve multi-modality integration while taking into account the block-wise missing nature of the data, we propose two novel learning frameworks. Our first system is based on a novel multi-task sparse learning framework, named incomplete multi-source feature learning (iMSF). Based on the availability of different feature sources, we divide the data set into several learning tasks, from each of which a unique classifier is learned. We then impose a structural sparse learning regularization onto these tasks, such that a common set of features is selected among these tasks. Therefore, we exploit the multi-task nature of the problem and the feature set is learned jointly among different tasks. In our second system, we tackle the difficulty of block-wise data completion by proposing a model score completion scheme (ScoreComp). Each data source is first treated independently, where a base classifier is learned such that the data source is converted into a single column consisting of model prediction scores. Thus, the missing data blocks become single missing values in the new representation of the data. Data imputation techniques are then applied and a new classifier is learned on the completed score matrix such that multiple sources are integrated together.

As an illustrative application, we study clinical group (diagnostic) classification problems in the ADNI baseline data set. Comprehensive experiments demonstrate the promising and stable performance of the proposed systems. 780 subjects in the ADNI baseline data set have their diagnosis (AD, MCI or Normal) available and have at least one type of features available (meaning an image or related clinical measure), including MRI, FDG-PET, CSF and proteomics. For the MCI subjects, we further make use of their 4-year follow up diagnosis to divide them into 2 sub-groups. We label those who had converted to AD by the time of a later visit as "converter" and those who stayed stable as "non-converter". We set out to use these data to solve clinical group classification problems (AD vs. Normal; AD vs. Non-converter and Converter vs. Normal). For our experiments, we obtained MRI, CSF and proteomics feature sets from the ADNI web site (`http://adni.loni.ucla.edu/`) and we processed FDG-PET data using the image analysis package, SPM (SPM8, `http://www.fil.ion.ucl.ac.uk/spm`) using the statistical region of interest (sROI) method. Besides our multi-source learning frameworks for incomplete data, we also include five other methods to estimate missing values: (1) the "Zero" method: a method that fills missing values with zeros; (2) EM: a missing value imputation method based on the expectation-maximization (EM) algorithm [34]; (3) SVD (singular value decomposition): a method for matrix completion using a low rank approximation to the full matrix; (4) KNN: a missing value imputation method based on the k-nearest neighbor principle [15]; and (5) SVT: a matrix completion method based on trace norm minimization [4, 5]. The experimental results show that our proposed methods are effective for incomplete multi-source data fusion.

The rest of the paper is organized as follows: in Section 2, we present two multi-source learning methods for a joint analysis of multi-modality data with a "block-wise" missing pattern; we have performed comprehensive experiments to evaluate the proposed methods and the results are reported in Section 3; finally, we conclude our paper in Section 4.

2. PROPOSED METHODS

In many applications, multiple data sources may suffer from a considerable amount of missing data. For example, in the ADNI data acquisition phase, many subjects lack a subset of measures, resulting in a scenario shown in Figure 2, where large chunks of missing data are marked by the white areas. A simple and popular

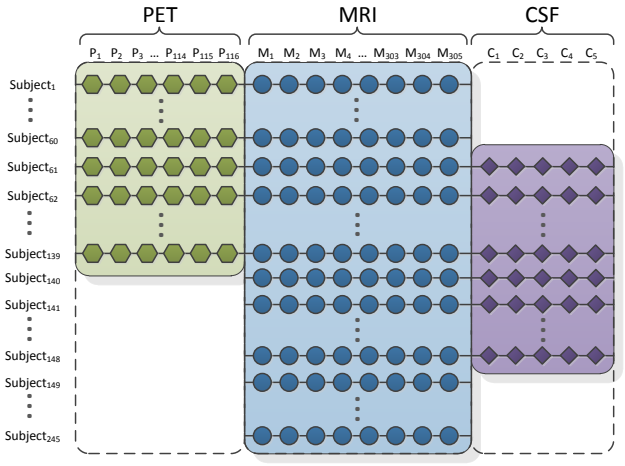

Figure 2: Illustration of the "block-wise" pattern of missing data for the ADNI data set. In this figure, we show AD and normal control subjects only. For simplicity, we focus on those subjects with complete MRI measures.

approach is to remove all the subjects with missing values, but this greatly reduces the number of samples and fails to fully use the information in the data set. In Figure 2, only 79 subjects (Subjects 61-139) out of a total of 245 subjects do not have missing values. Next, we present two methods for dealing with multi-source data with block-wise missing values.

2.1 Proposed Method I: Incomplete Multi-Source Feature Learning

In our feature learning framework described below, we fully use the multiple heterogeneous data with a block-wise missing pattern by exploiting the underlying structure in the multi-source data. Our proposed framework formulates the prediction problem as a multi-task learning problem [1, 2, 26, 44] by first decomposing the prediction problem into a set of tasks, one for each combination of data sources available, and then building the models for all tasks simultaneously.

For example, considering a data set with three sources (CSF, MRI, PET) and assuming all samples have MRI measures, we first partition the samples into multiple blocks (4 in this case), one for each combination of data sources available: (1) PET, MRI; (2) PET, MRI, CSF; (3) MRI, CSF; and (4) MRI. We then build four models, one for each block of data, resulting in four prediction tasks (Figure 3).

