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

Bo Cheng, Mingxia Liu, and Daoqiang Zhang(✉)

Department of Computer Science and Engineering,
Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
dqzhang@nuaa.edu.cn

Abstract. Recent machine learning based studies for early Alzheimer's disease (AD) diagnosis focus on the joint learning of both regression and classification tasks. However, most of existing methods only use data from a single domain, and thus cannot utilize the intrinsic useful correlation information among data from correlated domains. Accordingly, in this paper, we consider the joint learning of multi-domain regression and classification tasks with multimodal features for AD diagnosis. Specifically, we propose a novel multimodal multi-label transfer learning framework, which consists of two key components: 1) a multi-domain multi-label feature selection (MDML) model that selects the most informative feature subset from multi-domain data, and 2) multimodal regression and classification methods that can predict clinical scores and identify the conversion of mild cognitive impairment (MCI) to AD patients, respectively. Experimental results on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database show that the proposed method help improve the performances of both clinical score prediction and disease status identification, compared with the state-of-the-art methods.

1 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 older Americans will die with AD, and many of them will die from complications caused by AD in 2014 [1]. 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., mild cognitive impairment (MCI). Recently, many machine learning methods based on multimodal biomarkers have been used for early diagnosis of AD [2-5]. These multimodal data include the structural brain atrophy measurements from magnetic resonance imaging (MRI) scans, brain of functional changes by using the fluorodeoxyglucose positron emission tomography (FDG-PET), and pathological amyloid depositions measured through cerebrospinal fluid (CSF). Existing studies have shown that fusing multimodal biomarkers can provide complementary information for learning models, which helps improve the performances compared to methods using single-modality biomarkers [2-5].

In the literature, rather than only identifying disease status in classification problems, several studies begin to predict continuous clinical scores from brain images

L. Zhou et al. (Eds.): MLMI 2015, LNCS 9352, pp. 238–245, 2015.
DOI: 10.1007/978-3-319-24888-2_29

[3,6,13], such as Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog) and Mini-Mental State Examination (MMSE). It is worth nothing that, predicting clinical scores helps evaluate the stage of AD pathology and predict future progression [3,13]. On the other hand, some recent studies have indicated that the tasks of identifying disease status and predicting clinical scores are highly correlated, and the joint learning of regression and classification tasks can help alleviate the small-sample-size problem [3,7,8]. In these methods, the tasks of identifying disease status and predicting clinical scores are considered as different learning tasks, and multi-task learning methods are used to combine those different learning tasks.

However, most of existing studies on the joint learning of regression and classification only focus on using data from a single learning domain, and thus cannot utilize the intrinsic useful correlation information among data from different learning domains. In machine learning community, transfer learning provides an effective solution to deal with the problem involving multiple learning domains of data, and it assumes that the different learning domains have a certain correlation [15]. More recently, several studies have developed transfer learning-based methods for MCI converters (MCI-C) prediction, by treating AD/NC as auxiliary domain to help the learning problem in target domain of MCI-C/MCI non-converters (MCI-NC) [4,9,10]. Although these methods demonstrate that transfer learning methods yield better performance than conventional single-domain based methods, the underlying correlation among different domains is seldom considered in their learning models.

Inspired by the above problems, in this paper, we propose a novel multimodal multi-label transfer learning framework to jointly learn multi-domain regression and classification tasks by using multimodal data. Specifically, we first propose a Multi-Domain Multi-Label feature selection (MDML) method based on transfer learning and multi-label learning, which is used to select the most informative feature-subset from multi-domain data. Then, we employ the multi-kernel support vector machine (M-SVM) for classification and the multi-kernel relevance vector regression machine (M-RVR) for regression, which are used to identify MCI-C patients and to predict clinical scores, respectively. We validate the efficacy of our proposed method on both single-modality and multimodal data (including MRI, FDG-PET and CSF) from the ADNI database.

2 Method

In this section, we introduce our proposed multimodal multi-label transfer learning framework. Specifically, in Section 2.1, we first develop a multi-domain multi-label (MDML) feature selection method for selecting the most discriminative features. Then, in Section 2.2, we employ the multimodal regression and classification methods to predict clinical scores and identify the conversion of MCI to AD patients, respectively.

2.1 Multi-domain Multi-label Feature Selection

In the early AD diagnosis, it is very important to find discriminative brain regions from brain images (e.g., MRI and PET images). In the literature, Lasso-based sparse learning methods are widely used for feature selection to identify the most informa-

tive multimodal biomarkers [11], and have been shown effective in improving the classification performance. On the other hand, some studies suggest that the tasks of identifying disease status and predicting clinical scores are highly correlated. However, traditional Lasso-based methods cannot capture the intrinsic useful correlation information among different label groups (e.g., class labels and clinical score labels). For addressing that problem, we propose a sparse multi-label group Lasso model, by incorporating the underlying correlation information into the learning process.

