Manifold Regularized Multi-Task Feature Selection for Multi-Modality Classification in Alzheimer's Disease

Biao Jie[1,2], Daoqiang Zhang[1,*], Bo Cheng[1], and Dinggang Shen[2,*]

[1] Dept. of Computer Science and Engineering,
Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[2] Dept. of Radiology and BRIC, University of North Carolina at Chapel Hill, NC 27599
dqzhang@nuaa.edu.cn, dgshen@med.unc.edu

Abstract. Accurate diagnosis of Alzheimer's disease (AD), as well as its prodromal stage (i.e., mild cognitive impairment, MCI), is very important for possible delay and early treatment of the disease. Recently, multi-modality methods have been used for fusing information from multiple different and complementary imaging and non-imaging modalities. Although there are a number of existing multi-modality methods, few of them have addressed the problem of joint identification of disease-related brain regions from multi-modality data for classification. In this paper, we proposed a manifold regularized multi-task learning framework to jointly select features from multi-modality data. Specifically, we formulate the multi-modality classification as a multi-task learning framework, where each task focuses on the classification based on each modality. In order to capture the intrinsic relatedness among multiple tasks (i.e., modalities), we adopted a group sparsity regularizer, which ensures only a small number of features to be selected jointly. In addition, we introduced a new manifold based Laplacian regularization term to preserve the geometric distribution of original data from each task, which can lead to the selection of more discriminative features. Furthermore, we extend our method to the semi-supervised setting, which is very important since the acquisition of a large set of labeled data (i.e., diagnosis of disease) is usually expensive and time-consuming, while the collection of unlabeled data is relatively much easier. To validate our method, we have performed extensive evaluations on the baseline Magnetic resonance imaging (MRI) and fluorodeoxyglucose positron emission tomography (FDG-PET) data of Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Our experimental results demonstrate the effectiveness of the proposed method.

1 Introduction

Alzheimer's disease (AD) is the most common type of dementia, accounting for 60-80 percent of age-related dementia cases. It is predicted that the number of affected people will double in the next 20 years, and 1 in 85 people will be affected by 2050 [1]. Since the AD-specific brain changes begin years before the patient becomes symptomatic, early clinical diagnosis becomes a challenging task. Therefore, many

[*] Corresponding authors.

K. Mori et al. (Eds.): MICCAI 2013, Part I, LNCS 8149, pp. 275–283, 2013.

studies have focused on possible identification of such changes at the early stage, i.e., mild cognitive impairment (MCI), by leveraging neuroimaging data [2-4].

Recently, machine learning and pattern classifications methods have been widely used in neuroimaging analysis of AD and MCI, including both group comparison (i.e., between clinically different groups) and individual classification. Early researches mainly focus on extracting features (e.g., regions of interest (ROIs) based or voxel-based) from single imaging modality such as structural magnetic resonance imaging (MRI) or fluorodeoxyglucose positron emission tomography (FDG-PET), etc. More recently, researchers have begun to integrate multiple imaging modalities to further improve the accuracy of disease diagnosis.

Different imaging modalities provide different views of brain function or structure. For example, structural MRI provides information about the tissue type of the brain, while FDG-PET measures the cerebral metabolic rate for glucose. Intuitively, integration of multiple modalities may uncover the previously hidden information that cannot be found by using single modality. A number of studies have exploited the fusion of multiple modalities to improve the AD or MCI classification performance [2, 3, 5]. For example, Zhang et al. [2] combined three modalities, i.e., MRI, FDG-PET, and cerebrospinal fluid (CSF), to discriminate AD/MCI and normal controls. Existing studies have indicated that different imaging modalities can provide essential complementary information that can improve accuracy in disease diagnosis.

For imaging modalities, even after feature extraction (i.e., from brain regions), there may still exist the irrelevant features. So, feature selection is commonly used to remove the irrelevant features. However, due to the complexity of brain and the disease, it is challenging to detect all relevant disease-related regions from a single modality alone, especially in early stage of the disease. Different imaging modalities may provide essential complementary information that can help identify these dysfunctional regions implicated by the same underlying pathology. In addition, recent studies also show that there is overlap between the disease-related brain regions detected by MRI and FDG-PET, such as regions in the hippocampus and the mesia temporal lobe [3]. Some feature selection techniques (e.g., t-test) have been used for identifying the disease-related regions from multi-modality data, while an obvious disadvantage of these techniques is that they don't consider the intrinsic relatedness between features across different modalities. To the best of our knowledge, only a few works have exploited to jointly select features from multi-modality neuroimaging data for AD/MCI classification. For example, Huang et al. [3] proposed to jointly identify disease-related brain features from multi-modality data by using sparse composite linear discrimination analysis (SCLDA) method. Zhang et al. [5] proposed a multi-modal multi-task learning for joint feature selection for AD classification, and achieved the state-of-the-art performance in AD classification.

