Heterogeneous Data Fusion for Alzheimer's Disease Study

Jieping Ye [1,2], Kewei Chen [5], Teresa Wu [3], Jing Li [3], Zheng Zhao [2], Rinkal Patel [2], Min Bae [3], Ravi Janardan [4], Huan Liu [2], Gene Alexander [5], and Eric Reiman [5]
[1]Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University, Tempe, AZ 85287
[2]Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287
[3]Department of Industrial Engineering, Arizona State University, Tempe, AZ 85287
[4]Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN 55455
[5]Banner Alzheimer's Institute and Banner PET Center, Banner Good Samaritan Medical Center, Phoenix, AZ 85006

ABSTRACT

Effective diagnosis of Alzheimer's disease (AD) is of primary importance in biomedical research. Recent studies have demonstrated that neuroimaging parameters are sensitive and consistent measures of AD. In addition, genetic and demographic information have also been successfully used for detecting the onset and progression of AD. The research so far has mainly focused on studying one type of data source only. It is expected that the integration of heterogeneous data (neuroimages, demographic, and genetic measures) will improve the prediction accuracy and enhance knowledge discovery from the data, such as the detection of biomarkers. In this paper, we propose to integrate heterogeneous data for AD prediction based on a kernel method. We further extend the kernel framework for selecting features (biomarkers) from heterogeneous data sources. The proposed method is applied to a collection of MRI data from 59 normal healthy controls and 59 AD patients. The MRI data are pre-processed using tensor factorization. In this study, we treat the complementary voxel-based data and region of interest (ROI) data from MRI as two data sources, and attempt to integrate the complementary information by the proposed method. Experimental results show that the integration of multiple data sources leads to a considerable improvement in the prediction accuracy. Results also show that the proposed algorithm identifies biomarkers that play more significant roles than others in AD diagnosis.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - Data Mining; J.3 [**Life and Medical Sciences**]: Health, Medical information systems

General Terms

Algorithms

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

Keywords

Neuroimaging, tensor factorization, heterogeneous data source fusion, multiple kernel learning, biomarker detection

1. INTRODUCTION

Currently, approximately 5 million people in the US - about 10% of the population over 60 are afflicted by Alzheimer's disease (AD), the most common form of dementia. The direct cost to care the patients by family members or health care professional is estimated to be over $100 billion per year. As the population ages over the next several decades, it is expected that the AD cases and the associated costs will go up dramatically. Recognizing the urgent need to slow down or completely prevent from the occurrence of a health care crisis in US and worldwide, AD researchers have intensified their efforts to investigate ways to delay, cure, or prevent the onset and progression of AD. Objective and quantitative criteria, so called, biomarkers, are essential to evaluate the effectiveness of a potential treatment or prevention strategy. Thus, research on exploring biomarkers in the form of a test of cerebrospinal fluid (CSF) or blood, or images from brain scans has attracted great attention.

Recent studies have demonstrated that imaging parameters from brain scans are more sensitive and consistent measures of disease progression than cognitive assessment [32]. Some studies have shown that imaging measures correlate with cognitive test performance in Mild Cognitive Impairment (MCI)[1] and AD - an initial step in the validation of markers that accurately predict the course of the disease. Evidently, neuroimaging research offers great potential to identify the sensitive and specific biomarkers that can identify individuals early in the course of dementing illness. This opens up opportunities to implement treatments in the early stages of disease when intervention may be most beneficial.

The volumetric T1 weighted MRI is a high-resolution structural imaging technique that allows for the visualization of brain anatomy with a high degree of contrast between brain tissue types. It can be used to measure specific structures (e.g., hippocampus, entorhinal cortex, amygdale, etc.), a region of interest (ROI) approach, and detect the volume changes of the structures for AD vs. Normal [21]. Recently,

[1]Mild Cognitive Impairment (MCI) is a transition stage between the cognitive changes of normal aging and the more serious problems related to dementia.

structural MRI images have been used to quantify reductions of whole brain volume in sequentially acquired scans [8]. Promising methodological developments in the analysis of structural MRI data also include the use of probabilistic brain maps [3] to compute regional alterations in gray matter, white matter, CSF, and whole brain, and to examine cross-sectional difference or longitudinal changes on voxel-by-voxel basis, an approach referred to as the voxel-based morphometry (VBM) [36]. Another neuroimaging technique is the so called positron emission tomography (PET). With different radioactive tracers, PET provides information on various physiological, biochemical and/or metabolic processes. In addition, other types of data, e.g., demographic information, such as age, gender, education, genetic makeup (such as the possession of the allele of Apolipoprotein e4), etc, have also been shown to be associated with AD.

While promising, the research so far has mainly focused on studying only one type of neuroimaging, e.g., MRI, fMRI (functional MRI), or FDG-PET using either region of interest or voxel-based approach [2, 8, 9, 24, 35, 44]. It is expected that combining different types of neuroimaging will help the prediction. However, even for the same neuroimaging data, different features constructed by different approaches (region of interest versus voxel-based approaches) might complement each other. Integrating ROI and voxel-based information from the same type of neuroimaging data and incorporating demographic information is expected to improve the prediction.

In this paper, we propose a kernel method to fuse heterogeneous data (different types of features from single MRI data, together with demographic and genetic information) for accurately classifying subjects and discovering useful knowledge on biomarkers. More specifically, our main contributions to the AD research are summarized as follows:

- **Dimensionality Reduction via Tensor Factorization:** Neuroimages such as those from MRI are represented as a three-dimensional array, which contains a huge number of features (voxels). Due to the natural tensor representation of such images, dimensionality reduction based on tensor factorization is one effective approach for reducing data dimensionality [19, 25, 26]. We propose to apply N-mode SVD [41] and the out-of-core technique [43] for the factorization.

- **Multiple Data Source Fusion via Multiple Kernel Learning:** The integration of different types of features (region of interest and tensor features) from MRI data and non-imaging data such as demographic information is expected to improve the prediction accuracy for our AD study. Multiple kernel learning (MKL) provides a general framework for learning from multiple data sources. It has been applied for combining various biological data for enhanced biological inference [22]. We propose to apply a discriminant analysis-based formulation [47] for the integration.

- **Knowledge (Biomarker) Discovery via Multi-source Feature Selection:** In addition to offering a more accurate prediction of AD, another important component of the AD study is knowledge gained on the linkage between structural and functional abnormalities. Feature selection [13, 29] is an effective vehicle for such a discovery and the use of multiple data sources is expected to give a more accurate determination of biomarkers as well as to aid its understanding. Traditional feature selection algorithms work on a single data source only. We propose to integrate multiple kernel learning and traditional feature selection for biomarker detection from multiple data sources.