A simple approach to deal with the missing data is to build these four models separately, but that does not fully use the information in the multi-source data. Indeed, the sample size for each of these four tasks is even smaller, resulting in the *large dimension small sample size* problem. We address this by employing a joint feature learning formulation. We formulate our proposed framework as follows. Suppose the data set is divided into m tasks: $T^i = \{x_j^i, y_j^i\}, i = 1 \dots m, j = 1 \dots N_i$, where N_i is the number of subjects in the i-th task, and (x_j^i, y_j^i) is the j-th subject from the i-th task. For each task, we consider the following linear model:

$$f^i(x_j) = (\beta^i)^T x_j^i$$

where β^i is the weight vector, including the model parameters for the i-th task. Denote $\beta = \{\beta^1, \dots, \beta^m\}$ as the collection of all model parameters. Assume that we have a total of S data sources, and the feature dimensionality of the s-th source is denoted as p_s. For notational convenience, we introduce an index function $I(s,k)$ as follows: $\beta_{I(s,k)}$ denotes all the model parameters corresponding to the k-th feature in the s-th data source. The proposed multi-task feature learning framework is:

$$\min_{\beta} \frac{1}{m} \sum_{i=1}^{m} \frac{1}{N_i} \sum_{j=1}^{N_i} L(x_j^i, y_j^i, \beta_i) + \lambda \sum_{s=1}^{S} \sum_{k=1}^{p_s} \left\| \beta_{I(s,k)} \right\|_2 \quad (1)$$

where $L(\cdot)$ is the loss function, and we adopt the logistic loss in our study. The second part of the formulation, which is essentially an $\ell_{2,1}$-norm regularization on the model parameters [42], leads to a solution with the desired sparsity, that is, all models involving a specific source are constrained to select a common set of features for this particular source. The proposed formulation is novel as it (1) formulates the incomplete multi-source fusion as a multi-task learning problem, and (2) extends existing multi-task feature learning formulations to accommodate missing feature values.

The regularization parameter λ in (1) controls the sparsity of the solution. Generally speaking, the larger λ is, the sparser the solution will be. However, in practice, the same λ value will induce different sparsity for different data sets. To select a proper range of parameters, we follow a similar approach as discussed by Liu et al. [27] to obtain a value λ_{max} for each specific problem such that if $\lambda \geq \lambda_{max}$, the optimal solution to (1) is $\mathbf{0}$. Therefore, we just need to set a *ratio* r such that $\lambda = r\lambda_{max}$, and r is selected in the region $(0, 1)$.

2.1.1 *Efficient Optimization*

The optimization problem proposed in Section 2.2 is the composition of a smooth term and a non-smooth term, which is challenging to solve. In this paper, we propose to solve it using the accelerated gradient descent (AGD) method [31, 30] as in [26] because of its fast convergence rate. Denote the empirical loss as:

$$\ell(\beta) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{N_i} \sum_{j=1}^{N_i} L(x_j^i, y_j^i, \beta_i),$$

and the non-smooth regularization as:

$$\phi_\lambda(\beta) = \lambda \sum_{s=1}^{S} \sum_{k=1}^{p_s} \left\| \beta_{I(s,k)} \right\|_2 .$$

In AGD, we first approximate $\ell(\beta) + \phi_\lambda(\beta)$ by:

$$f_{L,\beta}(\theta) = \ell(\beta) + \langle \ell'(\beta), \theta - \beta \rangle + \theta_\lambda(\theta) + \frac{L}{2} \|\theta - \beta\|^2.$$

In the i-th step, a search point s_i is computed based on the past solutions of the previous steps by $s_i = \beta_i + \tau_i(\beta_i - \beta_{i-1})$. Then, the new solution β_{i+1} is obtained via the minimization of the model at the current search point, that is, $\beta_{i+1} = \arg\min_\theta f_{L,s_i}(\theta)$. This sub-problem is the key component to the optimization, and is often called the proximal operator [7]. A detailed discussion of how to solve this sub-problem efficiently is found in our previous work [26]. By doing so, we successfully bypass the difficulty of computing the sub-gradient of $\phi_\lambda(\cdot)$; algorithm details are summarized in Algorithm 1.

2.2 Proposed Method II: Model Score Completion

The iMSF framework proposed in Section 2.1 tackles the problem in a "row-wise" manner by dividing samples into different

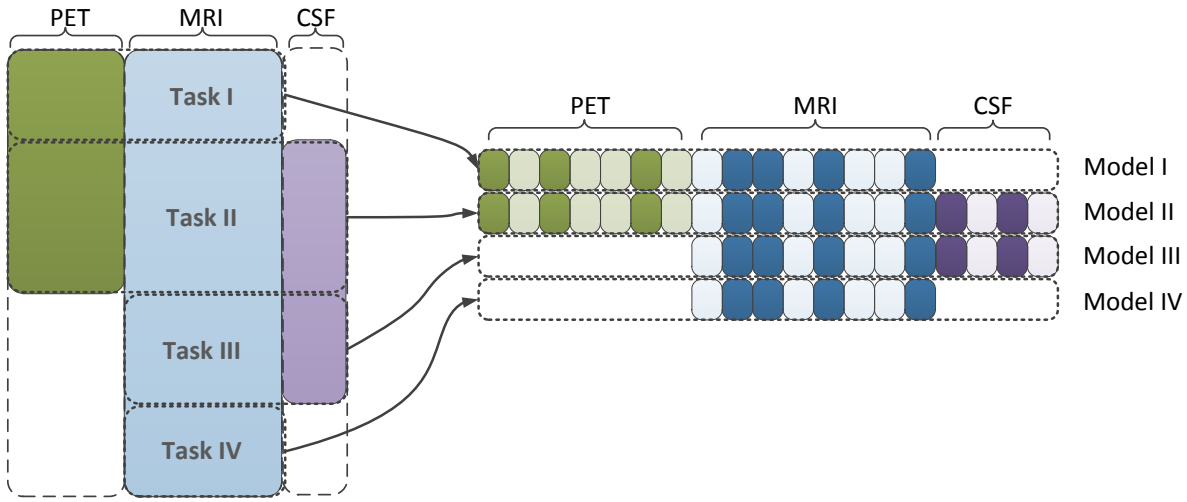

Figure 3: Illustration of the proposed multi-task feature learning framework for incomplete multi-source data fusion. In the proposed framework, we first partition the samples into multiple blocks (four blocks in this case), one for each combination of data sources available: (1) PET, MRI; (2) PET, MRI, CSF; (3) MRI, CSF; (4) MRI. We then build four models, one for each block of data, resulting in four prediction tasks. We use a joint feature learning framework that learns all models simultaneously. Specifically, all models involving a specific source are constrained to select a common set of features for that particular source.

