Multifold Bayesian Kernelization in Alzheimer's Diagnosis

Sidong Liu[1,⋆], Yang Song[1], Weidong Cai[1], Sonia Pujol[2], Ron Kikinis[2], Xiaogang Wang[3], and Dagan Feng[1]

[1] School of Information Technologies, University of Sydney, Australia
[2] Brigham & Women's Hospital, Harvard Medical School, Boston, USA
[3] Department of Electronic Engineering, Chinese University of Hong Kong, HK

Abstract. The accurate diagnosis of Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI) is important in early dementia detection and treatment planning. Most of current studies formulate the AD diagnosis scenario as a classification problem and solve it using various machine learners trained with multi-modal biomarkers. However, the diagnosis accuracy is usually constrained by the performance of the machine learners as well as the methods of integrating the multi-modal data. In this study, we propose a novel diagnosis algorithm, the Multifold Bayesian Kernelization (MBK), which models the diagnosis process as a synthesis analysis of multi-modal biomarkers. MBK constructs a kernel for each biomarker that maximizes the local neighborhood affinity, and further evaluates the contribution of each biomarker based on a Bayesian framework. MBK adopts a novel diagnosis scheme that could infer the subject's diagnosis by synthesizing the output diagnosis probabilities of individual biomarkers. The proposed algorithm, validated using multi-modal neuroimaging data from the ADNI baseline cohort with 85 AD, 169 MCI and 77 cognitive normal subjects, achieves significant improvements on all diagnosis groups compared to the state-of-the-art methods.

1 Introduction

Alzheimer's Disease (AD) is the most common neurodegenerative disorder among aging people and its dementia symptoms gradually deteriorate over years. Mild Cognitive Impairment (MCI) represents the transitional state between AD and cognitive normal (CN) with a high conversion rate to AD. The accurate diagnosis of AD, especially the early diagnosis of MCI converters who develop into AD in a short term, is important in identifying subjects at a high risk of dementia, thereby planning appropriate treatments accordingly.

Neuroimaging, such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), is a fundamental component in the diagnosis and prognosis of AD and MCI. More recently, the large neuroimaging data repositories, e.g., the Alzheimers Disease Neuroimaging Initiatives (ADNI) [1], boost the

⋆ Correspondence to S. Liu (sliu7418@uni.sydney.edu.au). This work was supported by ARC, AADRF, NA-MIC (NIH U54EB005149), and NAC (NIH P41RR013218).

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 303–310, 2013.

research in AD and MCI. Many non-imaging biomarkers, such as cerebrospinal fluid (CSF) measures, genetic biomarkers and clinical assessments, are also provided for the researchers to design algorithms to achieve more accurate diagnosis. Most of the current studies formulate the diagnosis scenario as a classification problem and solve it using various machine learners. These studies are conducted in a similar fashion. The primary features are usually extracted from the MRI data [2–9] and/or PET data [4–8], and sometimes combined with other biomarkers, e.g., CSF measures [4, 6, 8], genetic biomarkers [4, 6, 7] and clinical assessments [6]. The features are then fed into the classifiers, which are trained for future classifications. A challenge of this workflow is how to combine the multi-modal data. Many studies select a subset of features [5, 7, 9], based on the assumption that certain features are not important and therefore could be discarded. However, it is difficult to compare the multi-modal features on the same basis, and the grouping effects of features are usually ignored in feature selection. Several studies attempt to embed the multi-modal features into a unified feature space by linear analysis, e.g., Partial Least Squares (PLS) [4], or non-linear analysis, e.g., ISOMAP [2], yet the existing embedding algorithms could not sufficiently smooth the embeddings of multi-modal features. Another limitation is that the classification accuracy is always constrained by the performance of the classifiers, e.g., support vector machine (SVM) enforces the global consistency and continuity of the boundaries and ignores the local information. The domain knowledge can be used to manipulate the classifiers to further boost the performance [5]. However, the performance gain of such classifier-oriented manipulation might not be transferable when combined with other classifiers. In addition, the domain knowledge might lead to biased classification.

