Neuroinform
DOI 10.1007/s12021-016-9318-5

ORIGINAL ARTICLE

Multi-Domain Transfer Learning for Early Diagnosis of Alzheimer's Disease

Bo Cheng[1] · Mingxia Liu[2,3] · Dinggang Shen[3,4] · Zuoyong Li[5] · Daoqiang Zhang[2,5] · the Alzheimer's Disease Neuroimaging Initiative.

Abstract Recently, transfer learning has been successfully applied in early diagnosis of Alzheimer's Disease (AD) based on multi-domain data. However, most of existing methods only use data from a single auxiliary domain, and thus cannot utilize the intrinsic useful correlation information from multiple domains. Accordingly, in this paper, we consider the joint learning of tasks in multi-auxiliary domains and the target domain, and propose a novel Multi-Domain Transfer Learning (MDTL) framework for early diagnosis of AD. Specifically, the proposed MDTL framework consists of two key components: 1) a multi-domain transfer feature selection (MDTFS) model that selects the most informative feature subset from multi-domain data, and 2) a multi-domain transfer classification (MDTC) model that can identify disease status for early AD detection. We evaluate our method on 807 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database using baseline magnetic resonance imaging (MRI) data. The experimental results show that the proposed MDTL method can effectively utilize multi-auxiliary domain data for improving the learning performance in the target domain, compared with several state-of-the-art methods.

Keywords Transfer learning · Multi-domain · Alzheimer's disease (AD) · Feature selection

Introduction

Alzheimer's Disease (AD) is characterized by the progressive impairment of neurons and their connections, which leads to the loss of cognitive function and the ultimate death. It is reported that an estimated 700,000 elderly Americans will die with AD, and many of them will die from complications caused by AD in 2014 (Association, A.s 2014). Mild cognitive impairment (MCI) is a prodromal stage of AD, where some MCI patients will convert to AD over time, i.e., progressive MCI (pMCI), and other MCI patients remain stable, i.e., stable MCI (sMCI). Thus, for timely therapy that might be effective to slow the disease progression, it is important for early diagnosis of AD and its early stage (i.e., MCI). For the last decades, neuroimaging has been successfully used to investigate the characteristics of neurodegenerative progression in the spectrum between AD and normal controls (NC). In recent years, magnetic resonance imaging (MRI) data are widely applied to early diagnosis of AD, which can measure the structural brain atrophy (Fan et al. 2008; Misra et al. 2009; Risacher et al. 2009). For instance, several studies have shown that AD patients

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

✉ Dinggang Shen
dgshen@med.unc.edu

✉ Daoqiang Zhang
dqzhang@nuaa.edu.cn

the Alzheimer's Disease Neuroimaging Initiative.

[1] Key Laboratory of Advanced Network and Intellectual Technology, Chongqing Three Gorges University, Chongqing 404120, China

[2] Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

[3] Department of Radiology and BRIC, University of North Carolina, Chapel Hill, NC 27599, USA

[4] Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea

[5] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, China

Published online: 07 December 2016

exhibited significant decrease of gray matter volume (Chao et al. 2010; Chetelat et al. 2005; Guo et al. 2010).

Recently, many machine learning methods based on MRI biomarkers have been used for early diagnosis of AD (Cho et al. 2012; Coupé et al. 2012; Cuingnet et al. 2011; Eskildsen et al. 2013; Gaser et al. 2013; Li et al. 2014; Liu et al. 2014; Liu et al. 2016a, b; Ota et al. 2014; Wee et al. 2013; Westman et al. 2013; Westman et al. 2012; Zhang et al. 2016). According to the point of view in the machine learning field, the number of training samples available to build a generalized model is often overwhelmed by the feature dimensionality. In other words, the number of training samples is usually very limited, while the feature dimensionality is very high. This so-called small-sample-size problem has been one of the main challenges in neuroimaging data analysis, which may lead to over-fitting issue (Zhu et al. 2012). To overcome the small-sample-size problem, feature selection has been commonly used in many neuroimaging based studies (Cheng et al. 2015b; Eskildsen et al. 2013; Jie et al. 2015; Liu et al. 2014; Ye et al. 2012; Zhu et al. 2014), where various feature selection methods have been developed to select informative feature subset for reducing the feature dimensionality. Especially, in neuroimaging data analysis for disease diagnosis and therapy, features may be corresponding to brain regions. In such a case, feature selection can detect the regions with brain atrophy, thus potentially useful for timely therapy of brain diseases.

Besides feature selection, many studies have used multimodal data to improve classification performance (Jie et al. 2015; Liu et al. 2014; Ye et al. 2012; Zhang et al. 2012; Zhu et al. 2014). For example, to enhance the generalization of classifiers, some studies have used multi-task learning for multimodal feature selection (Jie et al. 2015; Liu et al. 2014; Zhang et al. 2012; Zhu et al. 2014). In all these studies using multimodal data, different biomarkers are regarded as different modalities, and each modality data is regarded as a learning task (Jie et al. 2015; Liu et al. 2014). On the other hand, several studies have considered each learning approach as a learning task (Zhang et al. 2012; Zhu et al. 2014). All these studies suggest that the use of multimodal data for multi-task learning of features can significantly improve classification performance and enhance generalization performance of classifiers. However, in the clinical practice of AD/MCI diagnosis, the collection of complete multimodal biomarkers from each subject is expensive and time-consuming; on the other hand, it is relatively easy to get single modality data (e.g., MRI) for different categories of subjects. Therefore, in this paper, we address the latter case to build respective classification models for early diagnosis of AD.

According to the pathology of AD, it is the progressive impairment of neurons, and MCI and advanced AD are thus highly related. In this way, several studies suggested that the learning domain of AD diagnosis is related to the learning domain of MCI diagnosis (Cheng et al. 2015b; Coupé et al. 2012; Da et al. 2014; Filipovych et al. 2011; Westman et al. 2013; Young et al. 2013). Also in machine learning community, transfer learning aims to extract the knowledge from one or more auxiliary domains and applies the extracted knowledge to a target domain (Duan et al. 2012; Pan and Yang 2010; Yang et al. 2007), where the auxiliary domain is assumed to be related to the target domain. However, in recent years, several transfer learning methods were developed just for AD/MCI diagnosis (Cheng et al. 2015a; Cheng et al. 2015b; Filipovych et al. 2011; Schwartz et al. 2012; Young et al. 2013). Although these studies suggested that the data from the auxiliary domain can improve the classification performance of target domain, the training data are often from just a single auxiliary domain. Actually, there are multiple auxiliary domain data that can be available in clinical practice. According to the principle of transfer learning, the application of multiple auxiliary domain data could further promote the performance of the target domain.

In addition, in our previous works (Cheng et al. 2015a; Cheng et al. 2015b; Cheng et al. 2012), we mainly consider the prediction of MCI conversion based on a single auxiliary domain data, to construct the respective transfer learning model. Although in our work (Cheng et al. 2015b) we proposed a domain transfer learning method for classification groups such as MCI vs. NC and MCI vs. AD, our proposed method still cannot acquire the deep structured information between the target domain and the auxiliary domain. Furthermore, few studies considered the heterogeneity of MCI to construct semi-supervised classification or regression models (where MCI subjects are regarded as unlabeled samples), which shows that using information of MCI diagnosis can help improve performance of classifying or estimating AD patients from NCs (Cheng et al. 2013; Filipovych et al. 2011; Zhang and Shen 2011). Inspired by the aforementioned issues and successes, in this paper, we assume that there is underlying relationship between each binary classification problem in the early diagnosis of AD, where each binary classification problem can be regarded as target domain, with the other binary classification problems as auxiliary domains. In Fig. 1, we illustrate this novel description of relationships between target domain and corresponding multi-auxiliary domains for early diagnosis of AD. Then, those single modal data that contain multiple data categories can be regarded as multiple related-learning-domains.

In particular, we develop a novel multi-domain transfer learning (MDTL) method for early diagnosis of AD, where training data from multiple auxiliary domains are jointly learned with the target domain. Specifically, we first develop a multi-domain transfer feature selection (MDTFS) model by using the training data from multiple auxiliary domains and target domain to select a subset of discriminative features. Then, we build a multi-domain transfer classifier (MDTC) that can conjointly

apply the training data from multi-auxiliary domains and target domain to construct the classifier. The proposed method is evaluated on the baseline Alzheimer's Disease Neuroimaging Initiative (ADNI) database of 807 subjects with MRI data. The experimental results demonstrate that the proposed method can further improve the performance of early diagnosis of AD, compared with several state-of-the-art methods.

Materials

ADNI Database

The data used in the preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/). ADNI researchers collect, validate and utilize data such as MRI and positron emission tomography (PET) images, genetics, cognitive tests, cerebrospinal fluid (CSF) and blood biomarkers as predictors for Alzheimer's disease. Data from the North American ADNI's study participants, including Alzheimer's disease patients, mild cognitive impairment subjects and elderly controls, are available in this database. In addition, 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 non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether the serial MRI, PET, other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of MCI and early 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.

The ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, aged 55 to 90, to participate in the research approximately

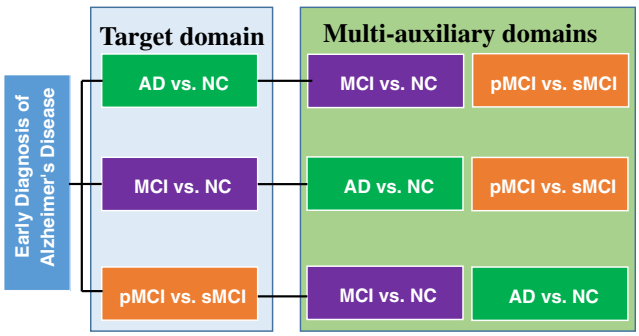


Fig. 1 Our proposed relationships between target domain and auxiliary domains

200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years (see www.adni-info.org for up-to-date information). The research protocol was approved by each local institutional review board, and the written informed consent is obtained from each participant.

Subjects

The ADNI general eligibility criteria are described at www.adni-info.org. Briefly, subjects are between 55 and 90 years of age, having a study partner able to provide an independent evaluation of functioning. Specific psychoactive medications will be excluded. General inclusion/ exclusion criteria are as follows: 1) healthy subjects: MMSE scores between 24 and 30, a Clinical Dementia Rating (CDR) of 0, non-depressed, non-MCI, and non-demented; 2) MCI subjects: MMSE scores between 24 and 30, a memory complaint, having objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia. MCI is a prodromal stage of AD, where some MCI patients will convert to AD, i.e., progressive MCI (pMCI), and other MCI patients remain stable, i.e., stable MCI (sMCI); and 3) Mild AD: MMSE scores between 20 and 26, CDR of 0.5 or 1.0, and meets the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS/ADRDA) criteria for probable AD.