Algorithm 1 Efficient Optimization for the Multi-Source Feature Learning Framework

Input: L_0, λ, β_0, n
Output: β_{n+1}
Initialize $\beta_1 = \beta_0$, $\alpha_{-1} = 0$, $\alpha_0 = 1$, and $L = L_0$
for $i = 1$ to n **do**
Set $\tau_i = \frac{\alpha_{i-2}-1}{\alpha_{i-1}}$, $s_i = \beta_i + \tau_i(\beta_i - \beta_{i-1})$
Find the smallest $L = L_{i-1}, 2L_{i-1}, \ldots$ such that $\ell(\beta_{i+1}) + \phi_\lambda(\beta_{i+1}) \leq f_{L,s_i}(\beta_{i+1})$ holds, where

$$\beta_{i+1} = \arg\min_\theta f_{L,s_i}(\theta)$$

Set $L_i = L$ and $\alpha_{i+1} = \frac{1+\sqrt{1+4\alpha_i^2}}{2}$
end for

groups. We next propose to tackle the problem in a "column-wise" manner. From Figure 2 we can observe another characteristic of this problem: if a certain data source is available for a particular sample (e.g., PET for subject 1), the complete set of features from this data source will be available. Intuitively, it is more challenging to estimate a complete block of missing values than a single value. This motivates us to design the model score completion scheme (ScoreComp). The overview of the ScoreComp is illustrated in Figure 4. Given an incomplete multi-source data set, we first train a base model on each individual data source using the available samples, and the base model is applied to produce prediction scores for the corresponding samples for this data source; thus each data source is represented by a single column of scores, and all data sources together are represented as a matrix of prediction scores with missing values. A missing value estimation method is then applied to obtain a complete set of model scores, which are treated as newly derived features to train our final classifier. That is, the prediction score from each data source is considered as a feature.

We formally describe our ScoreComp method as follows. Consider a labeled data set $\{\mathcal{D}_i, \mathcal{Y}_i\}$, $i = 1, \ldots, N$, with S incomplete

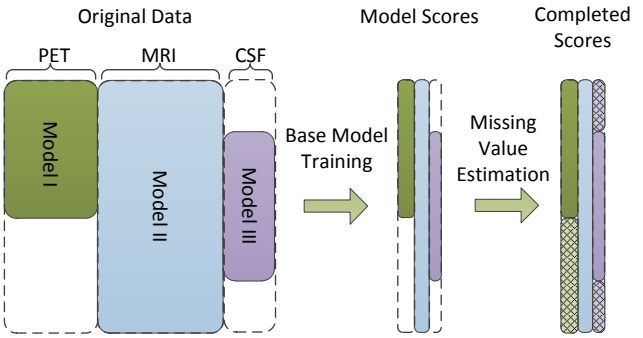

Figure 4: Illustration of the proposed model score completion scheme. We first train a base model on each individual data source using the available samples, and the base model is applied to produce prediction scores for this data source; thus each data source is represented by a single column of (incomplete) scores. A missing value estimation method is applied to obtain a complete set of model scores, which are treated as newly derived features to train our final classifier.

data sources. Let the set $sc(i) \subset \{1, \ldots, S\}$ denote the available data sources for the i^{th} subject, such that $\mathcal{D}_i = \{X_i^s | s \in sc(i)\}$, where X_i^s is the feature vector of the i^{th} subject from the s^{th} source. The goal of the ScoreComp method is to derive a completed prediction score matrix $\tilde{A} \in \mathbb{R}^{N \times S}$ from the original data set. Details are given below:

Base Model Training Step. We first choose a classifier learning algorithm $\mathcal{L}$, based on which a prediction model is constructed for each data source:

$$\mathcal{M}_s = \mathcal{L}\left(\{(X_i^s, \mathcal{Y}_i) | s \in sc(i)\}\right), \quad s = 1, \ldots S.$$

Then, we use these models to construct an incomplete prediction

score matrix $\hat{A} \in \mathbb{R}^{N \times S}$ given by:

$$\hat{A}_{i,s} = \begin{cases} \mathcal{M}_s(X_i^s) & \text{if } s \in sc(i) \\ \text{NaN} & \text{otherwise} \end{cases},$$

where $\mathcal{M}_s(X_i^s)$ is the prediction score of model $\mathcal{M}_s$ on feature vector X_i^s.

Missing Value Estimation Step. In this step, we choose a missing value estimation algorithm $\mathcal{E}$ such that $\tilde{A} = \mathcal{E}(\hat{A})$, where $\tilde{A}$ is the completed prediction score matrix. $\tilde{A}$ is then treated as the derived feature matrix for the original data set $\{\mathcal{D}_i, \mathcal{Y}_i\}$. The final model $\mathcal{M}$ is learned using $(\tilde{A}, \mathcal{Y})$ so that the data sources are integrated.

Prediction of Unlabeled Sample. Suppose we are given a set of unlabeled data $\{\mathcal{U}_j\}$, $j = 1, \ldots, M$, such that $\mathcal{U}_j = \{X_j^s | s \in sc(j)\}$. ScoreComp will first derive a completed feature matrix $\tilde{B} \in \mathbb{R}^{M \times S}$, which is then fed into $\mathcal{M}$ for prediction. To obtain $\tilde{B}$, we first use the models learned in the base model training step to construct an incomplete model score matrix $\hat{B} \in \mathbb{R}^{M \times S}$ given by:

$$\hat{B}_{j,s} = \begin{cases} \mathcal{M}_s(X_j^s) & \text{if } s \in sc(j) \\ \text{NaN} & \text{otherwise} \end{cases}.$$

Combining this with the previously obtained complete matrix $\tilde{A}$, we obtain:

$$C = \mathcal{E}\left(\begin{bmatrix} \tilde{A} \\ \hat{B} \end{bmatrix}\right), \quad C \in \mathbb{R}^{(N+M) \times S}.$$

Finally, by extracting the lower M rows of matrix C, we can obtain the derived feature matrix $\tilde{B}$ for the unlabeled data set.

