Multi-Modal Multi-Task Learning for Joint Prediction of Clinical Scores in Alzheimer's Disease

Daoqiang Zhang[1,2] and Dinggang Shen[1]

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

Abstract. One recent interest in computer-aided diagnosis of neurological diseases is to predict the clinical scores from brain images. Most existing methods usually estimate multiple clinical variables separately, without considering the useful correlation information among them. On the other hand, nearly all methods use only one modality of data (mostly structural MRI) for regression, and thus ignore the complementary information among different modalities. To address these issues, in this paper, we present a general methodology, namely Multi-Modal Multi-Task (M3T) learning, to jointly predict multiple variables from multi-modal data. Our method contains three major subsequent steps: (1) a multi-task feature selection which selects the common subset of relevant features for the related multiple clinical variables from each modality; (2) a kernel-based multimodal data fusion which fuses the above-selected features from all modalities; (3) a support vector regression which predicts multiple clinical variables based on the previously learnt mixed kernel. Experimental results on ADNI dataset with both imaging modalities (MRI and PET) and biological modality (CSF) validate the efficacy of the proposed M3T learning method.

1 Introduction

Alzheimer's disease (AD) is the most common form of dementia in elderly people worldwide. Over the past decades, many AD classification methods have been developed for early diagnosis of AD (including its prodromal stage, i.e., mild cognitive impairment (MCI)) [1-5]. Recently, some AD regression methods are also proposed for the prediction of clinical scores based on brain images [6-9]. Compared with classification, regression needs to estimate continuous rather than categorical variables and are thus more challenging. On the other hand, accurate estimation of clinical scores from brain images is important for helping evaluate the stage of AD pathology and predicting future progression.

In practical diagnosis of AD, generally multiple clinical scores are acquired, i.e., Mini Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale-Cognitive subtest (ADAS-Cog). Specifically, MMSE examines the orientation to time and place, the immediate and delayed recall of three words, the attention and calculations, language, and visuoconstructional functions, while ADAS-Cog is a

T. Liu et al. (Eds.): MBIA 2011, LNCS 7012, pp. 60–67, 2011.

global measure encompassing the core symptoms of AD [8]. It is known that there exist inherent correlations among multiple clinical scores of a subject, since the underlying pathology is the same. However, most existing methods model different clinical scores separately, without considering their inherent correlations that may be useful for robust and accurate estimation of clinical scores from brain images. Recently, a few studies on joint estimation of multiple clinical variables have appeared in imaging literature. For example, in [9], the authors assumed that the related clinical scores share a common relevant feature subset. However, for obtaining a common relevant feature subset, they still needed to perform separate feature selection for each clinical score, and then concatenated the same number of top-ranked features from each clinical score to build a joint regression model.

On the other hand, although multi-modal data are usually acquired for AD diagnosis, i.e., MRI, PET, and CSF biomarkers, nearly all the regression methods developed for estimation of clinical scores were based only on one imaging modality, i.e., the structural MRI. Recent studies have indicated that the biomarkers from different modalities provide complementary information, which is very useful for AD diagnosis [3]. More recently, a series of research works have started to use multi-modal data for AD classification and obtained the improved performance compared with the methods based only on single-modal data [4-5]. However, to the best of our knowledge, the same type of study in imaging-based regression, i.e., estimation of clinical scores from multi-modal data, was not investigated previously.

Inspired by the above problems, in this paper, we present a general methodology, namely Multi-Modal Multi-Task (M3T) learning, to jointly predict multiple clinical scores from multi-modal data. Here, we treat the estimations of multiple clinical scores as different tasks. Specifically, at first, as in the conventional multi-task feature learning methods [10-12], we assume that the related tasks share a common relevant feature subset but with a varying amount of influence on each task, and adopt a multi-task feature selection method [10-11] to obtain a common feature subset for different tasks simultaneously. Then, we use a kernel-based multimodal-data-fusion method to fuse the above-selected features from each individual modality. Finally, we use a support vector regression method [13] to predict multiple clinical scores based on the previously-learnt mixed kernel.

The rest of this paper is organized as follows. In Section 2, we present the proposed Multi-Modal Multi-Task (M3T) learning method in detail. Section 3 reports the experimental results on ADNI dataset using MRI, PET, and CSF biomarkers. Finally, we conclude this paper and indicate issues for future work in Section 4.

