Maximum Mean Discrepancy Based Multiple Kernel Learning for Incomplete Multimodality Neuroimaging Data

Xiaofeng Zhu, Kim-Han Thung, Ehsan Adeli, Yu Zhang, and Dinggang Shen*

Department of Radiology and BRIC, University of North Carolina at Chapel Hill, USA

Abstract. It is challenging to use incomplete multimodality data for Alzheimer's Disease (AD) diagnosis. The current methods to address this challenge, such as low-rank matrix completion (*i.e.,* imputing the missing values and unknown labels simultaneously) and multi-task learning (*i.e.,* defining one regression task for each combination of modalities and then learning them jointly), are unable to model the complex data-to-label relationship in AD diagnosis and also ignore the heterogeneity among the modalities. In light of this, we propose a new Maximum Mean Discrepancy (MMD) based Multiple Kernel Learning (MKL) method for AD diagnosis using incomplete multimodality data. Specifically, we map all the samples from different modalities into a Reproducing Kernel Hilbert Space (RKHS), by devising a new MMD algorithm. The proposed MMD method incorporates data distribution matching, pair-wise sample matching and feature selection in an unified formulation, thus alleviating the modality heterogeneity issue and making all the samples comparable to share a common classifier in the RKHS. The resulting classifier obviously captures the nonlinear data-to-label relationship. We have tested our method using MRI and PET data from Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset for AD diagnosis. The experimental results show that our method outperforms other methods.

1 Introduction

Alzheimer's Disease Neuroimaging Initiative (ADNI) has collected data from various modalities, such as Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), biospecimen, and many others, aiming to use these data to better understand the pathological progression of Alzheimer's Disease (AD) and to develop accurate AD biomarkers. However, due to budget limitation and other constraints, not all the modality data were collected for each subject in the study. For example, at baseline, while all the subjects underwent MRI

* Corresponding author: D. Shen (dgshen@med.unc.edu). This work was supported in part by NIH grants (EB006733, EB008374, EB009634, AG041721, and AG042599). X. Zhu was supported in part by the National Natural Science Foundation of China under grant 61573270.

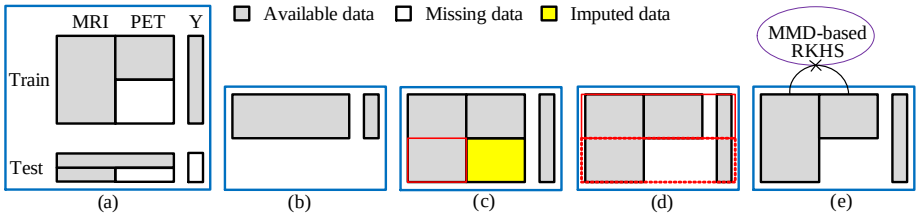

Fig. 1. (a) Block-wise incomplete multimodality data, (b) Disposal method, (c) Imputation method, (d) Multi-task learning method, and (e) Proposed method, which nonlinearly maps heterogeneous MRI and PET data into a common RKHS so that they are comparable, and thus allowing to learn a common MKL-based classifier.

scans, only half of them had PET scans. Assuming that MRI or PET data of one subject can be represented as a row vector, the ADNI neuroimaging multimodality data (*e.g.,* MRI and PET) is block-wise missing, as shown in Fig. 1(a). It is challenging to maximally utilize this kind of incomplete multimodality data (*i.e.,* some modalities are not available for certain subjects) for AD diagnosis. Many AD studies using multimodality data simply dispose the subjects with incomplete data and conduct AD study using only the subjects with complete data [1,5,8,14,12], as shown in Fig. 1(b). This "disposal" method not only significantly reduces the number of the subjects for AD analysis, but also wastes a lot of information in the incomplete subjects, *e.g.,* the red box in Fig. 1(c).