Like the iMSF method, all available information is used in the integration process. Note that in ScoreComp, we still estimate missing values; but instead of estimating blocks of missing data, we only need to impute the prediction scores. Another advantage of this framework is its simplicity. No additional parameters are introduced, and one can choose any classification algorithms and/or missing value estimation method that suit the data set at hand. In this study, we chose the random forest classifier [25] as the base model learning algorithm $\mathcal{L}$, and our final model $\mathcal{M}$ was trained using ridge regression. We also tried various missing data estimation methods, and more details may be found in Section 3.

3. EXPERIMENTAL RESULTS

In this section, we report the results of our experiments to evaluate the effectiveness of our proposed methods. As noted earlier, we used all the subjects who had at least one feature type available among four different data sources including MRI, PET, CSF and proteomics, and challenge our method with the problem of distinguishing AD, MCI and NC subjects from each other. In our binary classification test scenarios, the relative performances of different methods are evaluated using several metrics including accuracy, sensitivity and specificity.

3.1 Subjects

Data used in this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`http://adni.loni.ucla.edu`). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and nonprofit organizations, as a $60 million, 5-year public-private partnership. The primary 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 AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

In terms of data collected at baseline, a total of 822 ADNI participants were recruited from 59 sites across the U.S. and Canada. These including 229 Normal (normal elderly controls), 405 with MCI, and 188 with AD, ranging in age from 55 to 90 years. Phenotype data included structural MRI scans acquired on 1.5T MRI scanners, and clinical and neuropsychological assessments. Additional data such as 3-Tesla MRI, FDG-PET, PIB-PET (a PET scanning method using the amyloid-sensitive ligand, Pittsburgh compound B), and fluid biomarkers are available for some subjects as well. In our experiments, we use pre-processed 1.5 Tesla (T) MRI imaging features. Besides these data, we were able to include other subjects who had at least one of three data types available: FDG-PET, CSF, and/or proteomics. As a result, baseline data from a total of 780 participants (172 AD, 397 MCI, 211 Normal) were used to test our algorithms.

The MRI image features in this study were based on the imaging data from the ADNI database processed by the UCSF team, who performed cortical reconstruction and volumetric segmentations with the FreeSurfer image analysis suite (`http://surfer.nmr.mgh.harvard.edu/`). The processed MRI features come from a total of 648 subjects (138 AD, 319 MCI and 191 Normal), and can be grouped into 5 categories: average cortical thickness, standard deviation in cortical thickness, the volumes of cortical parcellations, the volumes of specific white matter parcellations, and the total surface area of the cortex. There were 305 MRI features in total. We also downloaded FDG-PET images of 327 subjects (74 AD, 172 MCI, and 81 Normal) from the ADNI website. With SPM8 (`http://www.fil.ion.ucl.ac.uk/spm/`), we processed these FDG-PET images. We applied Automated Anatomical Labeling (AAL) [36] to extract each of the 116 anatomical volumes of interest (AVOI) and derived average image values from each AVOI, for every subject. Baseline CSF samples were acquired from 416 subjects (102 AD, 200 MCI and 114 Normal) by the ADNI Biomarker Core laboratory at the University of Pennsylvania Medical Center [36]. In our study, we use 5 measures obtained from the CSF, including levels of beta amyloid 1-42 ($A\beta_{1-42}$), tau protein (Tau), phosphorylated-tau protein 181 (pTau$_{181p}$) along with two CSF ratios (Tau/$A\beta_{1-42}$ and pTau$_{181p}$/$A\beta_{1-42}$). The proteomics data set (97 AD, 345 MCI, and 54 Normal) was produced by the Biomarkers Consortium Project "Use of Targeted Multiplex Proteomic Strategies to Identify Plasma-Based Biomarkers in Alzheimer's Disease"[1] (see URL in footnote). We use 147 measures from the proteomic data downloaded from the ADNI web site. As a result, for a subject with all four types of data available, a total of 573 measures were studied in our classification experiment. The number of samples from each category corresponding to each type of feature utilized in this study is summarized in Table 1.

3.2 Comparison of iMSF, ScoreComp and imputation methods

In our first set of experiments, we apply our proposed methods to the full multi-source data set including MRI, PET, proteomics and CSF for solving clinical group classification problems (AD vs. Normal; AD vs. Non-converter and Converter vs. Normal). 780

[1] `http://adni.loni.ucla.edu/wp-content/uploads/2010/11/BC_Plasma_Proteomics_Data_Primer.pdf`

Table 1: The number of available samples and features used in this study.

	MRI	PET	CSF	Proteomics
# of AD Subjects	138	74	102	97
# of MCI Subjects	319	172	200	345
# of Normal Subjects	191	81	114	54
# of Features	305	116	5	147

subjects were analyzed. Among them, each subject has at least one of the four data sources (MRI, FDG-PET, CSF and proteomics features) available. We first randomly select a portion (from 50% to 75%) of samples as the training set to learn the model, and then apply the model to predict the labels on the remaining data, used as a non-overlapping test set. We repeated this process 30 times; the average performance is reported.

For comparison purposes, we included the following missing value estimation methods:

Zero: this is the most intuitive way to impute missing values - we assign zero to any element that is missing. When the data set is first normalized to have zero mean and unit standard deviation, this is equivalent to mean value imputation.

KNN: missing value imputation using the k-nearest neighbor method [15]. The KNN method replaces the missing value in the data matrix with the corresponding value from the nearest column. That is to say, KNN will first identify the most similar feature to the current one with a missing value, and then use this feature as a guess for the missing one.

EM: this method imputes missing values using the expectation-maximization (EM) algorithm [34]. An iteration of the EM algorithm includes two steps. In the E step, we estimate the mean and covariance matrix from the data matrix (with missing values filled with guesses from previous M step, or initialized as zeros); then in the M step, the missing value of each data column is filled in with their conditional expectation values based on the available values and the estimated mean and the covariance. We then re-estimate the mean and the variance based on the new estimates, therefore entering the next EM iteration.