To the best of our knowledge, our uses of (1) tensor factorization for extracting features from AD-related neuroimages; (2) multiple kernel learning for integrating AD-related multiple data sources; and (3) feature selection from multiple data sources are novel contributions to the AD research.

We evaluated the proposed method using data acquired under the support of ADNI[2] from 59 normal control (NC) and 59 AD patients. The data includes 118 MRI images represented as a three-dimensional array of size $181 \times 217 \times 181$ (in the standard Talairach space and with 1 cubic mm voxel size), as well as demographic information such as age, gender, and genetic information based on Apolipoprotein E e4 (APOE4). SPM5[3] was used together with the optimized voxel-based-morphometry to generate the modulated gray matter map in the customized template space[4]. Experimental results show that multiple kernel learning achieves a considerable improvement in the prediction accuracy in comparison with classification based on each data source individually. Results also show that the proposed algorithm is able to identify a number of brain regions that are known to be affected by AD.

The rest of the paper is organized as follows. Tensor factorization for feature extraction from MRI data is presented in Section 2. We introduce multiple kernel learning for heterogeneous data fusion in Section 3. Biomarker detection from multiple data sources is discussed in Section 4. Experimental results are presented in Section 5. Finally, Section 6 concludes this paper with discussions and future work.

2. IMAGE FEATURE EXTRACTION VIA TENSOR FACTORIZATION

As high resolution 3D data, volumetric MRI data have a huge number of voxels at the time they were acquired from each subject. In our VBM pre-processing, we kept the voxel size in the template space to be in 1 cubic mm resulting in the image dimension of $181 \times 217 \times 181$. Dimensionality reduction, which extracts a small number of features by removing the irrelevant, redundant, and noisy information, is crucial for the analysis of such data. Some neuroimaging studies use sub-sampling or a region of interest (ROI) based approach such as the Automated Anatomical Labeling (AAL) [40] to reduce the data dimensionality. Although the within-ROI variation is ignored, AAL ROI summarizes the information from multiple brain regions with much reduced dimension and these regions are representative over the whole brain volume. These techniques, however, may not be able to account for the information variation within each region of interest. Additionally, a traditional dimensionality reduction technique called Principle Component Analysis (PCA) [16] has been used widely in many applications [34, 39], including a well-known adaptation in the

[2]http://www.loni.ucla.edu/Research/Databases/
[3]http://www.fil.ion.ucl.ac.uk/spm/
[4]Gene Alexander's group processed the MRI images used in this study. More details on these data can be found in Section 5.1.

neuroimaing field, often referred to as the scaled sub-profile modeling (SSM) [1]. PCA adopts the vector representation for images by concatenating all voxels within a pre-defined brain volume into a single vector. One inherent problem with this approach is that some information on spatial relationships (such as the correlation among different slices of the 3D image) is not explicitly accounted for. One effective way to overcome these limitations is to treat a collection of three-dimensional images as a tensor and apply tensor factorization [19, 25, 26].

2.1 Background on Tensors

A *tensor*, also known as multidimensional matrix [45], is a higher order generalization of a vector (first order tensor) and a matrix (second order tensor). An Nth-order tensor is denoted as $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$. An element of the tensor $\mathcal{A}$ is denoted as $a_{i_1 \cdots i_n \cdots i_N}$, where $1 \leq i_n \leq I_n$, for $n = 1, \cdots, N$. An Nth-order tensor $\mathcal{A}$ is of rank-one if it can be expressed as the outer product of N vectors:

$$\mathcal{A} = \mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \cdots \otimes \mathbf{x}_N, \tag{1}$$

where $\mathbf{x}_n \in \mathbb{R}^{I_n}$, for all $1 \leq n \leq N$.

A generalization of the product of two matrices is the product of a tensor and a matrix. The mode-n product of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ and a matrix $Q \in \mathbb{R}^{J_n \times I_n}$, whose element is denoted as $q_{j_n i_n}$, where $1 \leq j_n \leq J_n$ and $1 \leq i_n \leq I_n$, is a tensor, denoted as

$$\mathcal{A} \times_n Q \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \cdots \times I_N}, \tag{2}$$

whose entries are given by

$$(\mathcal{A} \times_n Q)_{i_1 \cdots i_{n-1} j_n i_{n+1} \cdots i_N} = \sum_{i_n} a_{i_1 \cdots i_{n-1} i_n i_{n+1} \cdots i_N} q_{j_n i_n}. \tag{3}$$

Let $\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ be another tensor, whose general element is denoted as $b_{i_1 \cdots i_n \cdots i_N}$. The scalar product of two tensors $\mathcal{A}$ and $\mathcal{B}$ is defined as:

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i_1} \sum_{i_2} \cdots \sum_{i_N} a_{i_1 \cdots i_n \cdots i_N} b_{i_1 \cdots i_n \cdots i_N}. \tag{4}$$

The Frobenius norm of a tensor $\mathcal{A}$ is then defined as

$$||\mathcal{A}|| = \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}. \tag{5}$$

The mode-n vectors of $\mathcal{A}$ are the I_n-dimensional vectors obtained from $\mathcal{A}$ by varying index i_n while keeping the other indices fixed. They form the column vectors of matrix $A_{(n)} \in \mathbb{R}^{I_n \times (I_1 I_2 \cdots I_{n-1} I_{n+1} \cdots I_N)}$ that results from flattening the tensor $\mathcal{A}$. The n-th rank of $\mathcal{A}$, denoted as $\text{rank}_n(\mathcal{A})$, is defined as the dimension of the vector space spanned by the mode-n vectors: $\text{rank}_n(\mathcal{A}) = \text{rank}(A_{(n)})$.

2.2 Tensor Factorization

Given a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, a rank-$(R_1, \cdots, R_N)$ factorization of $\mathcal{A}$ is formulated as finding a lower-rank tensor $\tilde{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ with $\text{rank}_n(\tilde{\mathcal{A}}) = R_n \leq \text{rank}_n(\mathcal{A})$, for all n, such that the following least-squares cost function is minimized:

$$\tilde{\mathcal{A}} = \text{argmin}_{\hat{\mathcal{A}}} \left\| \mathcal{A} - \hat{\mathcal{A}} \right\|. \tag{6}$$

More specifically, $\tilde{\mathcal{A}}$ can be expressed as follows:

$$\tilde{\mathcal{A}} = \mathcal{C} \times_1 U^{(1)} \times_2 U^{(2)} \times \cdots \times_N U^{(N)}, \tag{7}$$

where $U^{(n)} \in \mathbb{R}^{I_n \times R_n}$ has orthonormal columns for $n = 1, \cdots, N$. When R_n is much smaller than I_n for all n, the core tensor $\mathcal{C}$ and the basis matrices $\{U^{(n)}\}_{n=1}^N$ give a compact representation of the original tensor $\mathcal{A}$, resulting in data compression.