In this study, we propose a novel diagnosis algorithm, the Multifold Bayesian Kernelization (MBK), to model the diagnosis process as a synthesis analysis of multi-modal biomarkers. MBK constructs non-linear kernels to obtain the diagnosis probabilities based on individual biomarkers. It derives the weights of the biomarker-specific kernels with the minimum cost of diagnostic errors and kernelization encoding errors using a Bayesian framework, and infers the subjects diagnosis by synthesizing the output diagnosis probabilities of individual biomarkers. One prominent advantage of MBK is its multi-class nature, unlike other multi-modal methods based on two-class classifications [6–8]. We evaluate the MBK algorithm with 4 diagnosis groups from the ADNI baseline cohort, and the preliminary results show that the MBK algorithm outperforms the state-of-the-art classification-based methods in the diagnosis of AD and MCI.

2 Multifold Bayesian Kernelization

2.1 Algorithm Overview

The goal of the Multifold Bayesian Kernelization (MBK) algorithm is to construct a set of kernels for multi-modal biomarkers and find an optimal way to integrate the diagnosis probabilities of individual biomarkers to enhance the AD and MCI diagnosis. It takes three steps to achieve this goal.

Assume we have a feature set X of N subjects with a collection of B biomarkers, M, the labels of the subjects represented as $Y = \{y_1, \ldots, y_N\}$, the feature for the i^{th} biomarker, $M^{(i)}$, represented as $X^{(i)} = \{x_1^{(i)}, \ldots, x_N^{(i)}\} \in \mathbb{R}^{V^{(i)} \times N}$, where $V^{(i)}$ is the dimension of the features. In the **K**-step, we aim to learn a kernel, $K^{(i)}$, for each biomarker to encode $X^{(i)}$ in such a way to maximize the local neighborhood affinity. Then in the **B**-step, the contribution of each kernel is evaluated based on the Bayesian framework by iteratively minimizing two types of errors: the overall diagnostic errors and the sum of individual kernelization encoding errors. Finally, in the **M**-step, MBK infers the diagnosis of an unknown subject, $\widetilde{x}$, by synthesizing the diagnosis probabilities of individual biomarkers available to $\widetilde{x}$. The proposed diagnosis scheme could take arbitrary biomarkers for analysis. Figure 1 illustrates the workflow of this algorithm.

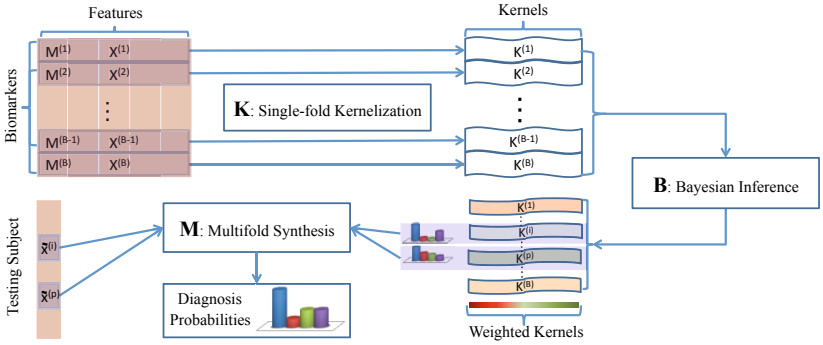

Fig. 1. The workflow of MBK algorithm with three steps shown by the boldfaced letters

2.2 K-Step: Single-Fold Kernelization

Single-fold kernelization aims to preserve the local information and provides a way to infer the subjects label from its affinity to its labeled neighbors. Such local information is essential in AD diagnosis because the features usually have high noise to signal ratio and the data points may not be linearly separable in the feature space.

We construct the kernels for the biomarkers individually by codebook quantization [10]. To begin with, we employ affinity propagation algorithm [11] to select a set of exemplars with least square errors to represent the dataset. The kernel, $K^{(i)}$, is defined as the kernelization codebook of the derived T exemplars, i.e., $K^{(i)} = \{\varepsilon_t\}_{t=1}^{T}$. Each exemplar, ε_t, represents a cluster, C_t, in the feature space, and the marginal distribution of labels given ε_t is defined as:

$$P(y|\varepsilon_t) = \frac{1}{N_t} \sum_{x^{(i)} \in C_t} P(x^{(i)}) \,. \tag{1}$$

where N_t is the number of members in C_t, and $P(x^{(i)})$ is the label distribution for $x^{(i)}$ estimated from itself and its k nearest neighbors. $K^{(i)}$ is used to encode the original features of an unknown subject, $\widetilde{x}^{(i)}$, into a new codeword as:

$$\mathrm{sig}(\widetilde{x}^{(i)}) = \arg\min_{\varepsilon_t} ||\varepsilon_t - \widetilde{x}^{(i)}||^2 \ . \tag{2}$$