Assume we have a training set $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N \in \mathbb{R}^{N\times F}$ with N samples, where $\mathbf{x}_n \in \mathbb{R}^F$ is a sample with F features. Denote the label matrix for the training data as $\mathbf{Y} = [\mathbf{y}^1, \ldots, \mathbf{y}^l, \ldots, \mathbf{y}^L] \in \mathbb{R}^{N\times L}$, where $\mathbf{y}^l = \{y_n^l\}_{n=1}^N \in \mathbb{R}^N$ is the l-th type of labels and L is the number of label groups. In this study, there are three different label groups ($L = 3$), including 1) class labels $\mathbf{y}^1$, 2) MMSE score labels $\mathbf{y}^2$, and 3) ADAS-Cog score labels $\mathbf{y}^3$. Let $\mathbf{W} \in \mathbb{R}^{F\times L}$ represent the weight matrix, with the row vector $\mathbf{w}_f$ denoting the coefficient vector associated with f-th feature across different label groups. Then, our proposed sparse multi-label group Lasso model is formulated as follows:

$$\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda_1\|\mathbf{W}\|_{1,1} + \lambda_2\|\mathbf{W}\|_{2,1} \tag{1}$$

where the second term $\|\mathbf{W}\|_{1,1} = \sum_{f=1}^F \sum_{l=1}^L |w_{f,l}|$ can select a discriminative subset of samples relevant self-label group, and the last term $\|\mathbf{W}\|_{2,1} = \sum_{f=1}^F \|\mathbf{w}_f\|_2$ is a 'group sparsity' regularizer that is used to simultaneously select a common feature subset relevant to all label groups. In addition, λ_1 and λ_2 are two regularization parameters that control the relative contributions of those three terms in Eq. (1). Here, the term $\|\cdot\|_F$ is the Frobenius norm of a matrix. By using a specific optimization algorithm [12,14] for solving the optimization problem of Eq. (1), we can get the sparse weight matrix $\mathbf{W}$, where features corresponding to those non-zero coefficients in $\mathbf{W}$ will be selected. In this way, we can find a common feature subset corresponding to all label groups.

Although the sparse multi-label group Lasso model can extract useful correlation information among different label groups, it only addresses the issue of single-domain learning. In the single-domain learning, we separately adopt the sparse multi-label group Lasso model to handle the multiple related domain data and get the multiple weight matrices $\{\mathbf{W}^1, \ldots, \mathbf{W}^d, \ldots, \mathbf{W}^D\}$, where D is the number of related learning domains with an index $d \in \{1, \cdots, D\}$, thus it cannot utilize the intrinsic useful correlation information among multiple related learning domains. To join the multiple related learning domain data, we extend the sparse multi-label group Lasso model to a multi-domain multi-label learning (MDML) model, which is formulated as follows:

$$\begin{aligned}\min_{\mathbf{W}} \tfrac{1}{2}\textstyle\sum_{d=1}^D \|\mathbf{Y}^d - \mathbf{X}^d\mathbf{W}^d\|_F^2 + \lambda_1 \sum_{d=1}^D \|\mathbf{W}^d\|_{1,1} + \lambda_2 \sum_{d=1}^D \|\mathbf{W}^d\|_{2,1} + \\ \lambda_3 \textstyle\sum_{l=1}^L \sum_{d=1}^{D-1} \|\mathbf{w}^{l,d} - \mathbf{w}^{l,d+1}\|_F^2\end{aligned} \tag{2}$$

where $\lambda_1, \lambda_2, \lambda_3 > 0$ are the regularization parameters that control the relative contributions of the four terms, and the last term $\|\mathbf{w}^{l,d} - \mathbf{w}^{l,d+1}\|_F^2$ is adopted to keep the temporal smoothness of multi-weight vector $\mathbf{w}^{l,d}$ among multiple related learning domains [6]. We propose to solve the optimization problem of Eq. (2) by the accele-

rated gradient method (AGM) [12]. In this study, there are two learning domains (i.e., $D = 2$, AD/NC subjects as the related learning domain, and MCI-C/MCI-NC subjects as the target domain).

2.2 Multimodal Regression and Classification

After MDML feature selection on the multi-domain training data, we will employ the multimodal regression and classification methods for combining with multimodal features in the target domain. Similar to the works in [2,3], in our multimodal multi-label transfer learning framework, we use the multi-kernel learning method to combine multimodal features. Specifically, we adopt the multi-kernel support vector machine (M-SVM) [3] to identify the MCI-C patients and employ the multi-kernel relevance vector machine regression (M-RVR) [2] method to predict the clinical scores.