Unlike the disposal method, the imputation method and multi-task learning method are designed to utilize all the samples in incomplete multimodality data for AD study. The imputation method imputes the missing data, as shown in Fig. 1(c), so that any machine learning method can be employed subsequently. Unfortunately, current imputation methods, such as expectation maximization and low-rank matrix completion, are only effective when the data are uniformly missing, and become less effective while the data is block-wise missing, as in our case [9,17]. Without the need of imputation, multi-task learning methods [15,6,16], as shown in Fig. 1(d), first divide the incomplete multimodality data into several subsets of complete data, and then jointly learn a classifier for each subset to conduct AD diagnosis. The main drawback for the imputation and the multi-task learning methods is their underlying assumption of linear data-to-label relationship, which is insufficient to model the complexity of AD progression. Moreover, the data heterogeneity across the modalities (modality heterogeneity for short) is also ignored in their formulations. On the other hand, though the advanced machine learning method such as Multiple Kernel Learning (MKL), is able to model the complex data-to-label relationship of heterogeneous multimodality data [7,2,13], it is currently only applicable to the set of complete data.

In this paper, we propose a new Maximum Mean Discrepancy (MMD) based MKL, so that we can use MKL to conduct AD diagnosis when the data are block-wise missing, as in our case. To do this, we design a new MMD mapping criterion to map the data from different modalities into a common Reproducing

Kernel Hilbert Space (RKHS), as shown in Fig. 1(e). The traditional MMD only considers to minimize the data distribution difference (*i.e.,* a type of high order data relationship) among the modalities [3], while our proposed MMD additionally enforces multiple kernel learning, feature selection and pair-wise sample mismatch minimization. Through the MMD non-linear mapping, the complex data-to-label relationship is captured, the modality heterogeneity is alleviated, and all the data in different modalities become comparable in the RHKS, where a common MKL-based classifier (for these data) is constructed for AD diagnosis.

2 Method

In this paper, we denote $\mathbf{X} = [\mathbf{x}_1^T, \cdots, \mathbf{x}_m^T]^T \in \mathbb{R}^{m \times p}$ and $\mathbf{U} = [\mathbf{u}_1^T, \cdots, \mathbf{u}_n^T]^T \in \mathbb{R}^{n \times q}$ as the Region of Interest (ROI)-based MRI and PET data, respectively, where m and n are the numbers of the samples of the MRI data and the PET data, respectively, p and q indicate the numbers of features in MRI and PET data, respectively, and the superscript T of a matrix indicates its transpose. In addition, $\mathbf{y} \in \{-1, 1\}^m$ and $\mathbf{v} \in \{-1, 1\}^n$ denote the diagnostic labels of the MRI and PET data, respectively.

2.1 Maximum mean discrepancy based MKL

Many studies minimize the heterogeneity among the modalities by using Canonical Correlation Analysis (CCA), which maps all the modalities into a common space [7] via pair-wise distance minimization of all the samples. Since CCA uses pair-wise distances, it is unable to deal with the multimodality data with different numbers of samples, as in our case. Thus, in this study, we design a new MMD criterion to relief the modality heterogeneity between MRI and PET. Traditional MMD criterion [4] uses the data distribution mismatch minimization to make the data from different modalities have similar data distribution in the common RKHS, which does not require equal number of the samples from each modality. The empirical estimation of MMD between $\mathbf{X}$ and $\mathbf{U}$ can be defined as the minimization of the following formulation:

$$\|\frac{1}{m}\sum\nolimits_{i=1}^{m}\phi(\mathbf{x}_i) - \frac{1}{n}\sum\nolimits_{i=1}^{n}\phi(\mathbf{u}_i)\|_{\mathcal{H}}, \tag{1}$$

where $\mathcal{H}$ is a universal RKHS and ϕ is a nonlinear feature mapping of an universal kernel. Recall from kernel methods, the inner product between $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$ is equivalent to a kernel function, *i.e.,* $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$.

In MMD, the empirical estimation distance in the RKHS is regarded as the distance between two different data distributions, as in Eq. (1). Actually, Eq. (1) captures high order statistics of multimodality data (*i.e.,* high order moments of probability distribution) [4], so that the multimodality data are effectively transformed into a high-dimensional or even infinite dimensional space through the nonlinear feature mapping ϕ, where their distributions will be close so that

the heterogeneous data are comparable. When the value of Eq. (1) is close to zero, the high order moments of the multimodality data (*i.e.,* their distributions) become matched. Mathematically, the minimization of Eq. (1) can be reduced to the minimization of the following term:

$$\|\frac{1}{m}\sum\nolimits_{i=1}^{m}\phi(\mathbf{x}_i)-\frac{1}{n}\sum\nolimits_{i=1}^{n}\phi(\mathbf{u}_i)\|_{\mathcal{H}} \Leftrightarrow tr(\mathbf{KS}), \tag{2}$$