In this work, we focus on using the baseline ADNI database with MRI data. Specifically, the structural MR scans were acquired from 1.5 T scanners. We downloaded raw Digital Imaging and Communications in Medicine (DICOM) MRI scans from the public ADNI website (www.loni.ucla.edu/ADNI), reviewed for quality, and corrected spatial distortion caused by gradient nonlinearity and B_1 field inhomogeneity. More detailed description can be found in (Zhang et al. 2011).

Method

In this section, we first briefly introduce our proposed learning method, and then present our proposed multi-modal transfer feature selection (MDTFS) model, as well as an optimization algorithm for solving the proposed objective function. Finally, we elaborate the proposed multi-domain transfer classification (MDTC) model.

Overview

In Fig. 2, we illustrate the proposed framework for early diagnosis of AD. Specifically, our framework consists of three

main components, i.e., (1) image pre-processing and feature extraction, (2) multi-domain transfer feature selection (MDTFS), and (3) multi-domain transfer classification (MDTC). As shown in Fig. 2, we first pre-process all MR images, and extract features from MR images as described in the Image Preprocessing and Feature Extraction section below. Then, we select informative features via the proposed MDTFS method. We finally build a multi-domain transfer classifier using both the target domain and multi-auxiliary domains data for the classification of AD and MCI.

Image Preprocessing and Feature Extraction

All MR images were pre-processed by first performing an anterior commissure-posterior commissure (AC-PC) correction using the MIPAV software (CIT 2012). The AC-PC corrected images were resampled to $256 \times 256 \times 256$, and the N3 algorithm (Sled et al. 1998) was used to correct intensity inhomogeneity. Then, a skull stripping method (Wang et al. 2011) was performed, and the skull stripping results were manually reviewed to ensure cleaning of skull and dura. The cerebellum was removed by first registering the skull-stripped image to a manually-labeled cerebellum template, and then removing all voxels within the labeled cerebellum mask. FAST in FSL (Zhang et al. 2001) was then used to segment human brain into three different tissues: grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF). We used HAMMER (Shen and Davatzikos 2002) for registration. After registration, each subject image was labeled using the Jacob template (Kabani et al. 1998) with 93 manually-labeled regions-of-interests (ROIs). Then, for each of 93 ROIs, we computed its GM tissue volume as a feature. As a result, for each subject, we have a 93-dimensional feature vector for representing it.

Multi-Domain Transfer Feature Selection (MDTFS)

Unlike previous methods that only considered a single auxiliary domain in model training, in this work, we use samples of target domain as well as multi-auxiliary domains to build a generalized model for feature selection. Hereafter, we denote D as the number of different domains with an index $d \in \{1, 2, \cdots, D\}$ throughout the whole paper. Assume that we have one target domain T, with N_T samples $\mathbf{x}_{T,i}$ and the class labels $\mathbf{y}_{T,i}$, denoted as $\mathrm{T} = \left\{\mathbf{x}_{T,i}, \mathbf{y}_{T,i}\right\}_{i=1}^{N_T}$, where $\mathbf{x}_{T,i} \in \mathbf{R}^F$ is the i-th sample with F features, and $\mathbf{y}_{T,i} \in \{+1, -1\}$ is its corresponding class label. Also, assume that we have $D-1$ auxiliary domains A, with $N_A^{(d)}$ samples $\mathbf{x}_{A,j}^{(d)}$ and the class labels $\mathbf{y}_{A,j}^{(d)}$ for each auxiliary domain, denoted as $A^{(d)} = \left\{\mathbf{x}_{A,j}^{(d)}, \mathbf{y}_{A,j}^{(d)}\right\}_{j=1}^{N_A^d}$, where $\mathbf{x}_{A,j}^{(d)} \in \mathbf{R}^F$ is the j-th sample with F features of the d-th auxiliary domain, and $\mathbf{y}_{A,j}^{(d)} \in \{+1, -1\}$ is the corresponding class label for the d-th auxiliary domain. Therefore, by adding up one target domain and $D-1$ auxiliary domains, we have D domains in total.

In this work, we use a traditional multi-task feature selection method (Obozinski et al. 2006) to design our model for feature selection, and use all the available domain data from the multi-auxiliary domains as well as the target domain to build a more generalized model. Since they are related between the target domain and each auxiliary domain, we need to utilize the intrinsic useful correlation information from multi-auxiliary domain, and introduce an L_2-norm regularizer based on weight vectors (i.e., $\sum_{d=1}^{D-1}\left\|\mathbf{w}_T - \mathbf{w}_A^{(d)}\right\|_2^2$) for different learning domains, which can capture the correlation information between the target domain and multi-auxiliary domains. To learn the common subset of features from all domains (i.e., target domain and all auxiliary domains), we also introduce an L_2/L_1-norm regularizer (i.e., $\|\mathbf{W}\|_{2,1} = \sum_{f=1}^{F}\|\boldsymbol{wf}\|_2$) based on the weight matrix $\mathbf{W}$, where $\boldsymbol{w}^f$ is the f-th row vector of weight matrix $\mathbf{W}$ and is associated with the f-th feature weight across all domains). In addition, to keep the useful decision information of itself, we also use the 'group sparsity' regularization of weight matrix for all domains (i.e., $\|\mathbf{W}\|_{1,1} = \sum_{f=1}^{F}\sum_{d=1}^{D}|w_{f,d}|$). Accordingly, the proposed multi-domain transfer feature selection (MDTFS) model $H(\mathbf{W})$ can be written as follows:

$$H(\mathbf{W}) = \min_{\mathbf{W}} \frac{1}{D}\sum_{d=1}^{D}\left\|\mathbf{y}^{(d)} - \mathbf{x}^{(d)}\mathbf{w}^{(d)}\right\|_2^2 \tag{1}$$

$$+ \lambda_1 \sum_{d=1}^{D-1}\left\|\mathbf{w}_T - \mathbf{w}_A^{(d)}\right\|_2^2 + \lambda_2\|\mathbf{W}\|_{1,1} + \lambda_3\|\mathbf{W}_{2,1}\|$$

where $\mathbf{y}^{(d)} \in \mathbf{R}^{N^{(d)} \times 1}$ is the class label vector of the d-th domain (including target domain and all auxiliary domains), and $\mathbf{x}^{(d)} \in \mathbf{R}^{N^{(d)} \times F}$ is the training dataset of the d-th domain. The 'group sparsity' regularizer matrix $\|\mathbf{W}\|_{1,1}$ $\left(\mathbf{W} \in \mathbf{R}^{F \times D}\right)$ can select a discriminative subset of features relevant to self-domain, and $\|\mathbf{W}\|_{2,1}$ $\left(\mathbf{W} = \left[\mathbf{w}_T, \ldots, \mathbf{w}_A^{(D-1)}\right]\right)$ can select a common subset of features relevant to all domains. The regularization term $\sum_{d=1}^{D-1}\left\|\mathbf{w}_T - \mathbf{w}_A^{(d)}\right\|_2^2$ can control the similarity of multiple weight vectors between the target domain and each auxiliary domain, thus keeping each weight vector of auxiliary domain close to the target domain (Zhou et al. 2013). The column vector $\mathbf{w}_A^{(d)}$ is the d-th auxiliary-domain weight vector, and the column vector $\mathbf{w}_T$ is the target-domain weight vector. In addition, $\lambda_1, \lambda_2, \lambda_3 > 0$ are the regularization parameters that

control the relative contributions of the three regularization terms. By minimizing Eq. (1), we can learn a converged $\mathbf{W}$ from the target domain and multi-auxiliary domain. It is worth noting that, because of using 'group sparsity', the elements of the weight matrix $\mathbf{W}$ will be zero. For feature selection, we just keep those features with non-zero weights.

To solve the optimization problem of Eq. (1), we employ an accelerated gradient descent (AGD) method (Chen et al. 2009; Nemirovski 2005). To be specific, we decompose the objective function of $H(\mathbf{W})$ in Eq. 1 into two parts, i.e., a smooth term $S(\mathbf{W})$ and a non-smooth term $G(\mathbf{W})$, as follows:

$$S(\mathbf{W}) = \frac{1}{D}\sum_{d=1}^{D}\left\|\mathbf{y}^{(d)}-\mathbf{x}^{(d)}\mathbf{w}^{(d)}{}_2^2\right\| + \lambda_1\sum_{d=1}^{D-1}\left\|\mathbf{w}_T-\mathbf{w}_A^{(d)}\right\|_2^2 \quad (2)$$

$$G(\mathbf{W}) = \lambda_2\|\mathbf{W}\|_{1,1} + \lambda_3\|\mathbf{W}\|_{2,1} \quad (3)$$

Then, we define the generalized gradient update rule as follows:

$$Q_h(\mathbf{W},\mathbf{W}_t) = S(\mathbf{W}_t) + \langle \mathbf{W}-\mathbf{W}_t, \nabla S(\mathbf{W}_t)\rangle + \frac{h}{2}\|\mathbf{W}-\mathbf{W}\|_{tF}^2 + G(\mathbf{W})$$

$$q_h(\mathbf{W}_t) = \arg min_{\mathbf{W}}\, Q_h(\mathbf{W},\mathbf{W}_t) \quad (4)$$

where $\nabla S(\mathbf{W}_t)$ denotes the gradient of $S(\mathbf{W})$ at the point $\mathbf{W}_t$ at the t-th iteration, h is a step size, $\langle \mathbf{W}-\mathbf{W}_t, \nabla S(\mathbf{W}_t)\rangle = \mathrm{tr}((\mathbf{W}-\mathbf{W}_t)'\nabla S(\mathbf{W}_t))$ is the matrix inner product, $\|\cdot\|_F$ denotes a Frobenius norm for matrix, and $\mathrm{tr}(\cdot)$ denotes a trace of a matrix. According to (Chen et al. 2009), the generalized gradient update rule of Eq. (4) can be further decomposed into N separate subproblems with a gradient mapping update approach. We summarize the details of AGD algorithm in **Algorithm 1**.