Given the basis matrices $\{U^{(n)}\}_{n=1}^N$, the core tensor $\mathcal{C}$ can be readily computed as $\mathcal{C} = \mathcal{A} \times_1 (U^{(1)})^T \cdots \times_N (U^{(N)})^T$. Thus, the optimization problem focuses on the computation of the basis matrices only. An iterative approach can be applied for the computation [25, 43]. Each iterative step optimizes only one of the basis matrices, while keeping the other $N-1$ basis matrices fixed.

The iterative algorithm above may be computationally expensive and the solution depends on the initialization. In this paper, we apply an approximation algorithm called N-mode SVD [41], which has been applied successfully in computer vision and computer graphics [41, 42]. Define D_n as an $I_n \times I_n$ matrix whose (u, v)-th entry $(1 \leq u, v \leq I_n)$ is given by:

$$\sum_{i_1} \cdots \sum_{i_{n-1}} \sum_{i_{n+1}} \cdots \sum_{i_N} a_{i_1 \cdots i_{n-1} u i_{n+1} \cdots i_N} \cdot a_{i_1 \cdots i_{n-1} v i_{n+1} \cdots i_N},$$

where $a_{i_1 \cdots i_{n-1} u i_{n+1} \cdots i_N}$ and $a_{i_1 \cdots i_{n-1} v i_{n+1} \cdots i_N}$ are elements of the Nth-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$. It follows that D_n is a symmetric and positive semi-definite matrix. Let $D_n = U_n \Sigma_n U_n^T$ be the SVD of D_n. Denote $U^{(n)}$ the basis matrix, which consists of the first R_n columns of U_n. The approximation of the original tensor $\mathcal{A}$ is given by

$$\tilde{\mathcal{A}} = \mathcal{C} \times_1 U^{(1)} \times_2 U^{(2)} \times \cdots \times_N U^{(N)}, \tag{8}$$

where $\mathcal{C} = \mathcal{A} \times_1 (U^{(1)})^T \times_2 (U^{(2)})^T \times \cdots \times_N (U^{(N)})^T$.

Since the size of the tensor $\mathcal{A}$ considered in this paper can easily exceed the memory capacity of a single machine, we develop an out-of-core algorithm by partitioning a tensor into smaller blocks as in [43].

3. MULTIPLE DATA SOURCE FUSION

Different neuroimaging features (voxel-based tensor and ROI-based AAL features from the same data source) may capture different but complementary characteristics of the data. For example, the voxel-based tensor features focus more on the global information, while AAL features focus on representative multiple ROI (local) information, though potential information overlaps exist between these two types of data (generated from the same MRI data set). A joint analysis of these data can potentially exploit their complementary information and improve the prediction. Such prediction can be further improved by incorporating additional non-imaging data sources, such as demographic information.

Multiple Kernel Learning (MKL) provides a general framework for learning from multiple data sources [23]. MKL works by first constructing a kernel from each of the data sources and then combining these kernels based on a certain criterion for improved classification performance. In addition to the SVM-based formulation in [23], we apply a discriminant analysis-based formulation.

3.1 Background on Kernel Methods

Kernel methods work by embedding the input data into some high-dimensional feature space and they are generally formulated as convex optimization problems [37, 38]. The key fact underlying the success of kernel methods is that the

embedding into feature space can be determined uniquely by specifying a kernel function that computes the dot product between data points in the feature space implicitly. In other words, the kernel function implicitly defines the nonlinear mapping to the feature space and expensive computations in the high-dimensional feature space can be avoided by evaluating the kernel function in the original attribute space. Thus one of the central issues in kernel methods is the selection of kernels.

We call $K : \mathcal{X} \times \mathcal{X} \to R$ a kernel function [37], where $\mathcal{X}$ is the input space, if it satisfies the finitely positive semidefinite property: for any $x_1, \cdots, x_m \in \mathcal{X}$, the *Gram* matrix $G \in \mathbb{R}^{m \times m}$, defined by $G_{ij} = K(x_i, x_j)$ is symmetric and positive semidefinite. Any kernel function K implicitly maps the input set $\mathcal{X}$ to a high-dimensional (possibly infinite) Hilbert space $\mathcal{H}_K$ equipped with the inner product $(\cdot,\cdot)_{\mathcal{H}_K}$ through a mapping $\phi_K : \mathcal{X} \to \mathcal{H}_K$ as

$$K(x, z) = (\phi_K(x), \phi_K(z))_{\mathcal{H}_K}.$$

In binary classifications, the algorithms learn a classifier $f : \mathcal{X} \to \{-1, +1\}$ whose decision boundary between the two classes is affine in the feature space: $f(x) = \mathrm{sgn}(w^T \phi_K(x) + b)$, where $w \in {}_K$ is the vector of feature weights, $b \in \mathbb{R}$ is the intercept, and $sgn(u) = +1$, if $u > 0$, and -1 otherwise.

3.2 Optimal Kernel Combination

Assume that we are given p kernel matrices $G_1, \cdots, G_p$. In MKL, the optimal kernel matrix G^* lies in the set $\mathcal{G}$ defined as

$$\mathcal{G} = \left\{ G \,\middle|\, G = \sum_{i=1}^{p} \theta_i G_i, \quad \sum_{i=1}^{p} \theta_i \, r_i = 1, \;\; \theta_i \geq 0 \right\}, \qquad (9)$$

where $r_i = \mathrm{trace}(G_i)$. In the following, we assume that all kernel matrices have been centered. This is equivalent to centering the data as the pre-processing step.