where $\mathbf{K}=\begin{bmatrix}\mathbf{K}^{(1,1)} & \mathbf{K}^{(1,2)}\\ \mathbf{K}^{(2,1)} & \mathbf{K}^{(2,2)}\end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}$ is a composite kernel matrix with $\{\mathbf{K}^{(1,1)}=[k(\mathbf{x}_i,\mathbf{x}_j)]\in\mathbb{R}^{m\times m}$, $\mathbf{K}^{(1,2)}=[k(\mathbf{x}_i,\mathbf{u}_g)]\in\mathbb{R}^{m\times n}$, $\mathbf{K}^{(2,1)}=\mathbf{K}^{(1,2)^T}$, $\mathbf{K}^{(2,2)}=[k(\mathbf{u}_g,\mathbf{u}_l)]\in\mathbb{R}^{n\times n}$, $i,j=1,...,m$, and $g,l=1,...,n\}$, and $\mathbf{S}=\mathbf{s}\times\mathbf{s}^T$ (where $\mathbf{s}=[\underbrace{1/m,...,1/m}_{m},\underbrace{-1/n,...,-1/n}_{n}]^T\in\mathbb{R}^{(m+n)}$), and $tr(\cdot)$ is the trace operator of a matrix.

Eq. (2) uses all the ROI-based features of MRI and PET data to build the kernel matrix $\mathbf{K}$. However, not all ROIs are related the AD [9,11], so the resulting $\mathbf{K}$ could be noisy. To address this, we design a feature-level version of Eq. (2) to select a subset of MRI features and PET features for AD diagnosis, via first building a kernel for each feature separately and then combining them through their summation. Specifically, we first extend $\mathbf{X}\in\mathbb{R}^{m\times p}$ and $\mathbf{U}\in\mathbb{R}^{n\times q}$, respectively, to $\tilde{\mathbf{X}}=[\mathbf{X},\mathbf{0}^{m\times q}]\in\mathbb{R}^{m\times(p+q)}$ and $\tilde{\mathbf{U}}=[\mathbf{0}^{n\times p},\mathbf{U}]\in\mathbb{R}^{n\times(p+q)}$, where $\mathbf{0}$ is a matrix with all zero elements. We then map $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{U}}$ into the RKHS by assigning a kernel function to each feature:

$$\min_{\boldsymbol{\alpha}}\ tr(\sum\nolimits_{i=1}^{(p+q)}\alpha_i\tilde{\mathbf{K}}_i\mathbf{S}), \tag{3}$$

where α_i is the weight of each kernel matrix $\tilde{\mathbf{K}}_i\in\mathbb{R}^{(m+n)\times(m+n)}$ (corresponding to each feature) and the kernel matrix $\tilde{\mathbf{K}}_i$ has four components as the kernel matrix $\mathbf{K}$ in Eq. (2). In addition, we also prefer to construct MKL for each feature, rather than fixing a single type of kernel for them, to more flexibly capture nonlinear data-to-label relationships, which leads to

$$\min_{\boldsymbol{\beta}}\ tr(\sum\nolimits_{i=1}^{(p+q)}\sum\nolimits_{j=1}^{M}\beta_{i,j}\hat{\mathbf{K}}_{i,j}\mathbf{S}) \Leftrightarrow \min_{\boldsymbol{\beta}}\ \boldsymbol{\beta}^T\mathbf{a}\mathbf{a}^T\boldsymbol{\beta}, \tag{4}$$