Algorithm 1 AGD algorithm for MDTFS in Eq. (1)

1: Initialization: $h_0 > 0, \eta > 1, \mathbf{W}_0 \in \mathbf{R}^{F\times D}, \overline{\mathbf{W}}_0 = \mathbf{W}_0, h{=}h_0$ and $\alpha_0 = 1$.

2: for $t = 0,1,2,\ldots$ until convergence of $\mathbf{W}_t$ do:

3: Set $h = h_t$

4: while $H\left(q_h(\overline{\mathbf{W}}_t)\right) > Q_h(q_h(\overline{\mathbf{W}}_t), \overline{\mathbf{W}}_t), \quad h = \eta h$

5: Set $h_{t+1} = h$ and compute

$$\mathbf{W}_{t+1} = \arg min_{\mathbf{W}}\, Q_{h_{t+1}}(\mathbf{W}, \overline{\mathbf{W}}_t),$$

$$\alpha_{t+1} = \frac{2}{t+3},\ \beta_{t+1} = \mathbf{W}_{t+1} - \mathbf{W}_t,$$

$$\overline{\mathbf{W}}_{t+1} = \mathbf{W}_{t+1} + \frac{1-\alpha_t}{\alpha_t}\alpha_{t+1}\beta_{t+1}$$

end-while

6: end-for

Multi-Domain Transfer Classification (MDTC)

After performing MDTFS, we can obtain the most discriminative common features, upon which we will build a multi-domain transfer classifier (MDTC) for final classification. Denote $\mathrm{X}_A^{(d)} = \left\{\mathrm{x}_{A,p}^{(d)}, \mathrm{y}_{A,p}\right\}_{p=1}^{N_A^{(d)}}$ and $\mathrm{X}^T = \left\{\mathrm{x}_q, \mathrm{y}_q\right\}_{q=1}^{N_T}$ as the new d-th auxiliary and target domains, with the corresponding labels denoted as $\mathrm{y}_A = \left\{\mathrm{y}_{A,p}\right\}_{p=1}^{N_A^{(d)}}$ and $\mathrm{y} = \left\{\mathrm{y}_q\right\}_{q=1}^{N_T}$, respectively. Here, $\mathrm{x}_{A,p}^{(d)} \in \mathbf{R}^{\overline{F}}$, $\mathrm{x}_q \in \mathbf{R}^{\overline{F}}$, and $\overline{F}$ denote the number of features for new d-th auxiliary and target domain after feature selection (via MDTFS). However, since we use the regularizer of $\|\mathbf{W}\|_{1,1}$ in the MDTFS step, the selected features from each domain are different. For simplicity, we roughly select same feature subset $\overline{F}$ for each auxiliary domain as the target domain.

Unlike our previous work (Cheng et al. 2015b) that only considered a single auxiliary domain in model training of classification. In this work, we will use multi-auxiliary domains for aiding the learning task of target domain. Due to the domain distribution relatedness between the target domain and each auxiliary domain, Yang et al. (Yang et al. 2007) consider that learning a multi-domain transfer classifier $f(\mathrm{x})$ is to learn the "delta function $\Delta f(\mathrm{x})$" between the target and auxiliary classifiers using an objective function similar to SVMs. To combine multi-auxiliary domain classifiers

$f_1^A(x), \ldots, f_d^A(x), \ldots, f_{D-1}^A(x)$, we construct an "ensemble" of auxiliary classifiers $\sum_{d-1}^{D-1} v_d f_d^A(x)$. Then, we employed the A-SVMs model of Yang et al. (Yang et al. 2007) to get the multi-domain transfer classifier, which has the following form:

$$\boldsymbol{f}(x) = \sum_{d=1}^{D-1} v_d f_d^A(x) + \Delta f(x) = \sum_{d=1}^{D-1} v_d f_d^A(x) + \boldsymbol{u}' \boldsymbol{\Phi}(x) \quad (5)$$

where $v_d \in (0, 1)$ is the weight of each auxiliary classifier $f_d^A(x)$, which sums to one as $\sum_{d=1}^{D-1} v_d = 1$. Also, $\boldsymbol{\Phi}(x)$ is a kernel-based mapping function, and $\boldsymbol{u}$ is the weight vector of target domain classifier. In addition, $\boldsymbol{u}'$ denotes the transpose of $\boldsymbol{u}$.

To learn the weight vector $\boldsymbol{u}$ in Eq. 5, we use the following objective function, similar to the SVM (Yang et al. 2007):

$$\begin{aligned} &\min_{u} \frac{1}{2}\|\boldsymbol{u}\|^2 + C\sum_{l=1}^{N_T} \beta_l \\ &s.t. \quad \beta_l \geq 0,\ y_l\, \boldsymbol{u}'\, \Phi(x_l) + y_l \sum_{d=1}^{D-1} v_d f_d^A(x_l) \geq 1 - \beta_l \end{aligned} \quad (6)$$

where l is the l-th sample in the target domain training subset $(x_l, y_l) \in X^T$, and β_l is the slack variable that represents the prediction error of objective function of Eq. 6, thus it can be used for nonlinear classification. The parameter C balances contributions between auxiliary classifier and target-domain training samples. According to (Yang et al. 2007), we can solve this objective function in Eq. 6 to obtain the solution for the weight vector $\boldsymbol{u}$. Then, we can obtain the final solution for $\boldsymbol{f}(x)$. In this study, $f_d^A(x)$ is trained by SVM, and $\Delta f(x)$ is solved by Eq.5 using the method of kernel learning.

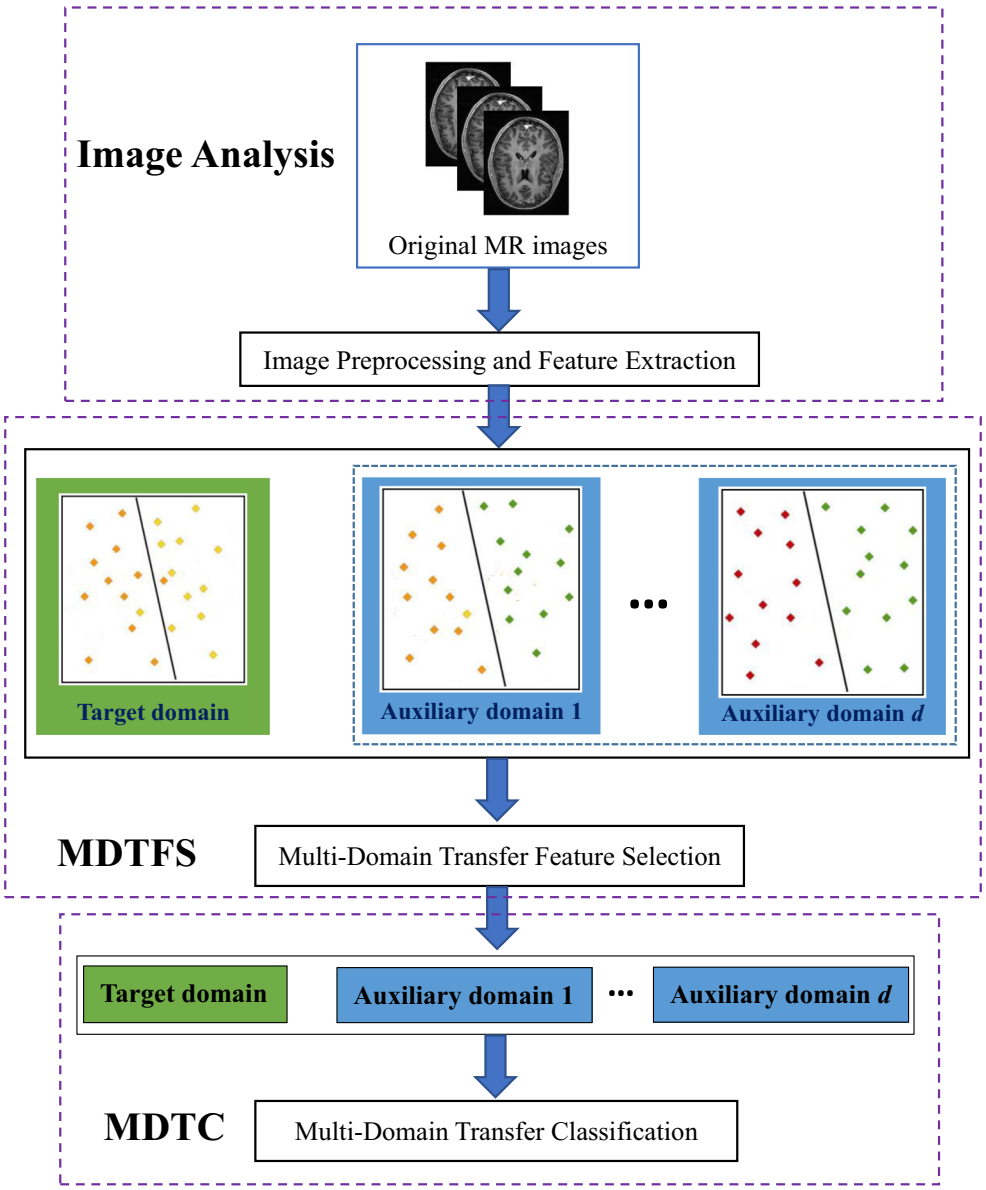


Fig. 2 Summary of our proposed framework for early diagnosis of AD using multi-domain transfer learning (MDTL) method

Results

In this section, we first describe experimental settings in our experiments. Then, we show the classification results on the ADNI database by comparing our proposed method with several state-of-the-art methods. In addition, we illustrate the most discriminative brain regions identified by our proposed method.

Experimental Settings

We use the samples of 807 subjects (186 AD, 395 MCI, and 226 NC), for whom the baseline MRI data were all available. It is worth noting that, for all 395 MCI subjects, during the 24-month follow-up period, 167 MCI subjects converted to AD (pMCI for short) and 228 MCI subjects remained stable (sMCI for short). In addition, we consider three binary classification problems, i.e., AD vs. NC classification, MCI vs. NC classification, and pMCI vs. sMCI classification. For our proposed multi-modal transferring method, we explicitly list the target domain and the corresponding auxiliary domains for each classification task in Table 1.