Given M sets of data extracted from different modalities, we first compute the multimodal kernel matrixes $\{\mathbf{K}^{(1)}, \ldots, \mathbf{K}^{(m)}, \ldots, \mathbf{K}^{(M)}\}$. Then, we use the multi-kernel learning method to define a new integrated kernel function for two subjects in the m-th modality (i.e., $\mathbf{x}_a^{(m)}$ and $\mathbf{x}_b^{(m)}$) as follows:

$$\boldsymbol{k}(\mathbf{x}_a, \mathbf{x}_b) = \sum\nolimits_{m=1}^{M} c_m k^{(m)}\left(\mathbf{x}_a^{(m)}, \mathbf{x}_b^{(m)}\right) \tag{3}$$

where $k^{(m)}$ denotes the kernel function for the m-th modality, and c_m denotes the weight for the m-th modality. From Eq. (3), we can achieve the integrated target domain kernel matrix $\mathbf{K} = \sum_{m=1}^{M} c_m \mathbf{K}^{(m)}$. To find the optimal values for weights c_m, we constrain them so that $\sum_m c_m = 1, 0 \le c_m \le 1$ and then adopt a *coarse-grid search* through cross-validation on the training data, which has been shown effective in extensive studies [2,3].

3 Experiments

In this section, we evaluate the effectiveness of our proposed multimodal multi-label transfer learning framework on multimodal data (including MRI, PET and CSF) from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. In the following, we first introduce the experimental settings, and then show the experimental results and discussion.

3.1 Experimental Settings

In our experiments, the baseline ADNI subjects with all corresponding MRI, PET, CSF, MMSE, and ADAS-Cog data are included, which leads to a total of 202 subjects (including 51 AD subjects, 99 MCI subjects, and 52 normal controls (NCs)). For the 99 MCI subjects, it includes 43 MCI converters and 56 MCI non-converters. We use 51 AD and 52 NC subjects as related learning domain, and 99 MCI subjects as target domain. Similar to [3], we adopt an image pre-processing procedure for all MRI and PET images to extract ROI-based features. In addition, three CSF biomarkers are also

used in this study, namely CSF $A\beta_{42}$, CSF t-tau, and CSF p-tau. As a result, for each subject, we have 93 features derived from MRI images, 93 features generated from PET images, and 3 features obtained from CSF biomarkers.

To evaluate the performance of different learning methods, we use a 10-fold cross-validation strategy and repeat this process 10 times to compute the average classification accuracy, sensitivity, specificity, and AUC (Area Under the ROC Curve) value. We also adopt the popular root-mean-square error (RMSE) and the correlation coefficient (CORR) as regression performance measures. In particular, for classifying MCI-C and MCI-NC, we use the 10-fold cross-validation on 99 MCI subjects, and we use the 10-fold cross-validation on all 202 subjects for regression. The SVM classifier is implemented using the LIBSVM toolbox (http://www.csie.ntu.edu.tw/~cjlin/libsvm/), with a linear kernel and a default value for the parameter C ($C = 1$). The regularization parameters (i.e., λ_1, $\lambda_2$and λ_3) can be chosen from the range of $\mathbf{\Omega}$[1] by an inner 10-fold cross-validation on the training data. The RVM regression is implemented using Sparse Bayesian toolbox (http://www.miketipping.com/), with the Gaussian kernel with the width parameter selected from range $\{2^1, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8\}$ that can be determined by an inner 10-fold cross-validation on the training data. In particular, the multi-kernel combination weights are determined via a grid search with range from 0 to 1 and step size 0.1 on the training data. In addition, we also perform the same feature normalization scheme as in [3] in our experiments.

3.2 Results

In the experiments of MCI-C vs. MCI-NC classification, we compare our proposed method with SVM/M-SVM, Lasso, ML-gLasso and Multi-task feature selection (MTFS) [3] methods by using both single-modality data and multimodal data, with results shown in Table 1. It is worth noting that, SVM and M-SVM denote methods using SVM for single-modality and M-SVM for multimodal data without feature selection, respectively. At the same time, Lasso, ML-gLasso, MTFS and MDML denote methods using corresponding feature selection (i.e., Lasso, ML-gLasso, MTFS and MDML) algorithms and adopting SVM/M-SVM methods for classification. In Fig. 1, we also present the ROC curves achieved by different methods for multimodal case. As we can see from Table 1 and Fig. 1, our proposed MDML method consistently achieves better results than SVM/M-SVM, Lasso, MTFS and ML-gLasso methods in terms of all performance measures, which validates the efficacy of our MDML method on using AD and NC subjects as related learning domain. Specifically, in multimodal case, our proposed MDML method can achieve a classification accuracy of 0.787, which is significantly better than M-SVM, Lasso, MTFS and ML-gLasso methods which achieve only 0.638, 0.673, 0.717 and 0.716, respectively. In addition, our proposed ML-gLasso method also achieves better results than SVM/M-SVM, and Lasso methods. It implies that multi-label learning can effectively utilize the intrinsic useful correlation information from multi-label groups.