SVD: this is a standard method for matrix completion based on a low rank approximation. The SVD based estimation works in a similar way to the EM method above. We first provide some initial guesses (such as 0) to the missing data values, and then we apply singular value decomposition (SVD) to obtain a low-rank approximation of the filled-in matrix. Next, we update the missing values using their corresponding values in the low-rank estimation. Finally, we apply SVD to the updated matrix again and the process is repeated until convergence.

SVT: Recently, trace norm minimization has been proposed for missing data estimation [4, 5]. This can be effective even when a large amount of data is missing. Therefore, it will be interesting to see how this algorithm (singular value thresholding or SVT) performs in our particular setting. We acquire the SVT program online (`http://svt.stanford.edu`) and follow their suggestions for parameter setting.

The classification results are summarized in Table 2. For our proposed iMSF method, five ratios (0.001, 0.01, 0.1, 0.2 and 0.4) were used for the regularization parameter λ. We first run our iMSF method to learn a set of features for each different task, and then train a random forest classifier [25] on each learning task using the selected features. The best and average performance obtained using these five ratios are reported in Table 2, as "iMSF-Best" and "iMSF-Average" respectively. For the proposed ScoreComp scheme, we first use random forest to obtain the prediction score for the incomplete data set, and apply three different missing value estimation methods (Zero, EM and KNN) to obtain the completed score matrix. Then, ridge regression is applied to integrate the data sources together. The performance using three different missing value estimation methods is reported in Table 2 as "ScoreComp-Zero", "ScoreComp-EM" and "ScoreComp-KNN" respectively.

From Table 2, we can observe that for the AD vs. Normal problem, our ScoreComp performs best, with about 90% accuracy. The top 2 methods in terms of performance are "ScoreComp-EM" and "ScoreComp-KNN" in all training ratios, and they tend to produce comparable results. The AD vs. Normal problem is considered to be less challenging, and this might explain why its simplicity may give ScoreComp method an edge over iMSF.

In the more challenging settings where MCI subjects are involved, we obtain low sensitivity in the AD vs. Non-converter case and low specificity in the Converter vs. Normal case. In these settings, iMSF performs much better – it provides more balanced classification results on both the positive and negative classes. For example, in the Converter vs. Normal problem, using 75% of data for training, missing value estimation based methods only achieve about 70% specificity, but our proposed iMSF achieves an average of 89% among the five different parameters. Therefore, even though ScoreComp achieves a higher accuracy in the AD vs. Non-converter case, we still consider iMSF as the best method for these settings. This may be due to the fact that our iMSF algorithm took a more systematic approach in using multiple sources of information for classification. A detailed comparison with published results using the ADNI data set may be found in the Discussion section.

Among different variations of the ScoreComp method illustrated in Table 2, "ScoreComp-EM" and "ScoreComp-KNN" produce more stable results than "ScoreComp-Zero", where the missing scores are substituted with zeros. For example, in the AD vs. Non-converter case, all three variations yield comparable performance, while in the Converter vs. Normal case, "ScoreComp-Zero" yields much lower specificity. We can conclude that in the missing value estimation step of our ScoreComp method, an effective algorithm can indeed enhance the classification performance by using the model scores from other samples. Interestingly, the five other different missing value estimation methods (Zero, EM, KNN, SVD and SVT) perform comparably to each other. Thus, estimating the block-wise missing values directly does not give much edge over simply substituting missing elements with zeros. This further justifies the effectiveness of our ScoreComp method.

3.3 Comparison with single source classification

Though it is a common belief that by integrating multiple heterogeneous sources, one can predict AD progression more accurately, it is still interesting to see how much we improve classification performance by using multi-modality data. Therefore, we extract the 648 subjects that have MRI features available (with complete data), and perform leave-one-out classification on the same problems we discussed before. We then extract the classification results for the same 648 subjects from our proposed methods, so that the comparison can be made using the same sample pool. Results are summarized in Table 3.

As we see in Table 3, using multiple data sources greatly improves the performance in each case. This is because we not only learn from additional information from the current sample (when data sources other than MRI are available), but we also use the information from other samples that are thrown away in the unimodal case.

Table 2: Classification performance comparison of the proposed iMSF, ScoreComp and missing value estimation methods (Zero, EM, KNN, SVD and SVT) in terms of accuracy, sensitivity and specificity when the training percentage varies from 50% to 75%.