In the experiments, we adopt a 10-fold cross-validation strategy to partition the target domain data into training and testing subsets. In particular, the target domain samples of each classification problem is partitioned into 10 subsets (each subset with a roughly equal size), and then one subset was successively selected as the testing samples and all the remaining subsets were used for training. To avoid the possible bias occurred during sample partitioning, we repeat this process 10 times. We report the average performances in terms of area under the receiver operating characteristic curve (AUC), accuracy (ACC), sensitivity (SEN), and specificity (SPE).

We compared the proposed method with a standard SVM (SVM for short), Lasso (Tibshirani 1996), MTFS (Zhang et al. 2012), and M2TFS (Jie et al. 2015). These methods are listed as follows.

- **SVM**: training samples only from the target domain, and without any feature selection before classification stage;
- **Lasso**: training samples only from the target domain, and the Lasso method conducted for feature selection before using SVM for classification;
- **MTFS** and **M2TFS**: training samples from the target and multi-auxiliary domains, and the MTFS and M2TFS methods conducted for feature selection before using the selected classification method in the literatures (Jie et al. 2015; Zhang et al. 2012).

The SVM method is implemented using the LIBSVM toolbox (Chang and Lin 2001) with a linear kernel and a default value for the parameter C. Also, other competing methods with the SVM for classification are implemented using the LIBSVM toolbox, with the same settings of parameters as the SVM method. For the Lasso and MTFS methods, we adopt the SLEP toolbox (Liu et al. 2009) to solve the optimization problem. In addition, we employ the accelerated proximal gradient (APG) method in the literature (Chen et al. 2009) to solve the optimization problem of M2TFS. There are multiple regularization parameters of these methods (including Lasso, MTFS, M2TFS, and proposed MDTL) to be optimized. All regularization parameters of these methods are chosen from the range of $\boldsymbol{P}$[1] by a nested 10-fold cross-validation on the training data. In the proposed MDTL frame, the weight v_d of auxiliary classifier $f_d^A(\mathrm{x})$ for MDTC is learned within a nested 10-fold cross-validation via a grid search in the range of 0 and 1 at a step size of 0.1, and adopted the SVM based linear kernel for training the target-domain and auxiliary-domain classifiers. Before training models, we normalized features following (Zhang et al. 2011).

Comparison between MDTL and Other Methods

To investigate the effectiveness of the proposed method, we compare the proposed method with several state-of-the-art methods. Table 2 shows the classification results achieved by six methods, including SVM (traditional SVM), Lasso, MTFS (Zhang et al. 2012), M2TFS (Jie et al. 2015), and the proposed method (i.e., MDTL and MDTC). In Table 2, the proposed 'MDTL' method first performs the MDTFS for feature selection and then adopts MDTC for classification, while the 'MDTC' method performs only MDTC for classification. Also, note that each value in Table 2 is the averaged result of the 10-fold cross validation, which was performed for ten different times. In addition, we plot the ROC curves achieved by these six methods in Fig. 3.

As can be seen from Table 2 and Fig. 3, for three binary classification problems, the proposed MDTL method consistently outperforms SVM, Lasso, MDTC, MTFS and M2TFS in terms of the classification accuracy, sensitivity, and AUC measures. We also perform DeLong's method (DeLong et al. 1988) on the AUC between the proposed method and each of other five competing methods, with the corresponding p-values shown in Table 2. The DeLong's test is a nonparametric statistical test for comparing AUC between two ROC curves, which can be employed to assess statistical significance by computing z-scores for the AUC estimate (Robin et al. 2011; Sabuncu et al. 2015). For both AD vs. NC and MCI vs. NC classification tasks, the proposed MDTL method consistently outperforms the competing methods in all classification measures. In

[1] $\mathbf{P} \in \{0.000001, 0.00001, 0.0001, 0.0003, 0.0007, 0.001, 0.003, 0.005, 0.007, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$

Table 1 Target domains and corresponding multiple auxiliary domains used in three binary classification tasks. The number in each bracket denotes the class label, where +1 denotes positive class and −1 represents negative class.

Classification problem	Target domain	Auxiliary domain
AD vs. NC	AD(+1) vs. NC(−1)	1) MCI(+1) vs. NC(−1), 2) pMCI(+1) vs. sMCI(−1)
MCI vs. NC	MCI(+1) vs. NC(−1)	1) AD(+1) vs. NC(−1), 2) pMCI(+1) vs. sMCI(−1)
pMCI vs. sMCI	pMCI(+1) vs. sMCI(−1)	1) AD(+1) vs. NC(−1), 2) MCI(+1) vs. NC(−1)

pMCI vs. sMCI classification, the proposed MDTL method outperforms the competing methods except for the specificity. Also, in Fig. 3, we can see from the ROC shown for pMCI vs. sMCI classification, which implies that the MDTL method can achieve better diagnostic performance in recognizing pMCI and sMCI patients than the competing methods. From the results in Table 2 and Fig. 3, it is clear that the proposed MDTL method can effectively integrate information of target domain and multi-auxiliary domains, which can achieve more significant performance improvement than the methods that use samples only from the target domain for training.

On the other hand, in Table 2 and Fig. 3, the proposed MDTC method consistently outperforms the SVM method in all classification measures for three binary classification problems. Also, there are slight differences of performance between the MDTC and Lasso method for three classification problems. These results imply that, compared with the case of only using SVM for performing classification, using MDTC can also improve the diagnostic performance, similar to the case of using the Lasso method for feature selection. We can see from Table 2 and Fig. 3, Lasso, MTFS, M2TFS, and MDTL methods also outperform the SVM method in all classification measures for three classification problems, which suggest that using feature selection on the high-dimensional features before performing classification can effectively improve the classification performance. In addition, from Table 2 and Fig. 3, MTFS, M2TFS, and MDTL methods can achieve better classification performance than the Lasso method, and the MDTL method also outperforms both the MTFS and M2TFS methods. These results also suggest that the inclusion of multi-auxiliary domains can improve the classification performance compared to the case of only using target domain, and that our proposed regularization factor based on multi-domain weight vector is more suitable than the manifold regularization factor for the transfer learning problem.

In addition, there is an interesting observation from Table 2 and Fig. 3. Specifically, different from conventional studies (Cheng et al. 2012; Coupé et al. 2012; Da et al. 2014; Young et al. 2013), using pMCI and sMCI subjects as auxiliary domain can also help improve the performance of AD and NC classification. The main reason for this observation is that we proposed a regularizer (i.e., $\sum_{d=1}^{D-1}\left\|\mathbf{w}_T-\mathbf{w}_A^{(d)}\right\|_{2,2}$) in step of MDTFS, which can use the weight vector from each of the multi-auxiliary domains to adjust the weight vector of target domain, and combine L_2/L_1-norm and L_1/L_1-norm regularizers to select features relevant to all domains (including self-domain), followed by using the MDTC based linear kernel SVM to keep these selected helpful features for classification. Furthermore, our proposed MDTFS model can also keep the target domain as the most important task in classification. Therefore, our proposed MDTL method can effectively use related multi-auxiliary domain data to improve the performance of target learning domain in early diagnosis of AD.

Comparison with MDTL and Other Variants

To investigate the relative contributions of the two components (i.e., MDTC and MDTFS) in our proposed method, we compare our method with its two variant methods. In Table 3, we give the classification measures by our proposed MDTL method, its variant methods (MDTC and MDTFS), and SVM (as a baseline method). Note that the proposed 'MDTL' method first performs feature selection using MDTFS model and then adopts MDTC for classification (i.e., MDTFS + MDTC, while the 'MDTC' method only performs classification using the proposed MDTC model. The 'MDTFS' method first performs feature selection using MDTFS model and then adopts SVM for classification. In Fig. 4, we also plot the ROC curves achieved by different methods. In addition, we also report the *p*-values, which are computed by DeLong's method (DeLong et al. 1988) on the AUC between the proposed method and its two variant methods, as well as baseline method, in Table 3. From Table 3 to Fig. 4, we can observe that each component can boost the classification performance compared with SVM method. However, using feature selection method (i.e., MDTFS) can achieve better improvement than the MDTC method for classification. In general, our proposed MDTL method that integrates all the two components together achieves the best performance.

Discriminative Brain Regions Detection

To evaluate the efficacy of our proposed multi-domain transfer feature selection (MDTFS) method in detecting the discriminative brain regions, we compare our proposed MDTFS method with the single-domain based feature selection method (i.e., Lasso) and the commonly used multi-domain based feature selection methods (i.e., MTFS and M2TFS). Table 4 shows the classification performances of four different methods, including

Table 2 Comparison of our proposed methods (MDTL and MDTC) and other 4 state-of-the-art methods (SVM, Lasso, MTFS and M2TFS) in three binary classification problems.

AD vs. NC Classification					
Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	*p*-value
SVM	85.8	84.6	85.9	0.933	<0.0001
MDTC	88.4	87.2	88.5	0.950	<0.0001
Lasso	87.9	87.8	88.1	0.951	<0.0001
MTFS	90.7	89.5	90.8	0.966	<0.001
M2TFS	91.5	91.4	91.6	0.979	<0.005
MDTL	**94.7**	**94.1**	**94.8**	**0.988**	-
MCI vs. NC Classification					
Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	*p*-value
SVM	72.5	78.8	60.4	0.769	<0.0001
MDTC	75.8	81.4	65.1	0.810	<0.0001
Lasso	74.1	80.1	62.7	0.787	<0.0001
MTFS	78.1	83.2	68.5	0.849	<0.0005
M2TFS	78.6	83.6	69.1	0.870	<0.005
MDTL	**81.5**	**85.8**	**73.3**	**0.882**	-
pMCI vs. sMCI Classification					
Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	*p*-value
SVM	67.1	61.0	71.5	0.728	<0.0001
MDTC	69.2	63.5	73.3	0.742	<0.0001
Lasso	69.4	63.8	73.5	0.744	<0.0001
MTFS	71.0	62.7	74.9	0.757	<0.0005
M2TFS	71.4	65.8	**77.8**	0.768	<0.001
MDTL	**73.8**	**69.0**	77.4	**0.796**	-