For binary-class problems, we are given $\{x_1^+, \cdots, x_{m_+}^+\}$ and $\{x_1^-, \cdots, x_{m_-}^-\}$, the collections of data points from the positive and negative (AD and normal) classes, respectively. The total number of data points is $m = m_+ + m_-$. For a given kernel function K, the basic idea of kernel discriminant analysis with regularization, called RKDA [47] is to find a direction w in the feature space $\mathcal{H}_K$ onto which the projections of the two sets $\{\phi_K(x_i^+)\}_{i=1}^{m_+}$ and $\{\phi_K(x_i^-)\}_{i=1}^{m_-}$ are well separated. Define the centroids of the two classes in the feature space as follows:

$$\mu_K^+ = \sum_{i=1}^{m_+} \phi_K(x_i^+)/m_+, \quad \mu_K^- = \sum_{i=1}^{m_-} \phi_K(x_i^-)/m_-.$$

In RKDA, the separation between the two classes is measured by the ratio of the variance $(w^T(\mu_K^+ - \mu_K^-))^2$ between the classes to the variance $w^T \Sigma_K w$ within the classes, where $\Sigma_K = \phi_K(X)\, P\, \phi_K(X)^T/m$ is the covariance matrix of the data in the feature space and $\phi_K(X)$ is the data matrix in the feature space. Specifically, RKDA in the binary-class case maximizes the following objective function:

$$F(w, K) = (w^T(\mu_K^+ - \mu_K^-))^2 / w^T(\Sigma_K + \mu I)w, \qquad (10)$$

where $\mu > 0$ is a regularization parameter.

It can be shown [47] that for a given set of p centered kernel matrices $G_1, \cdots, G_p$, the optimal kernel matrix $G^* = \sum_{i=1}^{p} \theta_i G_i \in \mathcal{G}$ that optimizes the criterion in Eq. (10) can be found by solving the following Semidefinite Program (SDP):

$$\begin{aligned} \min_{\theta, t} \quad & t \\ \text{subject to} \quad & \begin{pmatrix} I + \frac{1}{\mu}\sum_{i=1}^{p} \theta_i G_i & a \\ a^T & t \end{pmatrix} \succeq 0, \\ & \theta \geq 0,\, \theta^T r = 1, \end{aligned} \qquad (11)$$

where $\theta = [\theta_1, \cdots, \theta_p]^T$, $r = [\mathrm{trace}(G_1), \cdots, \mathrm{trace}(G_p)]^T$, and

$$a = [1/m_+, \cdots, 1/m_+, -1/m_-, \cdots, -1/m_-]^T \in \mathbb{R}^m.$$

The problem can be further formulated as a quadratically constraint quadratic programming (QCQP) problem [48], which is more efficient to solve than SDP.

4. BIOMARKER DETECTION FROM MULTIPLE DATA SOURCES

Recall from the introduction that identifying biomarkers which are sensitive to AD onset or progression [31] is extremely important in AD study. Feature selection is commonly used for selecting a small subset of features for building a comprehensible learning model with good generalization performance [13, 29, 30]. Such a small subset of features can then be used as 'biomarkers'. We propose to apply feature selection from multiple data sources, including the AAL data and the voxel-based tensor data, both from the same single MRI modality, as well as various types of demographic information.

4.1 Background on Feature Selection

Given a data set with d features $\{F_1, F_2, ..., F_d\}$, the task of a feature selection algorithm is to remove as many irrelevant (and redundant) features as it can and find a feature subset $\{F_{j_1}, ..., F_{j_r}\}$ $(r < d)$, such that with dimensionally-reduced data, a learning algorithm can achieve similar or better performance. Feature selection has been used widely in many applications including text mining [11, 17], image processing [12], and bioinformatics [6, 14, 27, 28]. Traditional feature selection algorithms work on a single data source only. The challenge is how to develop effective feature selection algorithms from multiple data sources, called "multi-source feature selection".

4.2 Feature Selection from Multiple Data Sources

Assume that among the p data sources $\{\mathcal{D}_i\}_{i=1}^p$ of m instances (subjects), $\mathcal{D}_t$ $(1 \leq t \leq p)$ is the target for feature (biomarker) selection. In our study, $p = 5$ data sources are involved. In feature selection from multiple data sources, we aim to remove irrelevant (and redundant) features according to the global pattern extracted from all p data sources. Clearly this is different from standard feature selection. To the best of our knowledge, feature selection from multiple data sources has not been well-addressed in the literature.

We propose to use multiple kernel learning for feature selection from multiple data sources. Specifically, multiple kernel learning is applied for information fusion from multiple data sources for pattern extraction. The combined kernel matrix extracts the pattern of the data in the form of pairwise similarities, which can then be used as the input for a

generic feature selection algorithm. We plan to study two feature selection algorithms, SPEC [50] and ReliefF [20], as both algorithms use the pairwise similarities (or distances) and the feature vectors as their input.

SPEC is a framework for both supervised and unsupervised feature weighting [50]. Given a data set $\mathcal{D}$, the similarities among instances can be captured by a set of pairwise instance similarities $\mathbb{S}$ and its induced graph $\mathbb{G}$. SPEC treats features as functions defined on $\mathcal{D}$ and selects features in terms of the smoothness on the manifold formed by the observed data instances. The smoothness of a feature f_i is evaluated by comparing the feature with the spectrums of $\mathcal{L}$, the normalized Laplacian matrix of $\mathbb{G}$:

$$r(f_i) = \varphi\left(f_i; \gamma(\lambda_1), \ldots, \gamma(\lambda_m); \xi_1, \ldots, \xi_m\right). \tag{12}$$

In Eq. (12), $(\lambda_i, \xi_i)_{i=1,\ldots,m}$ denotes the spectrum (or eigen) decomposition of the normalized Laplacian matrix $\mathcal{L}$. $\gamma(\cdot)$ is an increasing function which is used to modify the eigenvalues of $\mathcal{L}$ and has an effect of removing noise [49]. $\varphi(\cdot)$ is a predefined smoothness measure function, which compares the feature with the spectrums of $\mathcal{L}$. In [15, 50], a robust smoothness measure function is defined as:

$$\varphi(f_i) = \frac{\sum_{j=2}^{m} \alpha_j^2 \gamma(\lambda_j)}{\sum_{j=2}^{m} \alpha_j^2} = \frac{\widehat{\boldsymbol{f}}_i{}^T \gamma(\mathcal{L})\, \widehat{\boldsymbol{f}}_i}{1 - \left(\widehat{\boldsymbol{f}}_i{}^T \xi_1\right)^2}; \tag{13}$$

$$\hat{f}_i = \left\| D^{\frac{1}{2}} f_i \right\|^{-1} \cdot \left(D^{\frac{1}{2}} f_i \right), \quad \alpha_j = \xi_j^T \hat{f}_i \tag{14}$$

According to spectral clustering theories [33], the eigenvalues of $\mathcal{L}$ measure the separability of the components of the graph $\mathbb{G}$ and the eigenvectors are the corresponding soft cluster indicators. We compare the normalized feature vector $\hat{f}_i$ with the eigenvectors of $\mathcal{L}$ (measured by α_j's defined in Eq. (14)). The intuition behind Eq. (13) is that the better $\hat{f}_i$ aligns to the leading eigenvectors of $\mathcal{L}$, the better the feature f_i can separate the data as a function defined on $\mathcal{D}$. Note that in Eq. (13), we ignore the first eigenvector of $\mathcal{L}$. The reason is that the trivial eigenvector ξ_1 only carries density information around instances and does not determine separability. In the application, the set $\mathbb{S}$ of pairwise instance similarities can be obtained from the learned kernel matrix from the last section.