In this paper, as motivated by the work in [5], we proposed a new multi-task-based joint feature selection model that considers both the intrinsic relatedness among multi-modality data and the geometric distribution of each modality data. To this end, we formulate the classification of multi-modality data as a multi-task learning (MTL) problem, where each task focuses on the classification of each modality. The aim of MTL is to improve the generalization performance by jointly learning a set of related tasks [6]. Specifically, two regularization items are included in the proposed model. The first item is group Lasso regularizer [7], which ensure only a small number of features to be jointly selected across different tasks (i.e., modalities). The second item is Laplacian regularization term, which can preserve the geometric distribution information of the whole data from each task. This information may help to capture more discriminative features. Furthermore, we extend our method to the semi-supervised setting (i.e., learning from both labeled and unlabeled data), which is of great importance in practice since the acquisition of labeled data (i.e., diagnosis of disease) is generally expensive and time-consuming, while the collection of unlabeled data is relatively much easier.

2 Manifold Regularized Multi-Task Feature Selection

In this section, we first briefly introduce the existing multi-task feature selection method [5]. Then, we derive our proposed manifold regularized multi-task feature selection models as well as the corresponding optimization algorithm.

2.1 Multi-Task Feature Selection (MTFS)

Assume that there are M supervised learning tasks (i.e., the number of modalities), Denote $X^m = [x_1^m, x_2^m, \dots, x_N^m]^T \in R^{N \times d}$ as the training data matrix on m-th task (i.e., m-th modality) from N training subjects, and $Y = [y_1, y_2, \dots, y_N]^T \in R^N$ as the response vector from these training subjects, where x_i^m represents feature vector of the i-th subject, and y_i is the corresponding class label (i.e., patient or normal control). Let $w^m \in R^d$ parameterizes a linear discriminant function for task m. Then the multi-task feature selection (MTFS) model is to solve the following objective function:

$$\min_W \frac{1}{2} \sum_{m=1}^{M} \|Y - X^m w^m\|_2^2 + \lambda_1 \|W\|_{2,1} \tag{1}$$

where $W = [w^1, w^2, \dots, w^M] \in R^{d \times M}$ is the weight matrix whose row w_j is the vector of coefficients associated with the j-th feature across different tasks. Here, $\|W\|_{2,1} = \sum_{j=1}^{d} \|w_j\|_2$ is the sum of the ℓ_2-norms of the rows of matrix W, as was used in the Group Lasso [7]. The use of $\ell_{2,1}$-norm encourages matrix with many zero rows. In other words, this $\ell_{2,1}$-norm combines multiple tasks and ensures that a small number of common features will be selected across different tasks. The parameter λ_1 is a regularization parameter which balances the relative contributions of the two terms.

2.2 Manifold Regularized Multi-Task Feature Selection (M2TFS)

In the MTFS model, a linear mapping function (i.e., $f(x) = x^T w = w^T x$) was adopted to transform the data from the original high-dimensional space to one-dimensional space. In this model, for each task we only consider the relationship between data and class label, while the mutual dependence among data is ignored, which may result in large deviations even for very similar data after mapping.

To address this problem, we introduced a new regularization term which preserves the geometric distribution information of the whole data:

$$\min_{W} \sum_{m=1}^{M} \sum_{i,j}^{N} \left\| f(x_i^m) - f(x_j^m) \right\|_2^2 S_{ij}^m = 2 \sum_{m=1}^{M} (w^m)^T (X^m)^T L^m X^m w^m \qquad (2)$$

where $S^m = [s_{ij}^m]$ denotes a similarity matrix that defines the similarity on task *m* across different subjects. $L^m = D^m - S^m$ represents combinatorial Laplacian matrix for task *m*, where D^m is the diagonal matrix defined as $D_{ii}^m = \sum_{j=1}^{N} s_{ij}^m$. Here, the similarity matrix can be defined as:

$$s_{ij}^m = \begin{cases} 1\text{ , if } x_i^m \text{ and } x_j^m \text{ are from the same class.} \\ 0\text{ , otherwise.} \end{cases} \qquad (3)$$