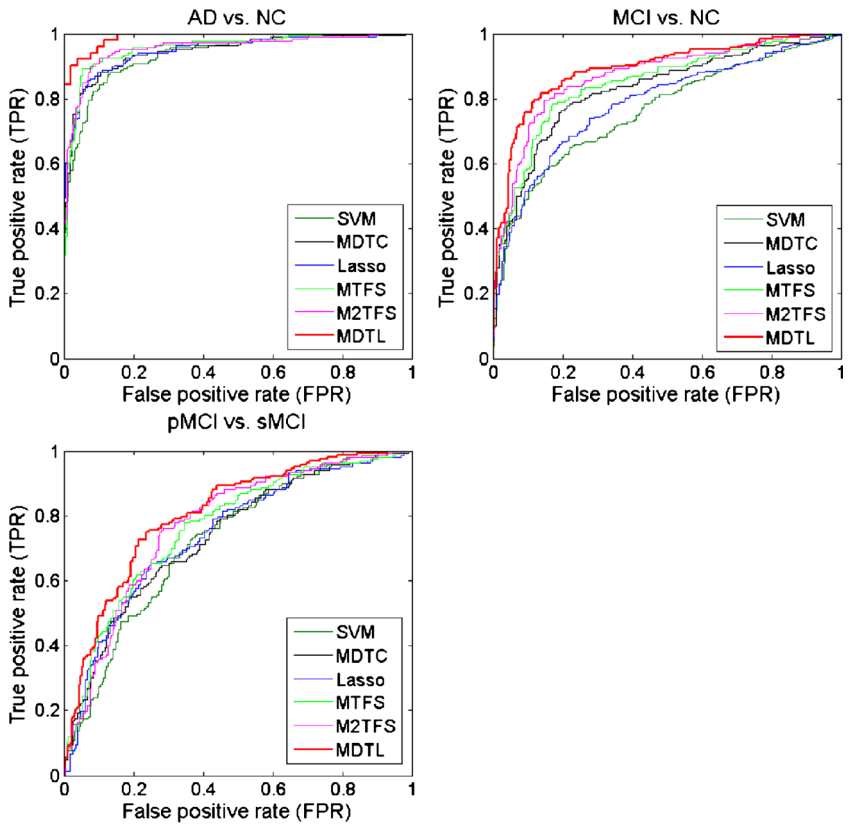


Fig. 3 ROC curves of different methods for three binary classification problems

Table 3 The comparison of our proposed methods (MDTL), its two variant methods (MDTC and MDTFS), and SVM (as a baseline method) in three binary classification problems.

AD vs. NC Classification					
Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	*p*-value
SVM	85.8	84.6	85.9	0.933	<0.0001
MDTC	88.4	87.2	88.5	0.950	<0.0001
MDTFS	93.4	93.3	93.5	0.982	<0.05
MDTL	**94.7**	**94.1**	**94.8**	**0.988**	-
MCI vs. NC Classification					
Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	*p*-value
SVM	72.5	78.8	60.4	0.769	<0.0001
MDTC	75.8	81.4	65.1	0.810	<0.0001
MDTFS	80.2	84.8	71.4	0.876	<0.05
MDTL	**81.5**	**85.8**	**73.3**	**0.882**	-
pMCI vs. sMCI Classification					
Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	*p*-value
SVM	67.1	61.0	71.5	0.728	<0.0001
MDTC	69.2	63.5	73.3	0.742	<0.0001
MDTFS	72.8	67.8	76.4	0.792	<0.05
MDTL	**73.8**	**69.0**	**77.4**	**0.796**	-

Lasso, MTFS (Zhang et al. 2012), M2TFS (Jie et al. 2015), and the proposed MDTFS, using classification accuracy, sensitivity, specificity and AUC measures. In addition, we also compute *p*-values on the AUC between the MDTFS method and other three methods via DeLong's method (DeLong et al. 1988), as also shown in Table 4. It is worth noting that, for fair comparison, we use SVM on the target domain in the classification step for our method and competing methods. Also, each value in Table 4 is the averaged result of 10-fold cross-validation strategy in 10 independent runs. As shown in Table 4, MDTFS, MTFS and M2TFS methods can achieve better classification performance than the Lasso method. The possible

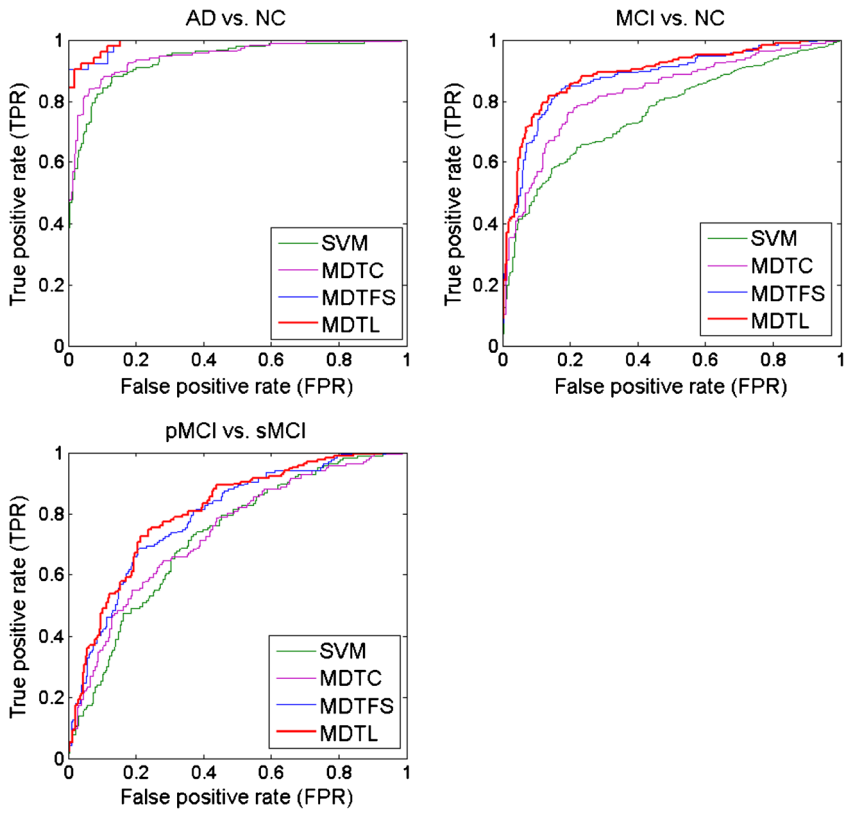


Fig. 4 ROC curves of different methods for three binary classification problems

Table 4 Comparison of our proposed feature selection method (MDTFS) and other state-of-the-art feature selection methods (Lasso, MTFS and M2TFS) in three binary classification problems.

AD vs. NC Classification					
Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	*p*-value
Lasso	87.9	87.8	88.1	0.951	<0.0001
MTFS	90.7	89.5	90.8	0.966	<0.001
M2TFS	91.5	91.4	91.6	0.979	<0.01
MDTFS	**93.4**	**93.3**	**93.5**	**0.982**	-
MCI vs. NC Classification					
Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	*p*-value
Lasso	74.1	80.1	62.7	0.787	<0.0001
MTFS	78.1	83.2	68.5	0.849	<0.0005
M2TFS	78.6	83.6	69.1	0.870	<0.01
MDTFS	**80.2**	**84.8**	**71.4**	**0.876**	**-**
pMCI vs. sMCI Classification					
Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	*p*-value
Lasso	69.4	63.8	73.5	0.744	<0.0001
MTFS	71.0	62.7	74.9	0.757	<0.0005
M2TFS	71.4	65.8	**77.8**	0.768	<0.001
MDTFS	**72.8**	**67.8**	76.4	**0.792**	**-**

reason could be that MDTFS, MTFS and M2TFS use data from multi-auxiliary domains. On the other hand, our proposed MDTFS method outperforms MTFS and M2TFS methods, suggesting that our method can better capture useful information between the target domain and multi-auxiliary domains.

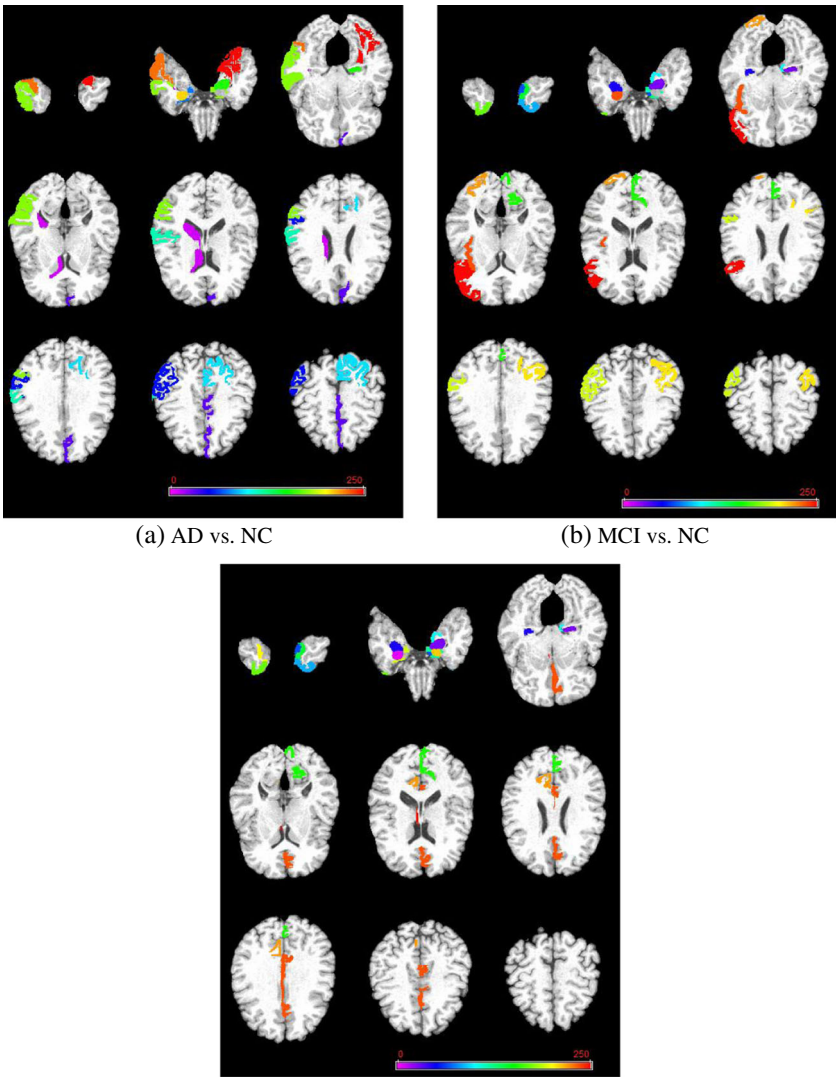
Furthermore, we also investigate the most discriminative regions identified by the proposed feature selection method. Since the feature selection in each fold was performed only based on the training set, the selected features could vary across different cross-validations. We thus defined the most discriminative brain regions based on the selected frequency of each region over the cross-validations. In Fig. 5, for three classification problems, we list all selected brain regions with the highest frequency of occurrence (i.e., each feature and selected across all folds and all runs) by MDTL (i.e., MDTFS + MDTC) on template MR image. As can be seen from Fig. 5, our proposed MDTL method successfully finds out the most discriminative brain regions (e.g., amygdala, hippocampal formation, entorhinal cortex, temporal pole, uncus, perirhinal cortex, cunecus, and temporal pole) that are known to be related to Alzheimer's disease (Davatzikos et al. 2011; Eskildsen et al. 2013; Jie et al. 2015; Ye et al. 2012; Zhang et al. 2012; Zhang et al. 2011; Zhu et al. 2014).