where M is the number of kernel types and $\hat{\mathbf{K}}_{i,j}\in\mathbb{R}^{(m+n)\times(m+n)}$ is the kernel matrix of the i-th feature and j-th kernel type, $\boldsymbol{\beta}=[\beta_{1,1},...,\beta_{1,M},....,\beta_{(p+q),M}]^T\in R^{((p+q)\times M)\times 1}$, $\mathbf{a}=[a_{1,1},...,a_{1,M},....,a_{(p+q),M}]^T\in R^{((p+q)\times M)\times 1}$ with its element given as $a_{i,j}=tr(\hat{\mathbf{K}}_{i,j}\mathbf{S})$. By comparing Eq. (2) and Eq. (3) with Eq. (4), we can see that the original MMD in Eq. (2) [4] has been extended to feature selection based MMD in the MKL framework (*i.e.,* in Eq. (4)), by replacing $\mathbf{K}$ with $\sum_{i=1}^{(p+q)}\alpha_i\tilde{\mathbf{K}}_i$, and then with $\sum_{i=1}^{(p+q)}\sum_{j=1}^{M}\beta_{i,j}\hat{\mathbf{K}}_{i,j}$. In this way, the problem of minimizing distribution mismatch via MMD is converted to the issue of a MKL with the optimal coefficient vector $\boldsymbol{\beta}$ in Eq. (4), which is called MMD based MKL in this paper and can be achieved using MKL algorithm [7]. Hence,

MMD is embedded into the framework of MKL to capture the nonlinear data-to-label relationship among incomplete multimodality data, where the modalities have different numbers of samples.

2.2 Subject consistency

In Section 2.1, we design a new MMD criterion in the MKL framework to map the available MRI and the PET data (*i.e.,* the left box in Fig. 1(e)) to the RKHS. The pair-wise information between the MRI and PET data of the same subject is not considered yet. As MRI and PET data are mapped into a common RKHS so that they are comparable, we would also like to include the subject consistency in our formulation, *i.e.,* samples from the same subject (but different modalities) should be close to each other in the RKHS. To do this, we constrain that the corresponding MRI and PET data to be consistent in the RKHS. Specifically, we consider the pair-wise sample mismatch minimizations (*i.e.,* minimizing the element-wise similarity for each of pair-wise samples) in the RKHS to conduct subject consistency, *i.e.,*

$$\min_{\boldsymbol{\beta}} \ \boldsymbol{\beta}^T \mathbf{d}\mathbf{d}^T \boldsymbol{\beta} \tag{5}$$

where $\mathbf{d} = [d_{1,1}, ..., d_{1,M}, ..., d_{(p+q),M}]^T \in R^{((p+q)\times M)\times 1}$, $d_{i,j} = (\hat{k}_{i,j} - \hat{k}_{i+m,j})$ (where $\hat{k}_{i,j}$ and $\hat{k}_{(i+m),j}$, respectively, are the kernel values of the MRI data and their corresponding PET data, $i = 1, ..., (p+q)$, and $j = 1, ..., M$).

2.3 Joint feature selection and classification

We use MKL-based max-margin classifier (*i.e.,* SVM) to conduct joint feature selection and classification under two constraints that have been described in the previous sections, *i.e.,* distribution mismatch minimization (Section 2.1) and subject consistency (Section 2.2). Thus the final objective function of our proposed method is defined as follows:

$$\begin{aligned} \min_{f,\boldsymbol{\beta}} \ & \tfrac{1}{2}\|f\|_{\mathcal{H}}^2 + C\sum_{i=1}^{(m+n)} L(\hat{y}_i, f(\hat{\mathbf{x}}_i)) + \lambda_1 \boldsymbol{\beta}^T \mathbf{a}\mathbf{a}^T \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^T \mathbf{d}\mathbf{d}^T \boldsymbol{\beta} + \lambda_3 \|\boldsymbol{\beta}\|_1, \\ & s.t., \ \beta_i \geq 0, \ i = 1, ..., (m+n). \end{aligned} \tag{6}$$

where $C > 0$ and $\lambda_j (j = 1, 2, 3)$ are the tuning parameters, $\hat{\mathbf{y}} = [\mathbf{y}; \mathbf{v}] \in R^{(m+n)}$ is a vector of diagnostic labels for MRI and PET samples, $\hat{\mathbf{X}} = [\tilde{\mathbf{X}}; \tilde{\mathbf{U}}] = [\hat{\mathbf{x}}_1^T; ...; \hat{\mathbf{x}}_{(m+n)}^T] \in R^{(m+n)\times(p+q)}$ is the concatenation of extended MRI and PET feature matrix (Section 2.1), f is the prediction function associated with a RKHS $\mathcal{H}$ (*i.e.,* $f \in \mathcal{H}$ and $f(\hat{\mathbf{x}}) = \mathbf{w}^T \phi(\hat{\mathbf{x}}) + \mathbf{b}$), and L is the hinge loss function.