This penalized item can be explained as follows. The more similar between x_i^m and x_j^m (i.e., x_i^m and x_j^m come from the same class), the distance between $f(x_i^m)$ and $f(x_j^m)$ shoud be smaller, and *vice versa*. It is easy to see that Eq. (2) aims to preserve the local neighboring structure of *same-class* data during the mapping. With the regularizer in Eq. (2), the proposed manifold regularized multi-task feature selection model (M2TFS) has the following objective function:

$$\min_{W} \frac{1}{2} \sum_{m=1}^{M} \|Y - X^m w^m\|_2^2 + \lambda_1 \|W\|_{2,1} + \lambda_2 \sum_{m=1}^{M} (w^m)^T (X^m)^T L^m X^m w^m \qquad (4)$$

where λ_1 and λ_2 are the two positive constants. Their values can be determined via inner cross-validation on training data.

2.3 Semi-supervised M2TFS (Semi-M2TFS)

Generally, semi-supervised learning methods attempt to exploit the intrinsic data distribution disclosed by the unlabeled data and thus help to construct a better learning model [8]. It is easy to find that, in the proposed M2TFS model, only the first item and the similarity matrix S^m in Eq. (2) involve the supervised information (i.e., the class labels of subjects), so we can easily extend our model to semi-supervised version as follows.

We first define a diagonal matrix $P \in R^{N \times N}$ to indicate labeled data, i.e., $P_{ii} = 0$ if the class label of subject *i* is unknown, and $P_{ii} = 1$ otherwise. Then, according to [9], we redefine the similarity matrix S^m with the following Gaussian function

$$s_{ij}^m = \exp\left(\frac{-\left\| x_i^m - x_j^m \right\|}{2\sigma^2} \right) \qquad (5)$$

Finally, based on the formulation in Eq. (4), the objective function of our semi-supervised M2TFS model (denoted as Semi-M2TFS) can be written as follows:

$$\min_{W} \frac{1}{2} \sum_{m=1}^{M} \|P(Y - X^m w^m)\|_2^2 + \lambda_1 \|W\|_{2,1} + \lambda_2 \sum_{m=1}^{M} (w^m)^T (X^m)^T L^m X^m w^m \qquad (6)$$

where L^m is the corresponding Laplacian matrix based on the new defined similarity matrix S^m in Eq. (5). It is worth noting that Eq. (4) is a special case of Eq. (6) except the definition of similarity matrix. Below, we will develop a new method for optimizing the objective function in Eq. (6).

2.4 Optimization Algorithm

To optimize the problem in Eq. (6), we resort to the widely applied Accelerated Proximal Gradient (APG) method [10]. In this paper, we have implemented an APG optimization procedure similar to that of [11]. Specifically, we first separate the objective function in Eq. (6) to the smooth part:

$$f(W) = \frac{1}{2}\sum_{m=1}^{M}(\|P(Y - X^m w^m)\|_2^2 + 2\lambda_2 (w^m)^T (X^m)^T L^m X^m w^m) \tag{7}$$

and non-smooth part:

$$g(W) = \lambda_1 \|W\|_{2,1} \tag{8}$$

Then, the following function is constructed for approximating the composite function $f(W) + g(W)$:

$$\Omega_l(W, W_k) = f(W_k) + \langle W - W_k, \nabla f(W_k)\rangle + \frac{l}{2}\|W - W_k\| + g(w) \tag{9}$$

where $\nabla f(W_k)$ denotes the gradient of $f(W)$ at point W_k of the k-th iteration, and l is the step size.

Finally, the update step of AGP algorithm is defined as:

$$W_{k+1} = arg\min_{W}\frac{1}{2}\|W - V_k\|_2^2 + \frac{1}{l}g(\mathrm{W}) \tag{10}$$

where l can be determined by line search, and $V_k = W_k - \frac{1}{l}\nabla f(W_k)$

The key of AGP algorithm is how to solve the update step efficiently. The study in [11] shows that this problem can be decomposed into d separate subproblems, and the analytical solutions of these subproblems can be easily obtained.

In addition, according to technique used in [10], instead of performing gradient descent based on W_k, we can compute the following formulation as:

$$Q_k = W_k + \alpha_k (W_k - W_{k-1}) \tag{11}$$

where $\alpha_k = \frac{(1-\beta_{k-1})\beta_k}{\beta_{k-1}}$ and $\beta_k = \frac{2}{k+3}$.

The algorithm for Eq. (6) can achieve a convergence rate of $O(1/K^2)$, where K is the maximum iteration.