	AD vs. Normal			AD vs. Non-converter			Converter vs. Normal		
Training Size 50%	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
iMSF-Best	86.37%	85.64%	**87.65%**	78.05%	**57.34%**	89.91%	**91.20%**	94.11%	**89.59%**
iMSF-Average	83.00%	81.56%	84.80%	77.51%	55.97%	88.64%	90.34%	93.31%	84.18%
ScoreComp-Zero	87.31%	87.61%	87.35%	81.25%	53.26%	**95.68%**	83.15%	**99.32%**	48.12%
ScoreComp-EM	89.74%	94.13%	85.32%	81.05%	53.35%	95.31%	88.99%	98.73%	68.30%
ScoreComp-KNN	**90.13%**	**94.87%**	85.37%	**81.46%**	56.02%	94.58%	87.60%	98.89%	63.49%
Zero	85.38%	87.74%	83.23%	77.25%	44.10%	94.14%	89.48%	97.93%	71.36%
EM	87.20%	88.76%	85.75%	78.11%	47.25%	93.80%	89.40%	97.97%	70.95%
KNN	85.79%	88.10%	83.66%	75.15%	36.78%	94.59%	86.40%	97.88%	61.41%
SVD	83.89%	84.94%	82.96%	75.73%	43.00%	92.37%	87.57%	97.09%	67.07%
SVT	55.75%	61.31%	50.70%	81.15%	56.82%	93.48%	71.69%	87.81%	36.34%
Training Size 66.7%	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
iMSF-Best	88.77%	86.77%	**91.25%**	80.55%	**65.83%**	89.03%	**93.17%**	95.36%	**90.52%**
iMSF-Average	86.07%	83.23%	89.23%	80.12%	63.70%	88.16%	92.40%	94.72%	87.37%
ScoreComp-Zero	88.18%	86.48%	90.10%	**83.48%**	59.90%	**94.96%**	84.48%	**99.53%**	50.13%
ScoreComp-EM	**90.28%**	93.64%	86.61%	83.25%	60.85%	94.12%	90.82%	99.16%	72.07%
ScoreComp-KNN	90.19%	**94.48%**	85.51%	83.30%	62.25%	93.56%	89.56%	99.24%	67.66%
Zero	86.79%	88.80%	84.61%	80.87%	52.21%	94.56%	92.30%	98.84%	77.32%
EM	88.46%	88.40%	88.52%	80.90%	53.28%	93.96%	91.69%	98.76%	75.75%
KNN	87.20%	87.60%	86.75%	78.46%	44.24%	94.75%	89.07%	98.68%	66.84%
SVD	85.25%	85.19%	85.36%	78.52%	48.89%	92.51%	90.44%	98.13%	72.75%
SVT	57.14%	61.08%	53.39%	83.25%	60.71%	94.01%	75.30%	91.66%	38.03%
Training Size 75%	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
iMSF-Best	88.79%	87.16%	**90.92%**	81.33%	**66.02%**	89.06%	**93.19%**	95.94%	**90.58%**
iMSF-Average	85.81%	82.77%	89.38%	80.78%	65.20%	88.70%	92.98%	95.03%	88.87%
ScoreComp-Zero	88.53%	87.67%	89.75%	**84.51%**	61.48%	**96.00%**	84.96%	**99.60%**	52.74%
ScoreComp-EM	89.91%	92.97%	86.77%	83.80%	61.24%	94.95%	89.93%	99.30%	70.21%
ScoreComp-KNN	**90.35%**	**93.89%**	86.74%	84.35%	63.79%	94.60%	88.89%	99.11%	66.91%
Zero	86.49%	88.06%	85.13%	81.76%	55.54%	94.74%	91.48%	97.68%	78.49%
EM	88.96%	88.54%	89.54%	81.25%	55.72%	93.83%	90.74%	98.44%	74.38%
KNN	87.32%	88.20%	86.66%	77.73%	44.97%	93.66%	88.89%	98.43%	68.12%
SVD	85.93%	85.78%	86.23%	78.55%	50.13%	92.29%	89.19%	97.48%	70.92%
SVT	56.58%	59.38%	53.73%	81.92%	56.93%	94.17%	75.56%	90.23%	42.59%

Table 3: Classification performance comparison of using multi-modality data (iMSF and ScoreComp) and just using MRI data. The classification is performed on the same set of samples, so that a fair comparison can be made. Leave-one-out is used and the accuracy, sensitivity and specificity are reported.

	AD vs. Normal			AD vs. Non-converter			Converter vs. Normal		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Baseline	86.45%	83.33%	89.36%	76.83%	50.60%	89.20%	85.80%	96.49%	60.42%
iMSF-Average	90.99%	90.28%	91.76%	77.84%	69.44%	83.84%	92.59%	94.34%	89.29%
ScoreComp-EM	90.82%	94.37%	87.02%	82.24%	63.89%	95.36%	90.66%	99.06%	71.43%

3.4 Comparison with methods that throw out missing data

An intrinsic advantage of our methods over throwing out missing data is that no sample will be wasted. The final learned model will benefit from all the samples, so long as at least one of the data sources is available. Also, unlike the one learned from a complete data set, our final model will be able to give a prediction for a newly arrived sample with any combination of the data sources. Still, it is interesting to investigate if this additional information can help improve the classification performance. Therefore, we extract the complete data set where each sample has all the four data sources (MRI, CSF, PET and proteomics) available. We then extract the classification results for the same 153 subjects from our methods (iMSF and ScoreComp), so that the comparison can be made using the same sample pool. Results are summarized in Table 4.

As we see in Table 4, using 153 subjects alone (about 20% of the 780 subjects we use) results in unsatisfactory performance. We can conclude that our method not only makes full use of the information available, but also greatly improves classification performance.

3.5 Effects of different λ ratios in the proposed iMSF

From Table 2, we can observe that in terms of classification performance, our iMSF method is not very sensitive to the parameter λ. Here we use a particular example to illustrate how the results vary when different λ ratios are chosen in our proposed iMSF method in Figure 5. We use the AD vs. Normal problem, and report leave-one-out results when we use different choices of λ ratio values (0.001, 0.01, 0.1, 0.2 and 0.4). As we can observe from the figure, as we increase λ, the number of features selected will gradually decrease, from about 25% of the features to 3%. This shows that by using a λ ratio instead of actual values, it is much easier to choose a range of parameters that lead to desired levels of sparsity. We can also observe that the best choice of λ lies in the middle of the region (in this example 0.2), but the performance is not very sensitive to the parameter.

4. DISCUSSION

This paper has two major contributions. First, we were able

Table 4: Classification comparison of using multi-modality data (iMSF and ScoreComp) and using the complete MRI + CSF + PET + Proteomics data. Classification is performed on the same set of samples, so that a fair comparison can be made. Leave-one-out is used and the accuracy, sensitivity and specificity are reported.