2 Method

In this section, we present the new Multi-Modal Multi-Task (M3T) learning method. Our method consists of three subsequent steps, i.e., multi-task feature selection, multiple-modal data fusion, and support vector regression. Specifically, we first assume that the related tasks share a common relevant feature subset but with varying amount of influence on each task, and adopt a multi-task feature selection method to obtain a common feature subset for different tasks simultaneously. Then, we use the kernel-based multimodal-data-fusion method to fuse the above-selected features from each individual modality. Finally, we use a support vector regression to predict multiple clinical scores based on the previously-learnt mixed kernel. Fig. 1 illustrates the flowchart of the proposed M3T method. Note that the feature extraction from MRI and PET images will be discussed in the next section.

2.1 Multi-Task Feature Selection

For imaging modalities such as MRI and PET, even after feature extraction, the number of features (extracted from brain regions) may be still large. Besides, not all features are relevant to the diseases, i.e., clinical scores (tasks). So, feature selection is commonly used for dimensionality reduction, as well as for removal of irrelevant features. Different from the conventional single-task feature selection, the multi-task feature selection simultaneously selects a common feature subset relevant to all tasks [10]. This point is especially important for diagnosis of neurological diseases, since multiple clinical scores are essentially determined by the same underlying pathology, i.e., the diseased brain regions. On the other hand, simultaneously performing feature selection for multiple clinical variables is also very helpful to suppress the noises in the individual clinical scores.

Suppose that we have N training subjects with M modalities and T tasks (clinical scores). Let $x_i^{(1)}, \ldots, x_i^{(M)}$ denote the M modalities of data and $t_i^{(1)}, \ldots, t_i^{(T)}$ the responses of T different tasks for the i-th subject, respectively. Denote $A^{(m)} = [x_1^{(m)}, \ldots, x_N^{(m)}]^{\mathrm{T}}$ as the training data matrix on the m-th modality and $y^{(j)} = [t_1^{(j)}, \ldots, t_N^{(j)}]^{\mathrm{T}}$ as the response vector on the j-th task, respectively. Following [10-11], linear models are used to model the multi-task feature selection (MTFS) as:

$$\hat{t}^{(j)}(x^{(m)}, w_j^{(m)}) = \left(x^{(m)}\right)^T w_j^{(m)}, \quad j = 1, \ldots, T, m = 1, \ldots, M \ , \tag{1}$$

where $w_j^{(m)}$ is the weight vector for the j-th task on m-th modality. The weight vectors for all T tasks form a weight matrix $W^{(m)} = [w_1^{(m)}, \ldots, w_T^{(m)}]$, which can be optimized by the following objective function:

$$\begin{aligned} \min_{W^{(m)}} \ & \frac{1}{2}\sum_{j=1}^{T}\sum_{i=1}^{N}\left(t_i^{(j)} - \hat{t}^{(j)}(x_i^{(m)}, w_j^{(m)})\right)^2 + \lambda\sum_{d=1}^{D_m}\left\|(w^d)^{(m)}\right\|_2 \\ & = \frac{1}{2}\sum_{j=1}^{T}\left\|y^{(j)} - A^{(m)}w_j^{(m)}\right\|_2^2 + \lambda\sum_{d=1}^{D_m}\left\|(w^d)^{(m)}\right\|_2 \end{aligned} , \tag{2}$$

where $(w^d)^{(m)}$ denotes the d-th row of $W^{(m)}$, D_m is the dimension of the m-th modal data, and λ is the regularization coefficient controlling the relative contributions of the two terms. Note that λ also controls the 'sparsity' of the linear models, with the high value corresponding to more sparse models (i.e., more values in $W^{(m)}$ are zero). It is easy to know that Eq. 2 reduces to the standard l_1-norm regularized optimization problem in Lasso [14] when there is only one task. In our case, this is actually a multi-task learning for given m-th modal data.