3 Experiments

We used the ADNI dataset ('`www.adni-info.org`') to conduct experiments and compare with various previous works. The used dataset includes 412 MRI

subjects (*i.e.,* 186 ADs and 226 Healthy Controls (HCs)) and 194 PET subjects (*i.e.,* 93 ADs and 101 HCs). More specifically, PET subjects have 218 missing subjects (*i.e.,* 93 ADs and 125 HCs), compared to 412 MRI subjects.

In this paper, we use ROI-based features from both MRI and PET images. The MRI data were sequentially preprocessed by anterior commissure and posterior commissure correction, skull-stripping, cerebellum removal, intensity inhomogeneity correction, segmentation, and registration. Subsequently, we dissected a cerebrum into 90 regions by the AAL template, followed by computing the gray matter tissue volume of each region to yield 90 features for an MRI image. We linearly aligned each PET image to its corresponding MRI image, and then used the mean intensity value of each ROI as PET feature. Finally, we used 90-dimension ROI-based features to represent, MRI and PET data, respectively.

3.1 Experiment setting

We tested our model by conducting two kinds of binary classification experiments, *i.e.,* AD diagnosis using the incomplete MRI and PET data (namely the incomplete data experiment) and AD diagnosis using the PET data with the help of the MRI data (namely the transfer learning experiment). We employed classification accuracy, sensitivity, specificity, and Area Under Curve (AUC) as performance metrics to compare our proposed method with the other methods.

The comparison methods for the two experiments including Baseline (*i.e.,* SVM classification using the MRI data for the incomplete data experiment, and using the PET data for the transfer learning experiment), Lasso [10] (*i.e.,* similar to the Baseline except it performs the Lasso feature selection prior classification), and a multi-task learning method (*i.e.,* regression-based incomplete Multi-Source Feature (iMSF) [11]). In addition, we also included an imputation method (*i.e.,* Low-Rank Matrix Completion with sparse feature selection (LRMC) [9]) for the incomplete data experiment and a popular multiple kernel learning method (*i.e.,* SimpleMKL [7]) for the transfer learning experiment.

3.2 Experimental results

We present the results of all the methods for the two classification experiments in Fig. 2. The results of the incomplete data experiment indicate that the proposed method performs consistently better than all the comparison methods in terms of four evaluation metrics. For example, in terms of accuracy, our method (*i.e.,* 90.9%) on average outperforms the Baseline, Lasso, iMSF, and LRMC methods by 9.1%, 5.9%, 4.7%, and 3.9%, respectively. The superiority of our proposed method is probably due to the nonlinear data-to-label mapping, modality heterogeneity alleviation, and joint feature selection and classification in RKHS of our proposed model. We also observe that all the methods with feature selection (*i.e.,* our proposed method, LRMC, iMSF, and Lasso) outperform the Baseline method, which did not conduct any feature selection. This shows that feature selection is necessary for AD study, which is consistent with the findings in [11,9].

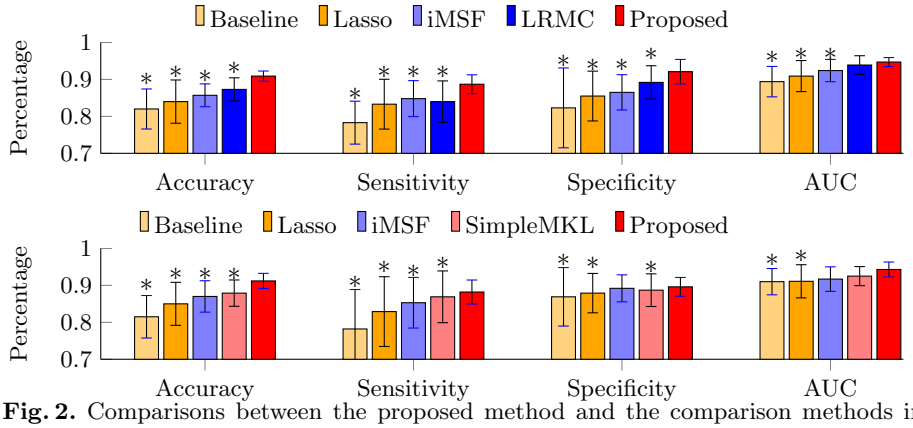

Fig. 2. Comparisons between the proposed method and the comparison methods in two classification experiments, *i.e.,* the incomplete data experiment (Upper row) and the transfer learning experiment (Bottom row). Error bars: standard deviations; *: statistically significant.