The diagnosis probability of $\widetilde{x}^{(i)}$ is derived as the label distribution of its nearest exemplar, i.e., $P(\widetilde{x}^{(i)}) = P(y|\mathrm{sig}(\widetilde{x}^{(i)}))$, and the predicted label of $\widetilde{x}^{(i)}$ is defined as:

$$\widehat{y}^{(i)} = \arg\max_{y}(y|\mathrm{sig}(\widetilde{x}^{(i)})) \ . \tag{3}$$

2.3 B-step: Bayesian Inference

In the **B**-step, we seek to optimally integrate the kernels, K, that could not only achieve more accurate diagnosis, but also preserve the local information of the original features [10], i.e., $K = \arg\max(\mathrm{I}(K, Y) + \mathrm{I}(X, K))$, where $\mathrm{I}(*, *)$ is the mutual information between two items. This optimization problem is equivalent to deriving the weights of each kernel, W, with the minimum cost of the two types of errors, i.e., the overall cost of diagnostic errors and the sum of cost of individual kernelization encoding errors, as in Eq. (4):

$$\arg\min_{W} \overbrace{[\frac{1}{N}\sum_{j=1}^{N}(\frac{1}{2}||\widehat{y}_{j,M,W} - y_j||^2)]}^{\text{Cost of Diagnostic Errors}} + \beta \overbrace{[\sum_{i=1}^{M} W(i) \sum_{j=1}^{N} \mathrm{D}(P(x_j{}^{(i)})|P(y|\mathrm{sig}(x_j{}^{(i)})))]}^{\text{Cost of Kernelization Errors}} \ . \tag{4}$$
$$\text{subject to } \textstyle\sum_{(i=1)}^{M} W(i)=1$$

where $\widehat{y}_{j,M,W}$ is the synthesized diagnosis using all the biomarkers as defined in Eq. (6), $\mathrm{D}(*, *)$ is the Kullback-Leibler divergence, and β is the trade-off parameter between these two types of errors. We initialize W equally, assuming the contributions of all the biomarkers are the same and then iteratively update W as follows: we recalculate the cost derived from each kernel after each iteration and then normalize the costs by the total cost as the inferred posterior weights, W'; we subtract the average weights of all kernels from W' to derive the change rates of the kernels, $\mathrm{d}W$, then use $(W - \mathrm{d}W)$ as the new input to the Bayesian framework; we repeat this process until the cost is minimized and no further improvement can be made.

2.4 M-step: Multifold Synthesis

The **M**-step is used to infer the diagnosis probabilities of a given testing subject with a set of biomarkers, $\widetilde{M}$. The subjects are first encoded into the codewords with the single-fold kernels of $\widetilde{M}$ to derive the diagnosis probabilities based on each biomarker. The diagnosis probabilities using individual kernels are further combined using W to compute the integrated diagnosis probabilities as:

$$P(y|\widetilde{x}, \widetilde{M}, W) = \sum_{i:\{M^{(i)} \in \widetilde{M}\}} W(i) P(y|\mathrm{sig}(\widetilde{x}^{(i)})) \ . \tag{5}$$

where $\mathrm{sig}(\widetilde{x}^{(i)})$ is the codeword of $\widetilde{x}$ derived from the i^{th} single-fold kernelization. Thus the synthesized diagnosis of $\widetilde{x}$ is defined as:

$$\widehat{y}_{j,\widetilde{M},W} = \arg\max_{y} P(y|\widetilde{x}, \widetilde{M}, W) \ . \tag{6}$$

Note that $\widetilde{M}$ is not required to be equal to M. This is because the outputs of the **M**-step are the diagnostic probabilities and the diagnosis can be made based on arbitrary number of biomarkers without a need to re-train the model, although more biomarkers may lead to more deterministic diagnoses. This flexibility makes the MBK algorithm more practical than the metric-based classifiers.