ReliefF is a well-known supervised feature selection algorithm derived as an extension of Relief [18]. It determines the relevance of a feature according to its contribution to the hypothesis margin [4] of the observed data. In ReliefF, the relevance of a feature f_i is defined as:

$$r(f_i) = \frac{1}{2} \sum_{t=1}^{m} (\|x_{t,i} - NM(x_t)_i\| - \|x_{t,i} - NH(x_t)_i\|).$$

where $x_{t,i}$ denotes the value of instance x_t on feature f_i, $NH(x)$ and $NM(x)$ denote the nearest points to x in the data with the same and different label respectively, and $\|\cdot\|$ is a distance measurement. In this application, the neighborhoods of instances can be determined by the learned kernel matrix from the last section.

5. EXPERIMENTS

We evaluate the effectiveness of the proposed methods on a collection of 118 samples consisting of 59 normal healthy controls and 59 AD patients.

5.1 Data Sources and Kernels

Five feature (data) sources are used in this study, including tensor and AAL features from MRI images, two types of demographic information related to AD: age and gender, and genetic information based on Apolipoprotein E e4 (APOE4)[5]. It is well known that Apolipoprotein E e4 (APOE4) is a risk factor for AD. Compared to APOE4 non-carriers whose onset age for AD of 84 and risk of 20%, people with one/two copy/copies of APOE4 get the disease at younger age (onset age of 75/68) and increased risk (47%/91%). We derive linear kernels for tensor and AAL features. A Gaussian kernel with an appropriate parameter value is used for the age feature. A simple binary kernel matrix is constructed based on gender feature: if two samples share a common gender, their corresponding kernel matrix entry is 1, otherwise it is set to 0. We use ApoE Genotyping Allele 1 and Allele 2 to divide the samples into three groups: APOE4 non-carriers, heterozygotes, and homozygote groups. A kernel matrix similar to the one for the gender feature is then constructed.

5.2 Performance of Tensor Factorization

We apply tensor factorization based on N-mode SVD for extracting features from the 118 MRI images. The whole collection of images can be represented as a 4th order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4}$ with $I_1 = 181$, $I_2 = 217$, $I_3 = 181$, and $I_4 = 118$. The fourth dimension of the 4th order tensor, which corresponds to the 118 subjects, is fixed (i.e., $R_4 = I_4 = 118$), while the other three dimensions (I_1, I_2, and I_3) which correspond to the size of a single 3D image are reduced to R_1, R_2, and R_3, respectively. Following Eq. (8), the approximation tensor $\tilde{\mathcal{A}}$ is given as

$$\tilde{\mathcal{A}} = \mathcal{C} \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}, \tag{15}$$

where $\mathcal{C} \in \mathbb{R}^{R_1 \times R_2 \times R_3 \times I_4}$ is the core tensor and $U^{(i)} \in \mathbb{R}^{I_i \times R_i}$ is the basis matrix along the i-th dimension. The j-th slice of $\tilde{\mathcal{A}}$ along the fourth dimension is of size $R_1 \times R_2 \times R_3$, which is the compressed representation for the j-th image.

We use the Tensor Toolbox in [5] as the building block for our implementation. The factorization performance is measured in terms of compression ratio and information loss [10, 46]. The information loss (IL) is given by

$$\text{IL} = \frac{\left\| \mathcal{A} - \hat{\mathcal{A}} \right\|}{\|\mathcal{A}\|}, \tag{16}$$

and the compression ratio (CR) is given by

$$\text{CR} = \frac{I_1 I_2 I_3 m}{R_1 R_2 R_3 m + I_1 R_1 + I_2 R_2 + I_3 R_3}, \tag{17}$$

where $m = I_4 = 118$ is the total number of samples.

For simplicity, we set the reduced dimensionalities, i.e, R_1, R_2, and R_3 to a common value in our experiment. The result is summarized in Table 1. We can observe that we achieve a compression ratio of about 263 when the size of each image is reduced to $30 \times 30 \times 30$, while the majority (96%) of the information from the original data is kept. This significantly

[5]Human apolipoprotein E (apoE) is a 34-kDa protein containing 299 amino acid residues. There are three major isoforms of human apoE (namely apoE2, apoE3, and apoE4), which are the products of three alleles (e2, e3, and e4) at a single gene locus on chromosome 19q13.2.

Table 1: Performance of tensor factorization on 118 MRI images.

Reduced Dimension	Compression Ratio	Information Loss
$20 \times 20 \times 20$	889	22%
$30 \times 30 \times 30$	263	4%
$40 \times 40 \times 40$	111	1%

reduces the memory and disk space, which is critical when analyzing and transmitting large volume neuroimaging.

To visually evaluate the compression performance, we randomly pick one image from the data set. We compare the original and reconstructed images (reduced dimension is $20 \times 20 \times 20$) using four slices in three different views (sagittal, coronal, and axial), as shown in Figure 1. We can observe from the figure that the reconstructed slices are visually very similar to the original ones, even with a compression ratio as high as 889. In the following experiment, we set $R_1 = R_2 = R_3 = 20$, resulting in a 8000-dimensional representation for each MRI image, as using a larger dimensionality doesn't improve the performance much.

5.3 Performance of MKL for AD Prediction

In this experiment, we evaluate the multiple kernel learning algorithm for integrating five data sources denoted as tensor, AAL, age, gender, and APOE4. The study is performed by repeated random splitting of the data into training and test sets of ratio of $2:1$. To reduce the variability, the splitting is repeated 20 times and the results are averaged.