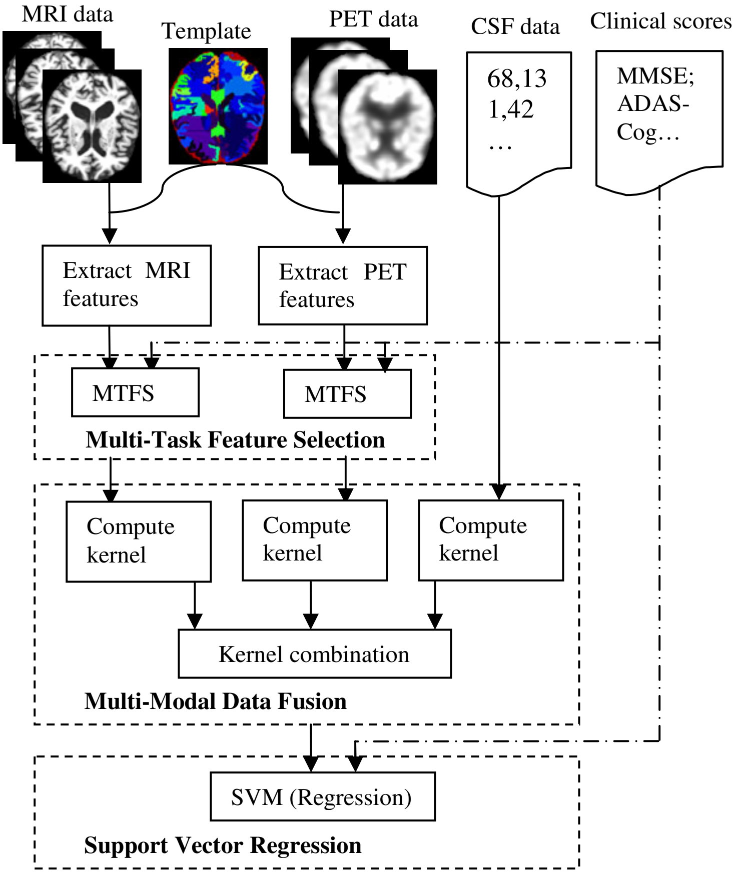

Fig. 1. Flowchart of the proposed M3T method

The key point of Eq. 2 is the use of l_2-norm for $(w^d)^{(m)}$, which forces the weights corresponding to the d-th feature (of the m-th modal data) across multiple tasks to be grouped together and tends to select features based on the strength of T tasks jointly. Note that, because of the characteristic of 'group sparsity', the solution of Eq. 2 results in a weight matrix $W^{(m)}$ whose elements in some rows are all zeros [11]. For feature selection, we just keep those features with non-zero weights. At present, there are many algorithms developed to solve Eq. 2, and in this paper we adopt the SLEP toolbox [15], which has been shown very effective on many datasets.

2.2 Multi-Modal Data Fusion

To effectively fuse data from different modalities, we adopt a multiple-kernels combination scheme in this paper. Assume that we have N training subjects, and each subject has M modalities of data, represented as $x_i=\{x_i^{(1)},\ldots, x_i^{(m)},\ldots, x_i^{(M)}\}$, $i=1,\ldots,N$ (similar to those defined above). Let $k^{(m)}(.,.)$ denote the kernel function on the m-th modality, and then we can define the kernel function $k(.,.)$ on two data x and z as:

$$k(x,z)=\sum_{m=1}^{M}\beta_m k^{(m)}(x^{(m)},z^{(m)}), \tag{3}$$

Where β_ms are the nonnegative weight parameters used to balance the contributions of different modalities. All β_ms are constrained by $\sum_m \beta_m = 1$.

Once we have defined the kernel function $k(.,\ .)$ on multimodal data, a N by N kernel matrix K on multimodal data of all training subjects can be straightforwardly obtained as $K=\{k(x_i,\ x_j)\}$. Then, the subsequent learner such as support vector regression (SVR) can be directly built with the kernel matrix K.

2.3 Support Vector Regression

After obtaining a common feature subset for all different tasks by MTFS and then using those selected features to generate the mixed kernel matrix through multimodal data fusion, we can now train the support vector regression (SVR) to get the final regression models. Here, for simplicity of implementation, we train a separate SVR model for each task. However, it is worth noting that, since we use the common subset of features (selected by MTFS during the feature selection stage) to train the regression models, our models are actually multi-task learning methods, rather than single-task ones. Moreover, one advantage of our models is that they can be easily solved by the standard SVM solvers, e.g., LIBSVM [16].

3 Experiments

In this section, we evaluate the effectiveness of the proposed method for Multi-Modal Multi-Task learning on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). In our experiments, we use three modalities of data including MRI, PET and CSF data. On the other hand, we use two clinical scores, i.e., MMSE and ADAS-Cog, as the two related tasks.