[1] $\mathbf{\Omega} \in \{0.00001, 0.0001, 0.0005, 0.001, 0.004, 0.007, 0.01, 0.02, 0.03, 0.05, 0.06, 0.08, 0.1, 0.2, 0.4, 0.6, 0.8\}$

Table 1. Comparison of performance of four methods for MCI-C vs. MCI-NC classification using different modalities. (ACC=Accuracy, SEN=Sensitivity, SPE = Specificity, ML-gLasso= Multi-Label group Lasso).

Modality	Method	ACC	SEN	SPE	AUC
MRI+PET+CSF	M-SVM	0.638	0.588	0.677	0.683
	Lasso	0.673	0.739	0.588	0.728
	ML-gLasso	0.716	0.763	0.654	0.761
	MTFS	0.717	0.768	0.649	0.766
	MDML	**0.787**	**0.822**	**0.738**	**0.843**
MRI	SVM	0.539	0.476	0.577	0.554
	Lasso	0.636	0.675	0.583	0.696
	ML-gLasso	0.688	0.722	0.643	0.715
	MTFS	0.667	0.703	0.619	0.718
	MDML	**0.732**	**0.761**	**0.694**	**0.774**
PET	SVM	0.580	0.521	0.625	0.612
	Lasso	0.609	0.651	0.554	0.662
	ML-gLasso	0.642	0.680	0.593	0.635
	MTFS	0.629	0.668	0.578	0.624
	MDML	**0.700**	**0.732**	**0.659**	**0.749**

Table 2. Comparison of performance of four methods for prediction of MMSE/ ADAS-Cog scores using different modalities, respectively.

Modality	Method	MMSE		ADAS-Cog	
		RMSE	CORR	RMSE	CORR
MRI+PET+CSF	M-RVR	2.171	0.526	4.753	0.623
	Lasso	2.039	0.622	4.550	0.686
	mLasso	2.003	0.647	4.415	0.714
	MTFS	1.975	0.655	4.346	0.708
	MDML	**1.963**	**0.663**	**4.258**	**0.725**
MRI	RVR	2.312	0.449	5.422	0.474
	Lasso	2.179	0.514	5.239	0.509
	mLasso	2.128	0.541	5.071	0.552
	MTFS	2.104	0.557	5.004	0.570
	MDML	**2.085**	**0.565**	**4.928**	**0.588**
PET	RVR	2.327	0.387	4.946	0.583
	Lasso	2.215	0.488	4.651	0.646
	mLasso	2.131	0.538	4.494	0.676
	MTFS	2.115	0.551	4.414	0.690
	MDML	**2.094**	**0.561**	**4.314**	**0.704**

On the other hand, our proposed method can be used to predict the clinical scores. Accordingly, in the second group of experiments, we compare our MDML method with RVR/M-RVR, Lasso, mLasso and MTFS [3] methods for both single-modality and multimodal cases. It is worth noting that, for MDML method, without related learning domain can be used for the prediction of clinical scores. In addition, the mLasso feature selection is a variant of ML-gLasso, which has no $L_{1,1}$-norm regularization term and only selects a common feature subset relevant to all label types. Table 2 shows their prediction performance for MMSE/ADAS-Cog scores

using different modalities. Similar to the classification experiments, we first use Lasso, mLasso, MTFS and MDML methods to perform feature selection, and then adopt RVR for single-modality and M-RVR for multimodal for regression, respectively. Note that RVR/M-RVR methods denote using RVR or M-RVR method without feature selection for MMSE/ADAS-Cog scores prediction. From Table 2, one can observe that our proposed MDML method consistently achieves better results than RVR/M-RVR, Lasso, mLasso and MTFS methods, which further validates the efficacy of our MDML method.