3 Classification

Following in [2], we adopted the multi-kernel based support vector machine (SVM) method for classification. Specifically, for each modality of training subjects, a linear kernel was first calculated based on features selected by the above-proposed method. Then, the multi-kernel SVM used in [2] was adopted to combine the multi-modality data for classification.

4 Experiments

To evaluate the effectiveness of our proposed method, we perform a series of experiments on the multi-modality data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). We used a total of 202 subjects with corresponding baseline MRI and PET data, which includes 51 AD patients, 99 MCI patients (including 43 MCI converters and 56 MCI non-converters), and 52 normal controls (NC).

Image pre-processing is performed for all MRI and FDG-PET images. Specifically, we use the specific application tool for image pre-processing as similarly used in [2], i.e., spatial distortion, skull-stripping, and removal of cerebellum. Then, for structural MR images, we use the FSL package [12] to segment each image into three different tissues: gray matter (GM), white matter (WM), and CSF. With atlas warping, each subject was registered to a template with 93 manually labeled regions-of-interest (ROIs) [13]. For each ROI, the volume of GM tissue in that ROI was computed as a feature. For FDG-PET image, we use a rigid transformation to align it onto its respective MR image of the same subject, and then compute the average intensity of each ROI in the FDG-PET image as a feature. Overall, for each subject, we can acquire 93 features from MRI image and another 93 features from PET image.

To evaluate the performance of proposed method, we adopt the classification accuracy, area under receiver operating characteristic (ROC) curve (AUC), sensitivity (i.e., the proportion of patients that are correctly predicted), and specificity (i.e., the proportion of normal controls that are correctly predicted), as performance measures. Two sets of experiments, i.e., supervised classification and semi-supervise classification, were performed on 202 ADNI baseline MRI and PET data, respectively. In both sets of experiments, multiple binary classifiers, i.e., AD vs. NC, MCI vs. NC, and MCI converters (MCI-C) vs. MCI non-converters (MCI-NC), are built, respectively.

4.1 Supervised Classification

In this experiment, 10-fold cross-validation strategy was adopted to evaluate the classification performance. This process is repeated for 10 times independently to avoid any bias introduced by randomly partitioning dataset in the cross-validation. In current studies, we compared our proposed method with the state-of-the-art multi-modality-based methods, including multi-modality method proposed in [2] (denoted as MM and MML, corresponding to 'without feature selection' and 'lasso as feature selection', respectively) and multi-task feature selection method [5] (denoted as MTFS). In addition, for more comparisons, we also concatenate all features from MRI and FDG-PET into a long feature vector, and then perform two different feature selection methods, i.e., t-test, Lasso and sequential floating forward selection (SFFS) [14]. Finally, the standard SVM with linear kernel was used for classification. The detailed experimental results are summarized in Table 1.

As we can see from Table 1, our proposed M2TFS method consistently outperforms the other methods on three classification groups. Specifically, our proposed M2TFS method achieves the classification accuracy of 95.03%, 79.27% and 68.94% for AD vs. NC, MCI vs. NC, and MCI-C vs. MCI-NC, respectively, while the best classification accuracy of other methods are 92.25%, 74.34% and 61.67%, respectively. Also, M2TFS is consistently superior to other methods in sensitivity measure as well as AUC value.

Besides, we performed the significance test between accuracy of our proposed and those of compared methods, using the standard paired t-test. The results show that our proposed method is significantly better than the comparison methods (i.e., all the corresponding p-value are less than 0.01). All these results show that our proposed M2TFS method can take advantage of geometric distribution of data to seek out the most discriminative subset of features.

Table 1. Classification performance of different methods

Methods	AD vs. NC (%)				MCI vs. NC (%)				MCI-C vs. MCI-NC (%)			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
CON-L	91.02	90.39	91.35	0.95	73.44	76.46	67.12	0.78	58.44	52.33	63.04	0.60
CON-T	90.94	91.57	90.00	0.97	73.02	78.08	63.08	0.77	59.11	53.49	63.57	0.64
SFFS	86.78	87.06	86.15	0.93	69.21	82.12	45.38	0.73	56.28	44.42	64.82	0.55
MM	91.65	92.94	90.19	0.96	74.34	85.35	53.46	0.78	59.67	46.28	69.64	0.60
MML	92.25	92.16	92.12	0.96	73.84	77.27	**66.92**	0.77	61.67	54.19	66.96	0.61
MTFS	92.07	91.76	92.12	0.95	74.17	81.31	60.19	0.77	61.61	57.21	65.36	0.62
M2TFS	**95.03**	**94.90**	**95.00**	**0.97**	**79.27**	**85.86**	66.54	**0.82**	**68.94**	**64.65**	**71.79**	**0.70**