3.1 Subjects and Settings

The ADNI database contains approximately 200 cognitively normal elderly subjects to be followed for 3 years, 400 subjects with MCI to be followed for 3 years, and 200 subjects with early AD to be followed for 2 years. In this paper, all ADNI baseline subjects with the corresponding MRI, PET, and CSF data are included. This yields a total of 202 subjects, including 51 AD patients, 99 MCI patients, and 52 healthy controls. Standard image pre-processing is performed for all MRI and PET images, including anterior commissure (AC) - posterior commissure (PC) correction, skull-stripping, removal of cerebellum, and segmentation of structural MR images into three different tissues: grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF). With atlas warping, we can partition each subject image into 93 regions of interests (ROIs). For each of the 93 ROIs, we compute the GM tissue volume from the subject's MRI image. For PET image, we first rigidly align it with its respective MRI image of the same subject, and then compute the average value of PET signals in each ROI. Therefore, for each subject, we can finally obtain totally 93 features from MRI image, other 93 features from PET image, and 3 features (Aβ_{42}, t-tau, and p-tau) from CSF biomarkers.

In our experiments, we compare our Multi-Modal Multi-Task learning method (denoted as M3T) with conventional Single-Modal Single-Task learning method (denoted as SMST). In SMST, a single-task feature selection (Eq. 2 with single task, i.e., Lasso) is first adopted to select the relevant features for each task separately, and then SVR is performed on those selected features. For comparison, we also implement two other variants, i.e., Single-Modal Multi-Task (SMMT) and Multi-Modal Single-Task (MMST) learning methods. Compared with M3T, single task feature selection (Lasso) is used in MMST, and a single-kernel SVR is used in SMMT. It is worth noting that neither method has been used for regression of clinical scores previously. In addition, we also compare M3T with a Recursive Feature Elimination (RFE) based feature selection also used for joint regression of multiple clinical scores [9], where for fair comparison, we implement the same procedure for the feature selection as in [9] and then use SVR for regression as in our method; we denote this method as RFE_SVR.

Five-fold cross-validation is adopted to evaluate the performances of different algorithms for estimation of multiple clinical scores by measuring the correlation coefficient between the actual clinical score and the predicted one [8]. For all respective methods, the values of the parameters (e.g., λ and β) are determined by performing another cross-validation on the training data. Also, linear kernel is used in SVR after performing a common feature normalization step, i.e., subtracting the mean and then dividing the standard deviation across all subjects for each feature value.

3.2 Results

Table 1 shows the comparison results on regression performances of five different methods. Here, for the multi-task methods (SMMT and M3T), we use the original features of CSF rather than the selected features by feature selection step, since there are only 3 features in this modality.

As can be seen from Table 1, the proposed M3T method is always superior to the conventional SMST method. Specifically, for the estimation of MMSE score, M3T achieves the highest correlation (0.532), while the best corresponding result of SMST is only 0.479 (on MRI modality); on the other hand, for the estimation of ADAS-Cog score, M3T achieves the highest correlation (0.598), while the best corresponding result of SMST is 0.577 (on PET modality). A closer observation on Table 1 shows that the performances of MMST are consistently better than those of other methods except M3T. This validates our assumption that the complementary information among different modalities is helpful for regression, which complements a similar well-known conclusion for classification. Furthermore, by simultaneously using multi-modal and multi-task information, M3T always achieves the best performance among all methods. On the other hand, Table 1 shows RFE_SVR is inferior to M3T which validates the advantage of multi-task feature selection over the strategy of concatenating features from separate single-task feature selection.

Finally, Fig. 2 shows the scatter plots of predicted clinical scores vs. actual scores by two methods for MMSE and ADAS-Cog, respectively. Here, due to space limit, we only list the results of SMST (on MRI modality) and M3T. As can be seen from Fig. 2, using multi-modal and multi-task information in M3T achieves better results than conventional SMST method.

Table 1. Comparison on regression performances of different methods. The reported values are correlation coefficients (mean ± standard deviation).

Methods	Modality	**Task 1 (MMSE)**	**Task 2 (ADAS-cog)**
SMST	MRI	0.479±0.112	0.433±0.164
	CSF	0.351±0.105	0.346±0.109
	PET	0.389±0.101	0.577±0.143
MMST	Multi-modal	0.513±0.078	0.581±0.148
SMMT	MRI	0.483±0.122	0.469±0.160
	CSF	0.351±0.105	0.346±0.109
	PET	0.409±0.116	0.580±0.147
M3T	Multi-modal	0.532±0.097	0.598±0.145
RFE_SVR	MRI	0.507±0.114	0.483±0.180