In the transfer learning experiment, we use the MRI data to assist AD diagnosis on PET data, so the method LMRC cannot be used for this experiment and we use SimpleMKL, which conducts MKL for AD diagnosis using all the PET and their corresponding MRI data, to be one of the comparison methods in our experiments. According to the experimental results, our proposed method still outperforms all the comparison methods. For example, the proposed method is improved by 8.9% and 3.9%, respectively, in terms of four evaluation metrics, if compared to Baseline and SimpleMKL (which achieves the best performance of all the comparison methods). By comparing the nonlinear feature selection methods (*i.e.,* our proposed method and SimpleMKL) with the linear feature selection methods (*i.e.,* iMSF and Lasso), the nonlinear methods are better than the linear methods in our experiments. This probably due to the fact that there is nonlinear relationship between the data features and the labels.

In addition, we also perform paired t-tests between our results and the results of other methods as significance test. We report the outcomes of the paired t-test in Fig. 2, by marking statistically significant difference results (between our method and all the comparison methods at 95% confidence level) with asterisks (*). The results show that the most of the improvement of the proposed method is statistically significant in our experiments.

4 Conclusion

In this paper, we proposed a MMD-based MKL method for AD diagnosis using *incomplete* multimodality neuroimaging data, which is able to capture the *nonlinear data-to-label relationship*, *relief modality heterogeneity*, and *utilize all the available samples* from different modalities to learn a classifier. To do so, we incorporate feature selection, data distribution and pair-wise sample mismatch minimizations, and classifier learning, in a MKL formulation, to concurrently map all the multimodality data into a common RKHS and learn a common

classifier for all the modalities. The experimental results also confirmed the superiority of our proposed method, compared with other methods.

References

1. Adeli, E., et al.: Joint feature-sample selection and robust diagnosis of parkinson's disease from MRI data. NeuroImage 141, 206–219 (2016)
2. Adeli, E., et al.: Kernel-based joint feature selection and max-margin classification for early diagnosis of parkinsons disease. Scientific reports 7 (2017)
3. Bach, F.R., et al.: Multiple kernel learning, conic duality, and the smo algorithm. In: ICML. p. 6 (2004)
4. Borgwardt, K.M., et al.: Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22(14), e49–e57 (2006)
5. Hor, S., Moradi, M.: Learning in data-limited multimodal scenarios: Scandent decision forests and tree-based features. Medical image analysis 34, 30–41 (2016)
6. Hu, R., et al.: Graph self-representation method for unsupervised feature selection. Neurocomputing 220, 130–137 (2017)
7. Rakotomamonjy, A., et al.: Simplemkl. Journal of Machine Learning Research 9, 2491–2521 (2008)
8. Thung, K., et al.: Neurodegenerative disease diagnosis using incomplete multimodality data via matrix shrinkage and completion. NeuroImage 91, 386–400 (2014)
9. Thung, K.H., et al.: Stability-weighted matrix completion of incomplete multimodal data for disease diagnosis. In: MICCAI. pp. 88–96 (2016)
10. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)
11. Yuan, L., et al.: Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. NeuroImage 61(3), 622–632 (2012)
12. Zhang, S., et al.: Learning k for knn classification. ACM TIST 8(3), 43:1–43:19 (2017)
13. Zhu, X., et al.: Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. IEEE Trans. Biomed. Engineering 63(3), 607–618 (2016)
14. Zhu, X., et al.: A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. Medical Image Analysis 38, 205–214 (2017)
15. Zhu, X., et al.: Robust joint graph sparse coding for unsupervised spectral feature selection. IEEE Trans. Neural Netw. Learning Syst. 28(6), 1263–1275 (2017)
16. Zhu, Y., et al.: Early diagnosis of alzheimers disease by joint feature selection and classification on temporally structured support vector machine. In: MICCAI. pp. 264–272 (2016)
17. Zhu, Y., et al.: Reveal consistent spatial-temporal patterns from dynamic functional connectivity for autism spectrum disorder identification. In: MICCAI. pp. 106–114 (2016)