Table 2 presents the prediction performance of various algorithms (RKDA and SVM[6] using each of the five data sources by itself and the combination of them based on MKL) in terms of sensitivity and specificity. We can observe from the table that multiple kernel learning based approaches (using the combination of all five data sources), outperform all other methods based on a single data source in terms of both sensitivity and specificity. For example, RKDA-based MKL achieves a sensitivity about 0.950, which is significantly higher than RKDA based on any single data source. We can obtain the same observation in the case of SVM. This implies that different data sources contain complementary information and the integration of them leads to a significant improvement for AD prediction.

5.4 Performance of Biomarker Detection

In this experiment, we evaluate the proposed algorithm for selecting features based on learning from multiple data sources. All five data sources are used to learn the kernel Gram matrix and AAL is used as the target data source with 116 brain regions as the feature set. We also report the feature selection result using AAL data source only for comparison.

Table 3 presents the top 20 regions (features) obtained by two feature selection algorithms: SPEC and ReliefF using G_2 (the kernel matrix constructed from the AAL data source only) and G^* (combined kernel matrix from all five data sources based on RKDA). It is important to note that

[6]We used the LIBSVM implementation in [7] and the SVM-based MKL in [23] with the regularization parameter estimated through 5-fold cross-validation.

Table 2: Performance of MKL based on RKDA and SVM in comparison with RKDA and SVM based on each of the five data sources alone. Prediction in terms of sensitivity and specificity.

Method	Data Source	Sensitivity	Specificity
RKDA	Tensor	0.785	0.790
	AAL	0.605	0.760
	APOE4	0.705	0.415
	Gender	0.740	0.545
	Age	0.505	0.595
	Combination	**0.950**	**0.895**
SVM	Tensor	0.800	0.795
	AAL	0.780	0.845
	APOE4	0.745	0.680
	Gender	0.440	0.460
	Age	0.665	0.515
	Combination	**0.945**	**0.850**

the comparison is based on the assumption that we use pre-existing AD domain knowledge from our collaborators at Banner Alzheimer's Institute at Phoenix as the gold standard. It is clear from Table 3 that both SPEC+G^* and ReliefF+G^* based on MKL perform significantly better than their counterparts SPEC+G_2 and ReliefF+G_2 based on a single data source. For example, among the top 20 regions from SPEC+G^*, 16 of them are confirmed to be AD-related, while there are only 11 AD-related regions from SPEC+G_2. Figure 2 highlights the top 12 regions detected by SPEC+G^*.

Our multiple kernel learning procedure not only provides adequate distinction of AD and normal subjects as shown in Table 2, but also identifies regions that play more significant roles than others in such classification (as shown in Table 3). These brain regions, interestingly enough, included left/right parahippocampla, hippocampus, amygdala, L/R Fusiform, various temporal regions, lingual, and occipital. It is worth noting that our MKL procedure was blind to the prior knowledge of brain regions associated with AD. Nevertheless, the regions that are known to be affected by AD are those that have contributed the most in our MKL analysis. This further confirms the promise of MKL for data fusion for the AD study.

6. DISCUSSIONS AND CONCLUSION

In this paper, we have proposed a kernel method for integrating heterogeneous data for AD prediction. We further extend the kernel framework for selecting features (biomarkers) from heterogeneous data sources. Our experiments show the integration of multiple data sources leads to a considerable improvement in the prediction accuracy. Results also show that the proposed multi-source feature selection algorithm identifies biomarkers (brain regions) that play more significant roles than others in AD diagnosis.

The tensor factorization used in this paper assumes no prior knowledge on the importance of entries from a given tensor. A uniform weight is applied to all entries. In our AD study, certain collections of entries in the brain are known to be more important. It is thus desirable to put higher weights to these voxels. We plan to examine weighted tensor factor-