3 Experiments

3.1 Data Acquisition and Feature Extraction

The experiment datasets were obtained from the ADNI database [1]. Totally 331 subjects were selected from the ADNI baseline cohort, including 85 AD-, 169 MCI- and 77 CN- subjects. The MCI group was further divided into two sub-groups. There were 67 MCI subjects converted to AD in half to 3 years from the first scan, and they were considered as the MCI converters (cMCI). The other 102 MCI subjects were then considered as the non-converters (ncMCI). For each subject, an FDG-PET image and a T1-weighted volume acquired on a 1.5 Tesla MR scanner were retrieved. All the 3D MRI and PET data were processed following the ADNI image correction protocols [1, 12]. The PET images were aligned to the corresponding MRI image using FSL FLIRT [13]. We then nonlinearly registered the MRI images to the ICBM_152 template [14] with 83 brain functional regions using the Image Registration Toolkit (IRTK) [15]. The outputted registration coefficients by IRTK were applied to warp the aligned PET images into the template space. We finally mapped all brain functional regions on each registered MRI and PET image using the multi-atlas propagation with enhanced registration (MAPER) approach [16]. Four types of features were extracted from each of the 83 brain regions, including the average cerebral metabolic rate from PET data, and the grey matter volume, solidity, and convexity features from MRI data. Totally 332 sets of features were extracted for each subject. In this study, we used each set of features to represent a biomarker, thus, the feature dimension was 1 for all biomarkers, i.e., $\{V^{(i)} = 1\}_{i=1}^{M}$. Figure 2 shows the process of the data pre-processing and feature extraction.

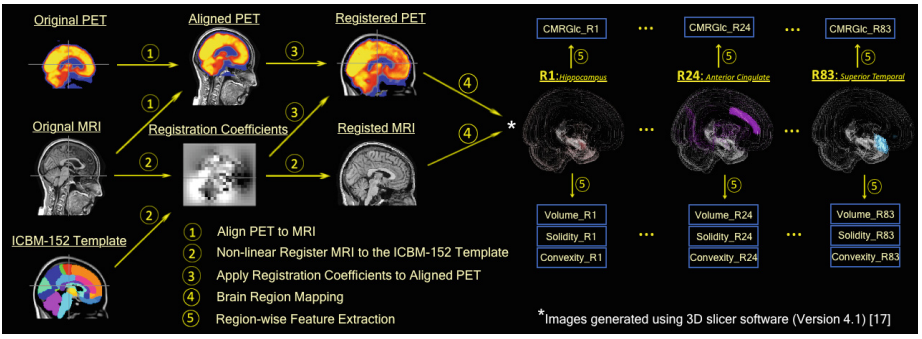

Fig. 2. The procedure for data pre-processing and feature extraction

3.2 Performance Evaluation

We compared the diagnosis performance of the proposed MBK algorithm to three state-of-the-art neuroimaging classification algorithms. We used ISOMAP, same as in [2], as the benchmark of the feature embedding algorithms. Elastic Net was used as the benchmark of the feature selection algorithms, same as in [7]. We further implemented a domain-knowledge-learning graph cuts (DKL-GC) algorithm, a variant of [5], as the benchmark of supervised learning algorithms. More specifically, we designed a cost function to encode the different AD conversion rates and minimize the type II error for cMCI, The features processed by EN and ISOMAP were fed into the SVM with Gaussian kernels. The optimal trade-off parameter (C) and the kernel parameter (γ) for Gaussians in SVM, and the cost function weight parameters in DKL-GC were estimated via grid-search. The parameter settings of MBK were set by pilot experiments ($[k, \beta] = [5, 0.5]$ in this study). A 5-fold cross-validation paradigm was adopted throughout all the algorithms for performance evaluation with a separate subset of the dataset as the testing set and the rest subset as training set each time. SVM was implemented using LIBSVM library [18] and the DKL-GC optimization was solved by the GCO_V3.0 library [19]. Note that for the MBK method, the same training set was used to construct the single-fold kernels in **K**-step as well as to derive the kernel parameters in **B**-step for each fold. The average classification accuracy of 4 diagnosis groups was used to evaluate the performance of different algorithms.

3.3 Results

We divided the biomarkers into two groups according to their modalities, including 83 biomarkers from PET data, and 249 biomarkers from MRI data. We then conducted the Bayesian inference in the **B**-step in MBK using the PET group, MRI group and the merged group (PET + MRI). Figure 3 demonstrates the average diagnosis accuracy and the cost of errors based on the updated weights derived during iteration. The error bars indicate the mean values and standard deviations of the 5 measures by cross-validation. We found that the merged group achieved the highest accuracy with the lowest error cost after 11 iterations and its performance stays stable after 15 iterations.