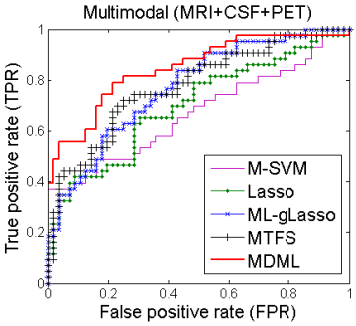

Fig. 1. ROC curves achieved by different methods using multimodal data.

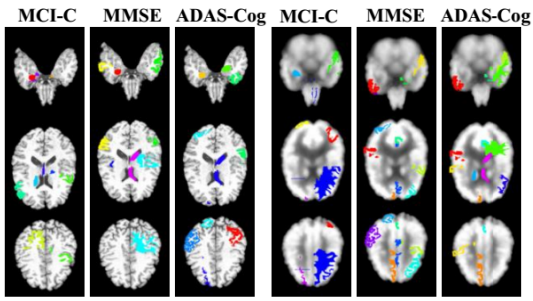

Fig. 2. Stable brain regions identified by MDML method on (a) MRI and (b) PET images. Here, 'MCI-C' is MCI-C vs. MCI-NC classification, and 'MMSE'/'ADAS-Cog' are MMSE/ADAS-Cog scores prediction, respectively.

Finally, in Fig. 2, we visually show the brain regions selected by our MDML method with the highest frequency of occurrence by MDML on MRI and PET images, respectively. Here, to get these features (i.e., brain regions), we count the frequency of each feature and selected across all folds and all runs (i.e., a total of 100 times), and then regard those features as stable features. As can be seen from Fig. 2, our proposed MDML method successfully finds out the most discriminative brain regions (e.g., hippocampal, amygdala, temporal lobe, precuneus, and insula) [3, 11].

4 Conclusion

This paper addresses the problem of jointly exploiting the use of related learning domain data and multi-label group information for early diagnosis of AD. By integrating multi-label learning and transfer learning, we develop a multi-domain multi-label feature selection (MDML) for acquiring the useful correlation information among different learning domains and multi-label groups, and then employ multi-kernel support vector machine for classification and multi-kernel relevance vector machine for regression, respectively. Experimental results on the ADNI database validate the efficacy of our proposed method.

Acknowledgements. This paper is supported by National Natural Science Foundation of China under grant numbers 61422204 and 61473149, by the Jiangsu Natural Science Foundation for Distinguished Young Scholar under grant number BK20130034, by the NUAA Fundamental Research Funds under grant number NE2013105, and by the Scientific and Technological Research Program of Chongqing Municipal Education Commission under Grant KJ1501014.

References

1. Alzheimer's Association: 2014 Alzheimer's disease facts and figures. Alzheimer's & Dement **10**(47), 92 (2014)
2. Cheng, B., Zhang, D., Chen, S., Kaufer, D.I., Shen, D.: Semi-supervised multimodal relevance vector regression improves cognitive performance estimation from imaging and biological biomarkers. Neuroinformatics **11**, 339–353 (2013)
3. Zhang, D., Shen, D.: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage **59**, 895–907 (2012)
4. Young, J., Modat, M., Cardoso, M.J., Mendelson, A., Cash, D., Ourselin, S.: Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. NeuroImage: Clinical **2**, 735–745 (2013)
5. Westman, E., Muehlboeck, J.S., Simmons, A.: Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. NeuroImage **62**, 229–238 (2012)
6. Zhou, J., Liu, J., Narayan, V.A., Ye, J.: Modeling disease progression via multi-task learning. NeuroImage **78**, 233–248 (2013)
7. Zhu, X., Suk, H., Shen, D.: A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. NeuroImage **100**, 91–105 (2014)
8. Wang, H., Nie, F., Huang, H., Risacher, S., Saykin, A.J., Shen, L.: Identifying AD-sensitive and cognition-relevant imaging biomarkers via joint classification and regression. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 115–123. Springer, Heidelberg (2011)
9. Filipovych, R., Davatzikos, C.: Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI). NeuroImage **55**, 1109–1119 (2011)
10. Cheng, B., Zhang, D., Shen, D.: Domain transfer learning for MCI conversion prediction. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part I. LNCS, vol. 7510, pp. 82–90. Springer, Heidelberg (2012)
11. Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., Narayan, V.A.: Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurology **12**, 1471-2377-1412-1446 (2012)
12. Nemirovski, A.: Efficient Method s in Convex Programming (2005)
13. Wang, Y., Fan, Y., Bhatt, P., Davatzikos, C.: High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables. NeuroImage **50**, 1519–1535 (2010)
14. Liu, J., Ji, S., Ye, J.: SLEP: sparse learning with efficient projections. Arizona State University (2009). http://www.public.asu.edu/~jye02/Software/SLEP
15. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**, 1345–1359 (2010)