4.2 Semi-supervised Classification

In the experiment, we validated the classification performance of our proposed method under semi-supervised setting. Specifically, we first fixed a ratio $r_1 = 50\%$ of positive and negative subjects as labeled data. At the following procedure, we used a fraction $r_2 \in \{10\%, 20\%, 40\%, 60\%, 80\%\}$ of the rest of subjects as unlabeled data. We evaluated our methods with selected labeled data and unlabeled data by using 10-fold cross validation. This process is also repeated 10 times independently. For any chosen fraction r_2 of unlabeled data, we also repeated 10 times to avoid any bias introduced by randomly choosing unlabeled data. The experiment was also repeated 10 times to avoid any bias introduced by randomly choosing labeled data. Fig. 1 shows the classification accuracy of our proposed method with respect to the use of different number of unlabeled samples.

As we can see from Fig. 1, the classification accuracy can be consistently improved with the increase of unlabeled samples on three classification groups, which show that the proposed method can lead to the selection of more discriminative features by using geometric distribution of data, and as a result the classification performance was significantly improved with increase of number of unlabeled data. These results also demonstrate the significant gain obtained by adding the distribution information of data.

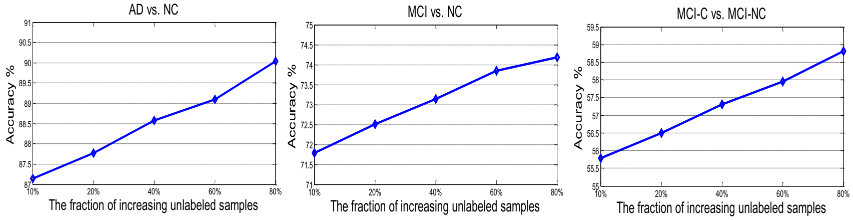

Fig. 1. Classification accuracy with different number of unlabeled samples

5 Conclusion

In summary, this paper addresses the problem of exploiting the geometric distribution of data to build the multi-task feature selection method for jointly selecting features from multi-modalities data. By introducing the manifold regularization item into the multi-task learning framework, we used the accelerated proximal gradient algorithm to seek the optimal solution for seeking out the most informative features subset. We have developed the manifold regularized multi-task feature selection method for both supervised and semi-supervised cases, and the corresponding algorithms are denoted as M2TFS and Semi-M2TFS, respectively. Experimental results on ADNI dataset validate the efficacy of our proposed method. Different from the existing multi-task feature selection method, our method utilizes the geometric distribution knowledge of data for early diagnosis of AD with better results.

Acknowledgments. This work was supported in part by NIH grants EB006733, EB008374, EB009634, and AG041721, SRFDP grant (No. 20123218110009), NUAAFRF grant (No. NE2013105), and also UNSFA grant (No. KJ2013Z095).

References

1. Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H.M.: Forecasting the global burden of Alzheimer's disease. Alzheimers & Dementia 3, 186–191 (2007)
2. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55, 856–867 (2011)
3. Huang, S., Li, J., Ye, J., Chen, K., Wu, T.: Identifying Alzheimer's Disease-Related Brain Regions from Multi-Modality Neuroimaging Data using Sparse Composite Linear Discrimination Analysis. In: Proceedings of Neural Information Processing Systems Conference (2011)
4. 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)
5. 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)
6. Argyriou, A., Evgeniou, T., Pontil, M.: Multi-task Feature Learning. In: NIPS (2006)

7. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. Roy. Stat. Soc. B 68, 49–67 (2006)
8. Zhu, X., Goldberg, A.B.: Introduction to semi-supervised learning. Morgan & Claypool, San Rafael (2009)
9. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. 7, 2399–2434 (2006)
10. Chen, X., Pan, W., Kwok, J.T., Carbonell, J.G.: Accelerated gradient method for multi-task sparse learning problem. In: ICDM (2009)
11. Liu, J., Ye, J.: Efficient L1/Lq Norm Regularization. Technical report, Arizona State University (2009)
12. Zhang, Y., Brady, M., Smith, S.: 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 (2001)
13. Shen, D., Davatzikos, C.: HAMMER: Hierarchical attribute matching mechanism for elastic registration. IEEE Transactions on Medical Imaging 21, 1421–1439 (2002)
14. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recognition Letters 15, 1119–1125 (1994)