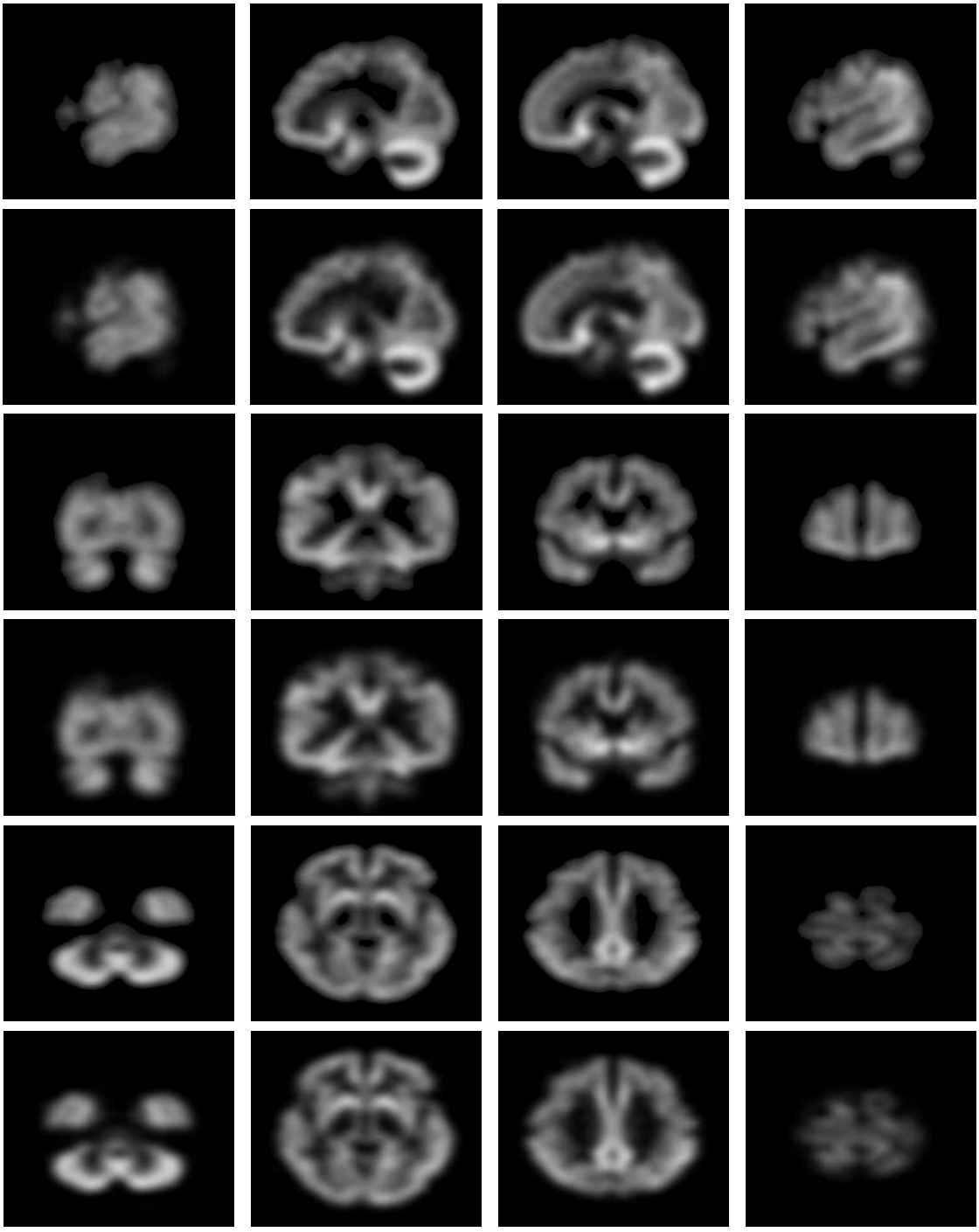

Figure 1: Original and reconstructed images (reduced dimension is $20 \times 20 \times 20$) shown as slices in sagittal view (rows 1 and 2), in coronal view (rows 3 and 4), and in axial view (rows 5 and 6). The first, third, and fifth rows represent 4 slices from the original image, while the second, fourth, and sixth rows represent the corresponding 4 slices from the reconstructed image. Observe that, even with a compression ratio as high as 889, the reconstructed slices are visually very similar to the original ones.

Table 3: The top 20 regions (ranked from top to bottom) from SPEC and ReliefF using G_2 (kernel matrix constructed from the AAL data source alone) and G^* (combined kernel matrix from all five data sources based on RKDA). Regions with bold typeface are relevant according to existing AD domain knowledge.

SPEC+G_2	SPEC+G^*	ReliefF+G_2	ReliefF+G^*
TempPlMidL	**ParaHippR**	**HippR**	**AmygdL**
TempPlMidR	**FusiformR**	**AmygdL**	**AmygdR**
FusiformR	**ParaHippL**	**AmygdR**	**HippR**
FusiformL	**FusiformL**	**HippL**	**HippL**
TempPlSupL	**TempInfR**	Cereb8L	Vermis9
TempInfR	**AmygdR**	Cereb9L	**ParaHippL**
LingualL	**TempInfL**	PutamenR	**ParaHippR**
Cereb6L	**TempPlMidR**	PutamenL	**TempPlSupR**
LingualR	LingualR	Vermis9	**TempPlSupL**
CerebCr1L	**TempPlMidL**	Cereb8R	**TempInfR**
ParaHippR	**OccInfR**	InsulaL	Cereb9L
TempInfL	**TempPlSupL**	PallidumR	Cereb8L
OccInfR	LingualL	PallidumL	Cereb9R
ParaHippL	**OccInfL**	Cereb7bR	Vermis10
TempPlSupR	**AmygdL**	InsulaR	**FusiformL**
Cereb6R	**HippR**	Cereb7bL	**TempPlMidR**
CerebCr1R	**TempPlSupR**	CuneusR	**TempInfL**
Vermis3	**TempMidR**	Vermis8	**TempPlMidL**
Vermis45	CerebCr1L	Cereb9R	Vermis8
CerebCr2L	Cereb6L	**TempMidL**	Vermis7

ization in the future. In this study, we have focused on MRI data. While MRI provides anatomical/structural information about the disease, the complementary PET technique with the positron-emitting radiotracer FDG allows researchers to examine the glucose hypometabolic pattern in AD patients in comparison with normals by measuring the cerebral metabolic rate for glucose (CMRgl). We expect that the fusion of MRI data (structural neuroimaging data) with PET data (functional neuroimaging data) as well as demographic data will further improve the prediction accuracy, and provide a more sensitive measure of longitudinal changes as well as a more powerful indication of any potential treatment/drug evaluations.

Acknowledgment

This research is supported in part by funds from the Arizona State University and the National Science Foundation (NSF) under Grant No. IIS-0612069.

7. REFERENCES

[1] G. Alexander and et al. Regional network of MRI gray matter volume in healthy aging. *Neuroreport*, 17:951–956, 2006.

[2] G. Alexander and J. Moeller. Application of the scaled subprofile model to functional imaging in neuropsychiatric disorder: a principal component approach to modeling brain function in disease. *Human Brain Mapping*, 2(1-2):79–94, 2004.

[3] G. Alexander and E. Reiman. *Neuroimaging*. M.F. Weiner, A.M. Lipton (eds.). The Dementias: Diagnosis, Treatment and Research, 3rd edition, 2003.

[4] R. G. Bachrach, A. Navot, and N. Tishby. Margin based feature selection - theory and algorithms. In *International Conference on Machine Learning (ICML)*, 2004.

[5] B. Bader and T. Kolda. MATLAB Tensor Toolbox Version 2.2. http://csmr.ca.sandia.gov/∼tgkolda/TensorToolbox/, January 2007.

[6] G. C. Cawley and N. L. C. Talbot. Gene selection in cancer classification using sparse logistic regression with bayesian regularization. *BIOINFORMATICS*, 22:2348–2355, 2006.

[7] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/∼cjlin/libsvm`.

[8] K. Chen, E. Reiman, G. Alexander, D. Bandy, R. Renaut, W. Crum, N. Fox, and M. Rossor. An automated algorithm for the computation of brain volume change from sequential MRI's using an iterative principal component analysis and its evaluation for the assessment of whole brain atrophy rates in patients with probable Alzheimer's disease. *Neuroimage*, 22(1):134–143, 2004.

[9] M. Davison. *Multidimensional Scaling*. New York: Wiley, 1983.

[10] C. Ding and J. Ye. 2-Dimensional singular value decomposition for 2D maps and images. In *Proceedings of the Fifth SIAM International Conference on Data Mining (SDM)*, pages 24–34, 2005.

[11] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.

[12] G. Fung and J. Stoeckel. Svm feature selection for classification of spect images of alzheimers disease using spatial information. *Knowledge and Information Systems*, 11:243–258, 2007.

[13] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3):389–422, 2002.

[15] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *Advances in Neural Information Processing Systems 18*. MIT Press, 2005.

[16] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[17] S. S. Keerthi. Generalized lars as an effective feature selection tool for text classification with svms. In *International Conference on Machine Learning (ICML)*, 2005.