Discussion

In this paper, we propose a multi-domain transfer learning (MDTL) method for early diagnosis of AD, in which we combine the data from multi-auxiliary domains and target domain in both feature selection and classification steps. We evaluate the performance of our method on 807 subjects from the publicly available ADNI database and compare our method with the state-of-the-art methods. The experimental results show that the proposed method consistently and substantially improved the performance of early diagnosis of AD.

Learning with Multi-Domain Data

In the field of neuroimaging-based early diagnosis of AD, multimodal biomarker is widely used for the model design of feature selection and classification (Cheng et al. 2015a; Cheng et al. 2015b; Hinrichs et al. 2011; Jie et al. 2015; Liu et al. 2014; Suk et al. 2014; Ye et al. 2012; Zhang et al. 2012; Zhang et al. 2011; Zhu et al. 2014), which can achieve better performance than the case of using single-modal biomarker. However, in clinical practice, the collection of multimodal biomarker from subject is expensive and time-consuming, and hence the size of collected complete multimodal biomarker dataset is often small. On the other hand, it is relatively easy to get more single-modal biomarker data (e.g., MRI) that contain different categories of subjects. Because of the characteristic of AD cohorts and the requirement of the clinical diagnosis of AD, these single modal data are classified as multiple learning domains that are related to each other. Some studies also show the effectiveness of transfer learning or semi-supervised learning technique in combining these data from related learning domains (Cheng et al. 2012; Da et al. 2014; Filipovych et al. 2011; Young et al. 2013; Zhang and Shen 2011). However, auxiliary data from a single related learning domain is often used in the aforementioned studies. In this paper, we developed MDTL method to enhance the generalization and accuracy of



(a) AD vs. NC (b) MCI vs. NC

(c) pMCI vs. sMCI

Fig. 5 The most discriminative brain regions identified by the proposed MDTL method for three classification tasks. Note that different colors indicate different brain regions

classifiers for the case of single modal data with multiple related learning domains,.

To integrate these data with multiple categories of subjects, the transfer learning can be used to build the learning model as done in our previous works (Cheng et al. 2015a; Cheng et al. 2015b; Cheng et al. 2012). However, in our previous works, we only adopted a single related domain data as auxiliary domain to help design classification model for MCI conversion prediction. For example, in our work (Cheng et al. 2015b), we gave an explanation that the domain of classifying pMCI and sMCI subjects was related to the domain of classifying AD and NC subjects, but it was only used for MCI

Table 5 Comparison of our proposed method (MDTL), SDTL1, SDTL2, and Lasso methods in three binary classification problems. The 'SDTL1' is the MDTL method using one auxiliary domain data, which actually is the first one in the auxiliary domains of Table 1. The 'SDTL2' is the MDTL method using one auxiliary domain data, which is the second one in the auxiliary domains of Table 1.

Method	AD vs. NC		MCI vs. NC		pMCI vs. sMCI	
	Accuracy (%)	AUC	Accuracy (%)	AUC	Accuracy (%)	AUC
Lasso	87.9	0.951	74.1	0.787	69.4	0.744
SDTL1	93.0	0.978	80.2	0.876	72.6	0.782
SDTL2	92.2	0.969	79.0	0.857	72.1	0.786
MDTL	**94.7**	**0.988**	**81.5**	**0.882**	**73.8**	**0.796**

Table 6 Comparison of our proposed method (MDTL) in three binary classification problems using different setting of regularization parameters.

Regularization parameter	AD vs. NC		MCI vs. NC		pMCI vs. sMCI	
	Accuracy (%)	AUC	Accuracy (%)	AUC	Accuracy (%)	AUC
$\lambda_1 = 0$	93.3	0.985	79.7	0.871	72.1	0.786
$\lambda_2 = 0$	93.7	0.986	80.6	0.875	73.1	0.792
$\lambda_3 = 0$	92.7	0.984	79.6	0.871	72.3	0.787
$\lambda_1, \lambda_2, \lambda_3 > 0$	**94.7**	**0.988**	**81.5**	**0.882**	**73.8**	**0.796**

conversion prediction. According to the pathology of AD and its progression, we extended our previous work (Cheng et al. 2015b), by assuming that the classification problem in the target domain is related to each classification problem in the auxiliary domains. To validate this assumption, we have listed results of our proposed method (MDTL), as well as SDTL1, SDTL2, and Lasso methods, in Table 5. For purpose of comparison, we used the Lasso as a baseline method. Also, SDTL1 and SDTL2 methods are the variants of our proposed MDTL method performed on single auxiliary domain and target domain. As we can see from Table 5, using multi-auxiliary domain data in the MDTL method can achieve better performance than using single auxiliary domain data in both SDTL1 and SDTL2 methods, and the baseline method (i.e., Lasso) is inferior to MDTL, SDTL1 and SDTL2 methods. These results make it clear that the use of multi-auxiliary domain can effectively improve the performance of early diagnosis of AD.

Multi-Domain Transfer Learning Model

To effectively integrate the domain knowledge between target domain and multi-auxiliary domain, we introduced the regularizer based on weight vector (i.e., $\sum_{d=1}^{D-1} \left\| \mathbf{w}_T - \mathbf{w}_A^{(d)} \right\|_{2,2}$), which can keep the similarity of multi-weight vectors between target domain and each auxiliary domain. In addition, we introduced two regularizations of weight matrix for all domains (i.e., $\|\mathbf{W}\|_{1,1}$ and $\|\mathbf{W}\|_{2,1}$) simultaneously, which can select a common feature subset relevant to all domains and also keep useful features relevant to self-domain. In this paper, we combined the three regularizers for feature selection from target-domain and multi-auxiliary-domain data, namely Multi-Domain Transfer Feature Selection (MDTFS). To evaluate the efficacy of each regularizer, we performed some experiments for testing the contribution of each regularizer. In Table 6, we give results of the proposed MDTL method for three classification problems using different setting of regularization parameters.

In the MDTFS model, the regularization parameters (i.e., $\lambda_1, \lambda_2, \lambda_3$) can control the relative contribution of the three regularizers. In Table 6, we investigate the contribution of each regularizer by setting the respective parameter to zero. For example, we set the regularization parameter λ_1 to zero (i.e., $\lambda_1 = 0$), which is used for evaluating contribution of the first regularizer. As can be seen from Table 6, combining three regularizers (i.e., $\lambda_1, \lambda_2, \lambda_3 \neq 0$) can achieve better performance for early diagnosis of AD. Specifically, for three classification problems, the minimum reduction of classification performance is without use of the second regularizer (i.e., $\lambda_2 = 0$), while the reduction of classification performance is small compared to the case without using the first regularizer (i.e., $\lambda_1 = 0$) and the case without the third regularizer (i.e., $\lambda_3 = 0$). These results suggest the importance of selecting the common feature subset relevant to all domains and keeping the similarity of weight vectors between target domain and each auxiliary domain, which also confirms the efficacy of using multi-auxiliary domain data.

Strategy for Selecting Feature Subset

Recently, many studies in early diagnosis of AD focus on designing feature selection methods to overcome the small-sample-size problem in neuroimaging data analysis (Cheng et al. 2015b; Li et al. 2014; Liu et al. 2014; Moradi et al. 2015; Ota et al. 2015; Ye et al. 2012; Zhu et al. 2014). In this paper, we develop a multi-domain transfer learning (MDTL)

Table 7 Comparison of two different strategies in selecting feature subset from each domain in our proposed MDTL method. AUC: area under the receiver operating characteristic curve; ACC: accuracy.

Strategy	AD vs. NC			MCI vs. NC			pMCI vs. sMCI		
	ACC (%)	AUC	*p*-value	ACC (%)	AUC	*p*-value	ACC (%)	AUC	*p*-value
Strategy1	94.7	0.988	>0.05	81.5	0.882	>0.05	73.8	0.796	>0.05
Strategy2	94.8	0.989	-	82.0	0.886	-	74.2	0.798	-

Table 8 Comparison with the state-of-the-art methods in early diagnosis of AD. AUC: area under the receiver operating characteristic curve; ACC: accuracy; SEN: sensitivity; SPE: specificity.