	AD vs. Normal			AD vs. Non-converter			Converter vs. Normal		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Baseline	87.72%	100.00%	63.16%	76.19%	65.79%	84.78%	83.02%	97.06%	57.89%
iMSF-Average	95.79%	97.37%	92.63%	84.05%	79.47%	87.83%	88.68%	90.59%	85.26%
ScoreComp-EM	98.25%	97.37%	100.00%	83.33%	65.79%	97.83%	84.91%	97.06%	63.16%

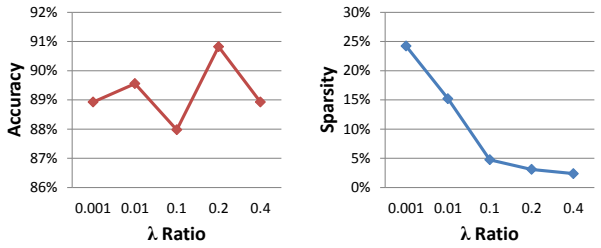

Figure 5: Illustration of the results obtained using different λ ratio values in our proposed iMSF method. We vary the λ ratio values from 0.001 to 0.4 (x-axis) and report the accuracy obtained (y-axis) in the left figure. In the right figure, we report the proportion of selected features (Sparsity, y-axis) when we increase λ ratio values from 0.001 to 0.4 (x-axis).

to use a large multi-modal data set for classification, even when large segments of the data were missing. Secondly, we propose two different multi-source learning schemes for block-wise incomplete data, and used them to construct automatic, robust classifiers. In our experiments, our systems greatly improved the classification accuracy on the ADNI data set. Our method (iMSF) has two major advantages: 1) All subjects, so long as at least one of the feature sources is available, can be used for feature learning, and all of them can contribute to the feature selection jointly; 2) the difficulty of guessing unknowns is bypassed, as the feature learning is only based on what data is available. To the best of our knowledge, in the ADNI data set, we are the first group who tried to utilize all the available information for classification by allowing the use of subjects with incomplete data. In our current pilot work, we assessed whether our multi-source learning models help to boost the statistical power in the whole data set. Except for some of the PET imaging measures (such as Pittsburgh compound B), we used all other measures that are publicly available at ADNI web site. We hope our work will increase interest in this important problem and that other groups might consider using this approach (not throwing out data) when performing future ADNI classification studies.

Pioneering work has been done on the automated diagnosis problem using the ADNI data set. López et al. (1995) applied principal component analysis to extract features from FDG-PET. In a data set of 211 subjects (53 AD, 114 MCI and 52 Normal), they achieved a best leave one out (LOO) accuracy 82% for classifying people into groups of AD vs. Normal and a best LOO accuracy of 81% on MCI vs. Normal. Cuingnet et al. [8] evaluated the performance of ten high dimensional classification methods using the ADNI MRI data. In their data set of 509 subjects, the best of the ten classifiers achieved 81%/95%, 65%/94% sensitivity/specificity for classification of AD vs. Normal, MCI vs. Normal, respectively. In our earlier work [21], we used support vector machines to combine several MRI measures, as well as PET and CSF biomarkers, etc. and we achieved a 90% LOO accuracy on AD vs. Normal and a 75% LOO accuracy on MCI vs. Normal classification. Notably, all of these studies were applied to a subset of the full available ADNI data used here. For example, in [21], 635 subjects were studied when only MRI-based measures were needed, but when both CSF and PET were also added to the set of predictors, the available sample size dropped dramatically to 166 subjects. Without a method to include subjects with missing data, it becomes quite difficult to train and test a classifier. The approach we outlined here still achieved comparable or better results than those in prior papers.

There are several possibilities for extending our current work. In this paper, we used numerical summary measures from MRI scans of 648 subjects, whose data were available at ADNI data set. In some of our earlier studies [16], we used tensor-based morphometry to study baseline and longitudinal MRI scans in ADNI, and these could be added to the feature set in the future. In addition, the second phase of the ADNI initiative is now collecting data from diffusion tensor imaging, arterial spin labeling, and resting state functional MRI. Although each of these features is likely to help with classification and for predicting decline, the 3 new imaging modalities will not all be performed on the same subjects - in fact, each of the ADNI subjects will be scanned using only one of the 3 additional modalities, because it was not feasible to prolong the scanning session to include all three in every subject. Such a situation lends itself to the data mining and machine learning approach developed here, as there will be considerable joint information available about the relationships between the new modalities and the traditional biomarkers, but not in the same subjects. Also, ensemble learning [10] can boost performance in general data mining and machine learning problems. By combining various models produced with different parameters, or even models from different methods, the ensemble method may produce even more stable and robust classifiers. In the future, we plan to enrich our model set and use ensemble method to tackle the incomplete data problem.

5. ACKNOWLEDGMENTS

This work was funded by the National Institute on Aging (AG016570 to PMT), the National Library of Medicine, the National Institute for Biomedical Imaging and Bioengineering, and the National Center for Research Resources (LM05639, EB01651, RR019771 to PMT), US National Science Foundation (NSF) (IIS-0812551, IIS-0953662, CCF-1025177 to JY), and National Library of Medicine (R01 LM010730 to JY).

6. REFERENCES

[1] R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6:1817–1853, 2005.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[3] M. Braskie, A. Klunder, K. Hayashi, H. Protas, V. Kepe, K. Miller, S. Huang, J. Barrio, L. Ercoli, P. Siddarth, et al. Plaque and tangle imaging and cognition in normal aging and Alzheimer's disease. *Neurobiology of Aging*, 31(10):1669–1678, 2010.

[4] J. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal of Optimization*, 20(4):1956–1982, 2010.

[5] E. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010.

[6] R. Casanova, R. Srikanth, A. Baer, P. Laurienti, J. Burdette, S. Hayasaka, L. Flowers, F. Wood, and J. Maldjian. Biological parametric mapping: a statistical toolbox for multimodality brain image analysis. *Neuroimage*, 34(1):137–143, 2007.

[7] P. Combettes and J. Pesquet. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212, 2011.

[8] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, and M. Habert. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *Neuroimage*, 2010.

[9] D. Devanand, G. Pradhaban, X. Liu, A. Khandji, S. De Santi, S. Segal, H. Rusinek, G. Pelton, L. Honig, R. Mayeux, et al. Hippocampal and entorhinal atrophy in mild cognitive impairment. *Neurology*, 68(11):828–836, 2007.

[10] T. Dietterich. Ensemble methods in machine learning. *Multiple Classifier Systems*, pages 1–15, 2000.