[18] K. Kira and L. Rendell. A practical approach to feature selection. In Sleeman and P. Edwards, editors, *ICML '92: Proceedings of the Ninth International Conference on Machine Learning*, pages 249–256. Morgan Kaufmann, 1992.

[19] T. Kolda. Orthogonal tensor decompositions. *SIAM J. Matrix Anal. Appl.*, 23(1):243–255, 2001.

[20] I. Kononenko. Estimating attributes: Analysis and extensions of relief. In *ECML*, page 171182, 1994.

[21] J. Krasaski and et al. Relation of medial temporal lobe volumes to aget and memory function in nondemented

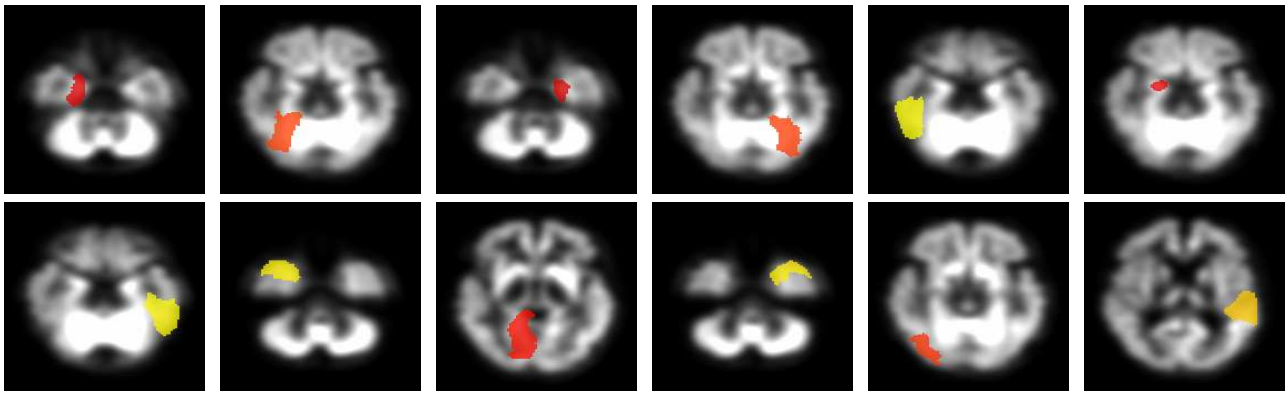

Figure 2: Top 12 regions (highlighted) detected by SPEC+G^* (ranked from left to right and top to bottom). Each region is shown in a different view to enhance the visualy quality.

adults with down's syndrome: implications for the prodromal phase of Alzheimer's disease. *American Journal of Psychiatry*, 159:74–81, 2002.

[22] G. Lanckriet, T. D. Bie, N. Cristianini, M. Jordan, and W. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[23] G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[24] O. Lange, A. Meyer-Baese, M. Hurdal, and S. Foo. A comparison between neural and fuzzy cluster analysis techniques for functional MRI. *Biomedical Signal Processing and Control*, 1(3):243–252, 2006.

[25] L. Lathauwer, B. Moor, and J. Vandewalle. On the best Rank-1 and Rank-(R1,R2,$\cdots$,RN) approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21(4):1324–1342, 2000.

[26] D. Leibovici and R. Sabatier. A singular value decomposition of k-way array for a principal component analysis of multiway data, PTA-k. *Linear Algebra and Its Applications*, 269:307–329, 1998.

[27] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *BIOINFORMATICS*, 20:2429–2437, 2004.

[28] Y. Li, C. Campbell, and M. Tipping. Bayesian automatic relevance determination algorithms for classifying gene expression data. *BIOINFORMATICS*, 18:1332–1339, 2002.

[29] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic Publishers, 1998.

[30] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17:491–502, 2005.

[31] H. Matsuda. Role of neuroimaging in Alzheimer's disease, with emphasis on brain perfusion SPECT. *Journal of Nuclear Medicine*, 48(8):1289–1300, 2007.

[32] S. Molchan. The Alzheimer's Disease Neuroimaging Initiative. *Business Briefing: US Neurology Review*, pages 30–32, 2005.

[33] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *The 14th Advances in Neural Information Processing Systems (NIPS)*, 2001.

[34] K. Nishino, Y. Sato, and K. Ikeuchi. Eigen-texture method: appearance compression based on 3d model. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 618–624, 1999.

[35] D. Pokrajac, V. Megalooikonomou, A. Lazarevic, D. Kontos, and Z. Obradovic. Applying spatial distribution analysis techniques to classification of 3d medical images. *Artificial Intelligence in Medicine*, 33(3):261–280, 2005.

[36] E. Reiman, R. Caselli, G. Alexander, and K. Chen. Tracking the decline in cerebral glucose metabolism in persons and laboratory animals at genetic risk for Alzheimer's disease. *Clinical Neuroscience Research*, 1:194–206, 2001.

[37] B. Schökopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[38] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[39] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[40] N. Tzourio-Mazoyer and et al. Automated anatomical labelling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single subject brain. *Neuroimage*, 15:273–289, 2002.

[41] M. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–460, 2002.

[42] M. Vasilescu and D. Terzopoulos. Tensortextures: multilinear image-based rendering. *ACM Trans. Graph.*, 23(3):336–342, 2004.

[43] H. Wang, Q. Wu, L. Shi, Y. Yu, and N. Ahuja. Out-of-core tensor approximation of multi-dimensional matrices of visual data. *ACM Trans. Graph.*, 24(3):527–535, 2005.

[44] K. Worsley and K. Friston. Analysis of fMRI time series revisited-again. *NeuroImage*, 2:173–181, 1995.

[45] R. Wrede. *Introduction to Vector and Tensor Analysis*. New York: Wiley, 1963.

[46] J. Ye. Generalized low rank approximations of matrices. *Machine Learning*, 61:167–191, 2005.

[47] J. Ye, J. Chen, and S. Ji. Discriminant kernel and regularization parameter learning via semidefinite programming. In *Proceedings of the twenty-fourth International Conference on Machine Learning*, pages 1095–1102, 2007.

[48] J. Ye, S. Ji, and J. Chen. Learning the kernel matrix in discriminant analysis via quadratically constrained quadratic programming. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 854–863, 2007.

[49] T. Zhang and R. Ando. Analysis of spectral kernel design based semi-supervised learning. In *Advances in Neural Information Processing Systems 18*, pages 1601–1608, 2006.

[50] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *International Conference on Machine Learning (ICML)*, 2007.