Reference	Data	Feature extraction	Result
Our proposed method	186 AD, 226 NC 167 pMCI, 228 sMCI	ROIs of GM	AD vs. NC ACC = 94.7 %, AUC = 0.988 MCI vs. NC ACC = 81.5 %, AUC = 0.882 pMCI vs. sMCI ACC = 73.8 %, AUC = 0.796
Hu et al. 2016	188 AD, 228 NC 71 pMCI, 62 sMCI	VBM of GM, WM, and CSF	AD vs. NC ACC = 84.13 %, AUC = 0.9 pMCI vs. sMCI ACC = 76.69 %, AUC = 0.79
Moradi et al. 2015	200 AD, 231 NC 164 pMCI, 100 sMCI	VBM of GM VBM of GM, age and cognitive measures	pMCI vs. sMCI ACC = 74.74 %, AUC = 0.7661 ACC = 82 %, AUC = 0.9
Khedher et al. 2015	188 AD, 229 NC 401 MCI	VBM of GM, and WM	AD vs. NC ACC = 88.49 % MCI vs. NC ACC = 81.89 % ACC = 77.57 % (only GM)
Ota et al. 2015	40 pMCI, 40 sMCI	VBM of whole brain	pMCI vs. sMCI AUC = 0.75
Zhu et al. 2014	51 AD, 52 NC 43 pMCI, 56 sMCI	ROIs of GM	AD vs. NC (Only MRI) ACC = 93.8 %, AUC = 0.979 MCI vs. NC (Only MRI) ACC = 79.7 %, AUC = 0.852 pMCI vs. sMCI (Only MRI) ACC = 70.8 %, AUC = 0.756
Liu et al. 2014	51 AD, 52 NC 43 pMCI, 56 sMCI	ROIs of GM	AD vs. NC (PET + MRI) ACC = 94.37 %, AUC = 0.9724 MCI vs. NC (PET + MRI) ACC = 78.8 %, AUC = 0.8284 pMCI vs. sMCI (PET + MRI) ACC = 67.83 %, AUC = 0.6957
Eskildsen et al. 2013	194 AD, 226 NC 161 pMCI, 227 sMCI	Cortical Thickness	AD vs. NC ACC = 86.7 %, AUC = 0.917 pMCI vs. sMCI ACC = 73 %, AUC = 0.803
Westman et al. 2013	187 AD, 225 NC 87 pMCI, 200 sMCI	Regional Volume and Cortical Thickness	AD vs. NC ACC = 91.5 %, AUC = 0.96 pMCI vs. sMCI (Conversion time of 24 months) ACC = 69.3 %, AUC = 0.748
Coupé et al. 2012	198 AD, 231 NC 167 pMCI, 238 sMCI	Nonlocal Image Patch of ROIs	AD vs. NC ACC = 89 % pMCI vs. sMCI ACC = 71 %
Cho et al. 2012	128 AD, 160 NC 72 pMCI, 131 sMCI	Cortical Thickness	AD vs. NC ACC = 86 % pMCI vs. sMCI ACC = 71 %
Cuingnet et al. 2011	137 AD, 162 NC 76 pMCI, 134 sMCI	Various (Voxel-based, hippocampus, and Cortical thickness)	AD vs. NC SEN = 81 %, SPE = 95 % pMCI vs. sMCI SEN = 62 %, SPE = 69 %
Duchesne and Mouiha 2011	75 AD, 75 NC 20 pMCI, 29 sMCI	Volume of ROIs	AD vs. NC ACC = 90 %, AUC = 0.9444 pMCI vs. sMCI ACC = 72.3 %, AUC = 0.794

Table 8 (continued)

Reference	Data	Feature extraction	Result
Wolz et al. 2011	198 AD, 231 NC 167 pMCI, 238 sMCI	Various (Hippocampal volume, Cortical thickness, MBL, TBM and CTH)	AD vs. NC ACC = 89 % pMCI vs. sMCI ACC = 68 %

method that can simultaneously utilize approaches of related multi-domain data and feature selection to improve generalization ability of classifiers.

In the MDTL framework, the multi-domain transfer feature selection (MDTFS) is developed, which can use the optimal weight matrix $\mathbf{W}$ to select informative feature subset. Since we use the regularizer of $\|\mathbf{W}\|_{1,1}$ in the MDTFS step, the selected features from each domain are different. For simplicity, in the current work, we select the same feature subset for each auxiliary domain as the target domain (i.e., Strategy1 in Table 7). Actually, we should consider the target domain more than the auxiliary domains in classification task.

To validate the above assumption, we also adopt another strategy for selecting feature subset, i.e., keeping just features with non-zero weights from each column weight vector of $\mathbf{W}$ for all domains (i.e., Strategy2 in Table 7). In Table 7, we list classification results obtained by two different strategies. As we can see from Table 7, the MDTL method using Strategy2 for feature subset selection can achieve slight improvement, compared with the case of using Strategy1. Generally, it is believed that using Strategy2 should be more effective than using Strategy1; but, when using the DeLong's test (DeLong et al. 1988) to assess the statistical difference between AUC values of two strategies, we found no statistical difference between these two strategies.

To further investigate the difference of using these two strategies, we do a statistical analysis on selected features of each domain. Specifically, we count the number of features selected across all folds and all runs (i.e., a total of 100 times for 10-fold cross-validation with 10 independent runs) on the training set. Then, those features with frequency of 100 (i.e., always selected in all folds and all runs) are regarded as stable features. Accordingly, we compute the average percentage of stable features in the target domain and also those stable features in each auxiliary domain, using MDTL method with the Strategy2 for selecting feature subset. For AD vs. NC, MCI vs. NC and pMCI vs. sMCI classification tasks, their corresponding mean ratios are 83 %, 88 % and 92 %. Similarly, we adopt the Strategy1 to select feature subset, and obtained the results that are slightly inferior to the case of the Strategy2. This implies that the target domain plays a critical role in the classification performance, compared with the auxiliary domain.

Comparison with Previous Methods

To further evaluate the efficacy of our proposed multi-domain transfer learning (MDTL) method for early diagnosis of AD, we list a comparison between the MDTL and some representative state-of-the-art methods in the recent 5 years (Cho et al. 2012; Coupé et al. 2012; Cuingnet et al. 2011; Duchesne and Mouiha 2011; Eskildsen et al. 2013; Hu et al. 2016; Khedher et al. 2015; Liu et al. 2014; Moradi et al. 2015; Ota et al. 2015; Westman et al. 2013; Wolz et al. 2011; Zhu et al. 2014), and show them in Table 8. Here, we provide two performance measurements (i.e., ACC: Accuracy; and AUC: Area Under the receiver operating characteristic Curve) in Table 8. Since it is not available the ACC and AUC from the paper of Cuingnet et al. 2011, we just list measurements of sensitivity (SEN) and specificity (SPE). Note that, in Table 8, for several studies using multimodal biomarker (Liu et al. 2014; Zhu et al. 2014), we report their results using only MRI data if available; otherwise, we report their results using multimodal data. Although feature extraction method is different for the comparison methods, this comparison also can show the efficacy of MDTL method at certain level. In most cases, the AUC and ACC of MDTL method are better than those of the comparison methods, indicating that MDTL has better diagnostic performance in early diagnosis of AD.

Table 9 Effects of using different imaging features in our proposed MDTL method. AUC: area under the receiver operating characteristic curve; ACC: accuracy.

Feature	AD vs. NC			MCI vs. NC			pMCI vs. sMCI		
	ACC (%)	AUC	*p*-value	ACC (%)	AUC	*p*-value	ACC (%)	AUC	*p*-value
GM	94.7	0.988	>0.05	81.5	0.882	>0.05	73.8	0.796	>0.05
GM + WM + CSF	95.1	0.988	–	82.1	0.892	–	74.0	0.797	–

Limitations

The current study is limited by several factors. First, our proposed method is based on the single modal (i.e., MRI) biomarker from the ADNI database. In the ADNI database, many subjects also have multimodal biomarkers. Also, many status-unlabeled subjects can be used to extend our current method. In the future work, we will investigate whether adding more auxiliary domain (e.g., multimodal biomarker, and status-unlabeled data) can further improve the performance.

Second, considering the small number of training samples, as well as the sensitivity of those very local features (i.e., thickness, and tissue density) to noises as well as errors in processing pipeline (including skull stripping, tissue segmentation, image registration, and regions-of-interest (ROI) labeling), our current study considers only using ROI features, and no surface-based cortical thickness features are extracted and used although some studies already show the sensitivity of cortical thickness in early diagnosis of AD (Cho et al. 2012; Cuingnet et al. 2011; Eskildsen et al. 2013; Querbes et al. 2009; Wee et al. 2013; Wolz et al. 2011). In the future work, we will consider extracting cortical thickness features from MR images and combine with volume-based features for early diagnosis of AD.

Finally, due to the small number of training samples, we adopted only the volume of gray matter (GM) tissue in each ROI as a feature and input the MDTL model for early diagnosis of AD. However, the study of Cuingnet et al. (Cuingnet et al. 2011) showed that other tissue volumetric feature (i.e., white matter (WM) and CSF) also contributed to AD and NC classification. Accordingly, we also used all types of volumetric features (i.e., GM + WM + CSF) to test the classification performance of our proposed MDTL model, by comparison with the MDTL model using only the GM features. In Table 9, we list their respective ACC and AUC, and further perform the DeLong's test (DeLong et al. 1988) on the AUC to test their statistical difference, with *p*-values provided. The results in Table 9 suggest that using three types of volumetric features can improve the performance, but the corresponding *p*-values show no statistical significant improvement by using three types of volumetric features. In future work, we will improve the MDTL model and combine with the improvement of neuroimaging pre-processing pipeline to enhance the final classification results.

Conclusion

In this paper, we propose a novel multi-domain transfer learning (MDTL) method for early diagnosis of AD, which consists of multi-domain transfer feature selection (MDTFS) and multi-domain transfer classifier (MDTC). The main idea of our multi-domain transfer learning based method is to exploit the multi-auxiliary domain data to improve classification performance (e.g., AD vs. NC, MCI vs. NC and pMCI vs. sMCI) in the target domain. Also, we further combine the source data from multi-auxiliary domain and target domain to guide both feature selection and classification steps. We evaluate our method on the baseline ADNI database with MRI data, and the experimental results demonstrate the efficacy of our method by comparison with several state-of-the-art methods.

Information Sharing Statement

The dataset used in this paper are from the Alzheimer's Disease Neuroimaging Initiative (ADNI, RRID:SCR_003007) which are available at http://adni.loni.usc.edu/. Source code and binary programs developed in this paper are available via our website (http://ibrain.nuaa.edu.cn/, RRID:SCR_014691), and also via email, cb729@nuaa.edu.cn.

Acknowledgments Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc., F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuron Imaging at the University of California, Los Angeles. This work was supported in part by the National Natural Science Foundation of China (Nos. 61602072, 61422204 and 61473149), the Chongqing Cutting-edge and Applied Foundation Research Program (Nos. cstc2016jcyjA0063, cstc2014jcyjA1316, and cstc2014jcyjA40035), the Scientific and Technological Research Program of Chongqing Municipal Education Commission (Nos. KJ1501014, KJ1401010, and KJ1601003), the NUAA Fundamental Research Funds (No. NE2013105), and NIH grants (AG041721, AG049371, AG042599, AG053867).

References

Association, A.s (2014). 2014 Alzheimer's disease facts and figures. *Alzheimers Dement, 10*, 47–92.

Chang, C.C., Lin, C.J., (2001). LIBSVM: a library for support vector machines *http://www.csie.ntu.edu.tw/~cjlin/libsvm/*.

Chao, L. L., Buckley, S. T., Kornak, J., Schuff, N., Madison, C., Yaffe, K., Miller, B. L., Kramer, J. H., & Weiner, M. W. (2010). ASL perfusion

MRI predicts cognitive decline and conversion from MCI to dementia. *Alzheimer Disease and Associated Disorders, 24*, 19–27.