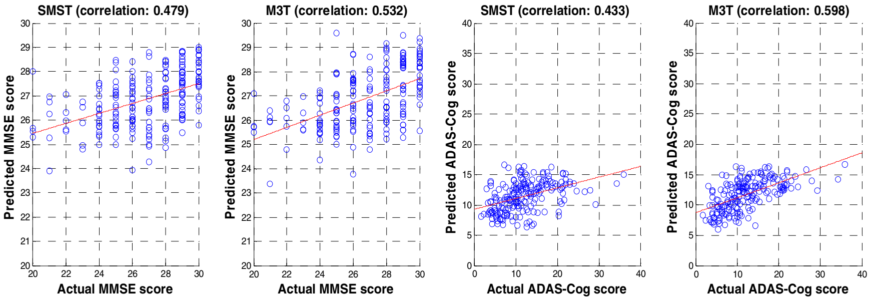

Fig. 2. Scatter plots of predicted **MMSE** and **ADAS-Cog** scores vs. actual scores by SMST (column 1 and 3) and M3T (column 2 and 4). The red solid lines represent the regression lines.

4 Conclusion

We have formulized a new learning problem, called Multi-Modal Multi-Task (M3T) learning, which originates naturally from the practical neurological diseases. Then, we use a new learning framework to jointly predict multiple clinical scores for AD. Specifically, we have developed a new method by combining multi-task feature selection, kernel-based multi-modal data fusion, and support vector regression within an integrated framework. Experimental results on the ADNI dataset have validated the efficacy of the Multi-Modal Multi-Task learning method. In future work, we will develop models which can iteratively use the multi-modal and multi-task information rather than the sequential combination in this paper to further improve performances.

Acknowledgments. This work was supported in part by NIH grants EB006733, EB008374, EB009634 and MH088520, and also by National Science Foundation of China under grant No. 60875030.

References

1. Kloppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack Jr., C.R., Ashburner, J., Frackowiak, R.S.: Automatic classification of MR scans in Alzheimer's disease. Brain 131, 681–689 (2008)
2. Vemuri, P., Wiste, H.J., Weigand, S.D., Shaw, L.M., Trojanowski, J.Q., Weiner, M.W., Knopman, D.S., Petersen, R.C., Jack Jr., C.R.: MRI and CSF biomarkers in normal, MCI, and AD subjects: predicting future clinical change. Neurology 73, 294–301 (2009)
3. Walhovd, K.B., Fjell, A.M., Dale, A.M., McEvoy, L.K., Brewer, J., Karow, D.S., Salmon, D.P., Fennema-Notestine, C.: Multi-modal imaging predicts memory performance in normal aging and cognitive decline. Neurobiol. Aging 31, 1107–1121 (2010)
4. Hinrichs, C., Singh, V., Xu, G., Johnson, S.: MKL for robust multi-modality AD classification. Med. Image Comput. Comput. Assist. Interv. 12, 786–794 (2009)
5. Ye, J., Chen, K., Wu, T., Li, J., Zhao, Z., Patel, R., Bae, M., Janardan, R., Liu, H., Alexander, G., Reiman, E.M.: Heterogeneous data fusion for Alzheimer's disease study. In: ACM International Conference on Knowledge Discovery and Data Mining (2008)
6. Duchesne, S., Caroli, A., Geroldi, C., Frisoni, G.B., Collins, D.L.: Predicting clinical variable from MRI features: Application to MMSE in MCI. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3749, pp. 392–399. Springer, Heidelberg (2005)
7. Wang, Y., Fan, Y., Bhatt, P., Davatzikos, C.: High-dimensional pattern regression using machine learning: from medical images to continuous variables. NeuroImage 50, 1519–1535 (2010)
8. Stonnington, C.M., Chu, C., Kloppel, S., Jack Jr., C.R., Ashburner, J., Frackowiak, R.S.J.: Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. NeuroImage 51, 1405–1413 (2010)
9. Fan, Y., Kaufer, D., Shen, D.: Joint estimation of multiple clinical variables of neurological diseases from imaging patterns. In: The 2010 IEEE International Conference on Biomedical Imaging: from Nano to Macro, pp. 852–855 (2010)
10. Obozinski, G., Taskar, B., Jordan, M.I.: Multi-task feature selection. Technical report, Statistics Department, UC Berkeley (2006)
11. Liu, J., Ji, S., Ye, J.: Multi-task feature learning via efficient l2,1-norm minimization. In: Uncertainty in Artificial Intelligence (2009)
12. Yang, X., Kim, S., Xing, E.P.: Heterogeneous multitask learning with joint sparsity constraints. In: Advances in Neural Information Processing Systems (2009)
13. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
14. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B 58, 267–288 (1996)
15. Liu, J., Ji, S., Ye, J.: SLEP: Sparse Learning with Efficient Projections. Technical report, Arizona State University (2009)
16. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001)