[11] Y. Fan, S. Resnick, X. Wu, and C. Davatzikos. Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. *Neuroimage*, 41(2):277–285, 2008.

[12] C. Fennema-Notestine, D. Hagler Jr, L. McEvoy, A. Fleisher, E. Wu, D. Karow, and A. Dale. Structural MRI biomarkers for preclinical and mild Alzheimer's disease. *Human Brain Mapping*, 30(10):3238–3253, 2009.

[13] S. Gao. A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data. *Statistics in Medicine*, 23(2):211–219, 2004.

[14] S. Hardy, H. Allore, and S. Studenski. Missing data: A special challenge in aging research. *Journal of the American Geriatrics Society*, 57(4):722–729, 2009.

[15] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein. Imputing missing data for gene expression arrays. *Technical Report*, 1999.

[16] X. Hua, B. Gutman, C. Boyle, P. Rajagopalan, A. Leow, I. Yanovsky, A. Kumar, A. Toga, C. Jack Jr, N. Schuff, et al. Accurate measurement of brain changes in longitudinal MRI scans using tensor-based morphometry. *Neuroimage*, 2011.

[17] V. Ibanez, P. Pietrini, G. Alexander, M. Furey, D. Teichberg, J. Rajapakse, S. Rapoport, M. Schapiro, and B. Horwitz. Regional glucose metabolic abnormalities are not the result of atrophy in Alzheimer's disease. *Neurology*, 50(6):1585–1593, 1998.

[18] C. Jack, V. Lowe, M. Senjem, S. Weigand, B. Kemp, M. Shiung, D. Knopman, B. Boeve, W. Klunk, C. Mathis, et al. [11C]-PiB and structural MRI provide complementary information in imaging of Alzheimer's disease and amnestic mild cognitive impairment. *Brain*, 131(3):665, 2008.

[19] C. Jack Jr, F. Barkhof, M. Bernstein, M. Cantillon, P. Cole, C. DeCarli, B. Dubois, S. Duchesne, N. Fox, G. Frisoni, et al. Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. *Alzheimer's and Dementia*, 7(4):474–485, 2011.

[20] C. Jack Jr, M. Bernstein, N. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. Britson, J. L Whitwell, C. Ward, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

[21] O. Kohannim, X. Hua, D. Hibar, S. Lee, Y. Chou, A. Toga, C. Jack Jr, M. Weiner, P. Thompson, et al. Boosting power for clinical trials using classifiers based on multiple biomarkers. *Neurobiology of Aging*, 31(8):1429–1442, 2010.

[22] R. Kuljiš. Grand challenges in dementia 2010. *Frontiers in Neurology*, 1, 2010.

[23] S. Landau, D. Harvey, C. Madison, R. Koeppe, E. Reiman, N. Foster, M. Weiner, and W. Jagust. Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI. *Neurobiology of Aging*, 32(7):1207–1218, 2011.

[24] S. Lemm, B. Blankertz, T. Dickhaus, and K. Muller. Introduction to machine learning for brain imaging. *Neuroimage*, 2010.

[25] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[26] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient L2,1-Norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 339–348. AUAI Press, 2009.

[27] J. Liu, L. Yuan, and J. Ye. An efficient algorithm for a class of fused lasso problems. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 323–332. ACM, 2010.

[28] E. Martinez-Montes, P. Valdés-Sosa, F. Miwakeichi, R. Goldman, and M. Cohen. Concurrent EEG/fMRI analysis by multiway partial least squares. *Neuroimage*, 22(3):1023–1034, 2004.

[29] S. Mueller, M. Weiner, L. Thal, R. Petersen, C. Jack, W. Jagust, J. Trojanowski, A. Toga, and L. Beckett. The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging Clinics of North America*, 15(4):869, 2005.

[30] Y. Nesterov. Gradient methods for minimizing composite objective function. *ReCALL*, 76(2007076), 2007.

[31] Y. Nesterov and I. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

[32] R. Palmer and D. Royall. Missing data? Plan on it! *Journal of the American Geriatrics Society*, 58:S343–S348, 2010.

[33] E. Reiman, J. Langbaum, and P. Tariot. Alzheimer's prevention initiative: a proposal to evaluate presymptomatic treatments as quickly as possible. *Biomarkers*, 4(1):3–14, 2010.

[34] T. Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14(5):853–871, 2001.

[35] J. Schott, J. Bartlett, J. Barnes, K. Leung, S. Ourselin, N. Fox, and ADNI. Reduced sample sizes for atrophy outcomes in Alzheimer's disease trials: baseline adjustment. *Neurobiology of Aging*, 31(8):1452, 2010.

[36] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[37] P. Van Ness, T. Murphy, K. Araujo, M. Pisani, and H. Allore. The use of missingness screens in clinical epidemiologic research has implications for regression modeling. *Journal of clinical epidemiology*, 60(12):1239–1245, 2007.

[38] P. Vemuri, H. Wiste, S. Weigand, L. Shaw, J. Trojanowski, M. Weiner, D. Knopman, R. Petersen, C. J. Jack, and ADNI. MRI and CSF biomarkers in normal, MCI, and AD subjects: diagnostic discrimination and cognitive correlations. *Neurology*, 73(4):287–293, 2009.

[39] Y. Wang, Y. Fan, P. Bhatt, and C. Davatzikos. High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. *Neuroimage*, 50(4):1519–1535, 2010.

[40] K. Worsley, J. Poline, K. Friston, and A. Evans. Characterizing the response of PET and fMRI data using multivariate linear models. *Neuroimage*, 6(4):305–319, 1997.

[41] J. Ye, K. Chen, T. Wu, J. Li, Z. Zhao, R. Patel, M. Bae, R. Janardan, H. Liu, G. Alexander, et al. Heterogeneous data fusion for Alzheimer's disease study. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 1025–1033. ACM, 2008.

[42] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[43] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, et al. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage*, 2011.

[44] J. Zhou, J. Chen, and J. Ye. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011.