Chen, X., Pan, W., Kwok, J. T., & Carbonell, J. G. (2009). Accelerated gradient method for multi-task sparse learning problem. In *Proceeding of ninth IEEE international conference on data mining and knowledge discovery* (pp. 746–751).

Cheng, B., Zhang, D., & Shen, D. (2012). Domain transfer learning for MCI conversion prediction. In *Proceeding of International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI 2012 7510* (pp. 82–90).

Cheng, B., Zhang, D., Chen, S., Kaufer, D. I., Shen, D., & ADNI (2013). Semi-supervised multimodal relevance vector regression improves cognitive performance estimation from imaging and biological biomarkers. *Neuroinformatics, 11*, 339–353.

Cheng, B., Liu, M., Suk, H., Shen, D., & Zhang, D. (2015a). Multimodal manifold-regularized transfer learning for MCI conversion prediction. *Brain Imaging and Behavior, 9*, 913–926.

Cheng, B., Liu, M., Zhang, D., Munsell, B. C., & Shen, D. (2015b). Domain transfer learning for MCI conversion prediction. *IEEE Transactions on Biomedical Engineering, 62*, 1805–1817.

Chetelat, G., Landeau, B., Eustache, F., Mezenge, F., Viader, F., de la Sayette, V., Desgranges, B., & Baron, J. C. (2005). Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study. *NeuroImage, 27*, 934–946.

Cho, Y., Seong, J. K., Jeong, Y., Shin, S. Y., & ADNI (2012). Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. *NeuroImage, 59*, 2217–2230.

CIT, (2012). Medical Image Processing, Analysis and Visualization (MIPAV) *http://mipav.cit.nih.gov/clickwrap.php*.

Coupé, P., Eskildsen, S. F., Manjón, J. V., Fonov, V. S., Pruessner, J. C., Allard, M., & Collins, D. L. (2012). Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. *NeuroImage: Clinical, 1*, 141–152.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehericy, S., Habert, M. O., Chupin, M., Benali, H., & Colliot, O. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *NeuroImage, 56*, 766–781.

Da, X., Toledo, J. B., Zee, J., Wolk, D. A., Xie, S. X., Ou, Y., Shacklett, A., Parmpi, P., Shaw, L., Trojanowski, J. Q., & Davatzikos, C. (2014). Integration and relative value of biomarkers for prediction of MCI to AD progression: spatial patterns of brain atrophy, cognitive scores, APOE genotype and CSF biomarkers. *NeuroImage: Clinical, 4*, 164–173.

Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., (2011). Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging* 32, 2322.e2319–2322.e2327.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics, 44*, 837–845.

Duan, L. X., Tsang, I. W., & Xu, D. (2012). Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*, 465–479.

Duchesne, S., & Mouiha, A. (2011). Morphological factor estimation via high-dimensional reduction: prediction of MCI conversion to probable AD. *International Journal of Alzheimer's Disease, 2011*, 914085.

Eskildsen, S. F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J. C., & Collins, D. L. (2013). Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *NeuroImage, 65*, 511–521.

Fan, Y., Batmanghelich, N., Clark, C. M., Davatzikos, C., & Initia, A. D. N. (2008). Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage, 39*, 1731–1743.

Filipovych, R., Davatzikos, C., & Initia, A. D. N. (2011). Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI). *NeuroImage, 55*, 1109–1119.

Gaser, C., Franke, K., Kloppel, S., Koutsouleris, N., & Sauer, H. (2013). BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. *PloS One, 8*, e67346.

Guo, X., Wang, Z., Li, K., Li, Z., Qi, Z., Jin, Z., Yao, L., & Chen, K. (2010). Voxel-based assessment of gray and whitematter volumes in Alzheimer's disease. *Neuroscience Letters, 468*, 146–150.

Hinrichs, C., Singh, V., Xu, G. F., Johnson, S. C., & Neuroimaging, A. D. (2011). Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *NeuroImage, 55*, 574–589.

Hu, K., Wang, Y., Chen, K., Hou, L., & Zhang, X. (2016). Multi-scale features extraction from baseline structure MRI for MCI patient classification and AD early diagnosis. *Neurocomputing, 175*, 132–145.

Jie, B., Zhang, D., Cheng, B., & Shen, D. (2015). Manifold regularized multitask feature learning for multimodality disease classification. *Human Brain Mapping, 36*, 489–507.

Kabani, N., MacDonald, D., Holmes, C. J., & Evans, A. (1998). A 3D atlas of the human brain. *NeuroImage, 7*, S717.

Khedher, L., Ramírez, J., Górriz, J. M., Brahim, A., & Segovia, F. (2015). Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images. *Neurocomputing, 151*, 139–150.

Li, H., Liu, Y., Gong, P., Zhang, C., & Ye, J. (2014). Hierarchical interactions model for predicting mild cognitive impairment (MCI) to Alzheimer's disease (AD) conversion. *PloS One, 9*, e82450.

Liu, J., Ji, S., Ye, J., (2009). SLEP: sparse learning with efficient projections. Arizona State University, *http://www.public.asu.edu/~jye02/Software/SLEP*.

Liu, F., Wee, C. Y., Chen, H. F., Shen, D. G., & ADNI (2014). Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. *NeuroImage, 84*, 466–475.

Liu, M., Zhang, D., & Shen, D., (2016a). Inherent structure based multi-view learning with multi-atlas feature representation for alzheimer's disease diagnosis. *IEEE Transactions on Biomedical Engineering, 63*, 1473–1482.

Liu, M., Zhang, D., & Shen, D., (2016b). Relationship induced multi-template learning for diagnosis of alzheimer's disease and mild cognitive impairment. *IEEE Transactions on Medical Imaging, 35*, 1463–1474.

Misra, C., Fan, Y., & Davatzikos, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. *NeuroImage, 44*, 1415–1422.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., & Tohka, J. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage, 104*, 398–412.

Nemirovski, A., (2005). Efficient method s in convex programming.

Obozinski, G., Taskar, B., & Jordan, M. I. (2006). *Multi-task feature selection*. Statistics Department, UC Berkeley: Technical report.

Ota, K., Oishi, N., Ito, K., Fukuyama, H., & Grp, S.-J. S. (2014). A comparison of three brain atlases for MCI prediction. *Journal of Neuroscience Methods, 221*, 139–150.

Ota, K., Oishi, N., Ito, K., & Fukuyama, H. (2015). Effects of imaging modalities, brain atlases and feature selection on prediction of Alzheimer's disease. *Journal of Neuroscience Methods, 256*, 168–183.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*, 1345–1359.

Querbes, O., Aubry, F., Pariente, J., Lotterie, J.-A., Demonet, J.-F., Duret, V., Puel, M., Berry, I., Fort, J.-C., Celsis, P., & ADNI (2009). Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain: A Journal of Neurology, 132*, 2036–2047.

Risacher, S. L., Saykin, A. J., West, J. D., Shen, L., Firpi, H. A., & McDonald, B. C. (2009). Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Current Alzheimer Research, 6*, 347–361.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*.

Sabuncu, M. R., Konukoglu, E., & ADNI (2015). Clinical prediction from structural brain MRI scans: a large-scale empirical study. *Neuroinformatics, 13*, 31–46.

Schwartz, Y., Varoquaux, G., Pallier, C., Pinel, P., Poline, J., & Thirion, B. (2012). Improving accuracy and power with transfer learning using a meta-analytic database. In *Proceeding of International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI 2012 7512* (pp. 248–255).

Shen, D., & Davatzikos, C. (2002). HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging, 21*, 1421–1439.

Sled, J. G., Zijdenbos, A. P., & Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging, 17*, 87–97.

Suk, H., Lee, S. W., Shen, D., & ADNI (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage, 101*, 569–582.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

Wang, Y., Nie, J., Yap, P.-T., Shi, F., Guo, L., Shen, D., 2011. Robust Deformable-Surface-Based Skull-Stripping for Large-Scale Studies. In: Fichtinger, G., Martel, A., Peters, T. (Eds.), Medical Image Computing and Computer-Assisted Intervention. *Springer Berlin / Heidelberg*, Toronto, Canada, pp. 635–642.

Wee, C. Y., Yap, P. T., Shen, D. G., & ADNI (2013). Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. *Human Brain Mapping, 34*, 3411–3425.

Westman, E., Muehlboeck, J. S., & Simmons, A. (2012). Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. *NeuroImage, 62*, 229–238.

Westman, E., Aguilar, C., Muehlboeck, J. S., & Simmons, A. (2013). Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer's disease and mild cognitive impairment. *Brain Topography, 26*, 9–23.

Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D. P., Rueckert, D., Soininen, H., & Lotjonen, J. (2011). Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. *PloS One, 6*, e25446.

Yang, J., Yan, R., Hauptmann, A.G., (2007). Cross-domain video concept detection using adaptive SVMs. *Proceedings of the 15th international conference on Multimedia*, 188–197.

Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., Narayan, V.A., ADNI, (2012). Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. *BMC Neurology* 12, 1471–2377–1412-1446.

Young, J., Modat, M., Cardoso, M. J., Mendelson, A., Cash, D., & Ourselin, S. (2013). Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical, 2*, 735–745.

Zhang, D., Shen, D., (2011). Semi-supervised multimodal classification of Alzheimer's disease. *Proceeding of IEEE International Symposium on Biomedical Imaging* 1628–1631.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. *IEEE Transactions on Medical Imaging, 20*, 45–57.

Zhang, J., Gao, Y., Munsell, B.C., & Shen, D., (2016). Detecting anatomical landmarks for fast alzheimer's disease diagnosis. *IEEE Transactions on Medical Imaging*. Doi: 10.1109/TMI.2016.2582386.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., & ADNI (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage, 55*, 856–867.

Zhang, D., Shen, D., & ADNI (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage, 59*, 895–907.

Zhou, J., Liu, J., Narayan, V. A., Ye, J., & ADNI (2013). Modeling disease progression via multi-task learning. *NeuroImage, 78*, 233–248.

Zhu, X., Huang, Z., Shen, H. T., Cheng, J., & Xu, C. (2012). Dimensionality reduction by mixed kernel canonical correlation analysis. *Pattern Recognition, 45*, 3003–3016.

Zhu, X., Suk, H., & Shen, D. (2014). A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. *NeuroImage, 100*, 91–105.