Table 1 shows the results of the proposed MBK algorithm compared to ISOMAP and EN with SVMs, and DKL-GC. The MBK algorithm outperformed the other classification-based algorithms in all diagnostic groups, achieving an average accuracy of 74.2% compared to 38.4% of the ISOMAP, 54.3% of EN, and 63.29% of DKL-GC. The ISOMAP method had the lowest performance, which indicated that it was not suitable for multi-modal feature embedding. EN introduced l_1 and l_2 penalties on the feature variables to encourage the grouping effect, therefore the correlation between features were better preserved and it achieved better results than ISOMAP. DKL-GC algorithm was specifically designed for prediction of cMCI, as a result the cMCI classification rate of DKL-GC was markedly higher than ISOMAP and EN. However, it required the prior knowledge to assign higher penalty for a cMCI type II error to achieve better cMCI detection; the performance of ncMCI classification was compromised due

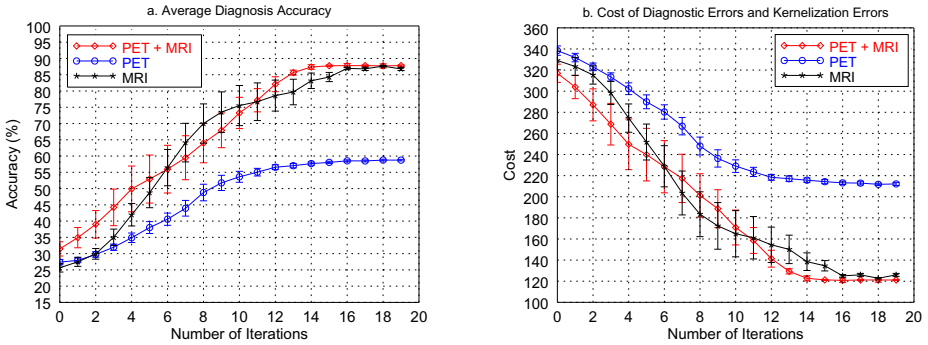

Fig. 3. The cost and accuracy of **B**-step outputs in MBK

Table 1. The diagnosis accuracy (%) of all algorithms, evaluated using PET+MRI biomarkers. **Dgns.** is the ground truth diagnosis, **Prdt.** is the predicted diagnosis.

Algorithm	Dgns. \Prdt.	CN	*nc*MCI	*c*MCI	AD
Feature Embedding: ISOMAP-SVM	CN	**34.33**	38.80	15.60	11.27
	*nc*MCI	26.64	**38.86**	15.12	19.38
	*c*MCI	20.30	34.46	**21.08**	24.16
	AD	16.81	25.66	18.56	**38.96**
Feature Selection: EN-SVM	CN	**60.57**	29.13	4.13	6.17
	*nc*MCI	27.43	**43.56**	11.69	17.32
	*c*MCI	17.96	33.64	**25.06**	23.33
	AD	5.71	19.05	11.43	**63.81**
Supervised Learning: DKL-GC	CN	**64.29**	0.00	0.65	35.06
	*nc*MCI	26.96	**38.24**	2.94	31.86
	*c*MCI	21.64	6.72	**51.49**	20.15
	AD	8.24	7.06	2.94	**81.76**
The Proposed: **MBK**	CN	**86.00**	6.50	1.00	6.50
	*nc*MCI	10.00	**66.96**	0.43	22.61
	*c*MCI	8.48	8.48	**60.61**	22.42
	AD	5.65	8.70	2.17	**83.48**
PET Biomarkers MBK	CN	**59.74**	15.58	9.09	15.58
	*nc*MCI	24.51	**43.14**	3.92	28.43
	*c*MCI	16.42	8.96	**46.27**	28.36
	AD	3.53	16.47	8.24	**71.76**

to such penalty function design. The MBK algorithm requires no domain knowledge and it will not bias the performance of certain diagnosis groups. Table 1 also shows the performance of MBK on 83 PET biomarkers alone using the average weights derived by 5-fold cross-validation for 332 PET+MRI biomarkers. The performance of PET biomarkers alone is not as high as the merged PET+MRI biomarkers, but is comparable with other algorithms. This demonstrates that the MBK works well with varying biomarker set.

4 Conclusions

In this study, we presented a novel diagnosis algorithm, the Multifold Bayesian Kernelization, for the diagnosis of AD and MCI. It differs from the classification-based methods in that: 1) it models the diagnosis process as a synthesis analysis of multi-modal biomarkers; 2) it adopts a novel diagnosis scheme synthesizing the outputted diagnosis probabilities of individual biomarkers instead of combining the inputted features of the biomarkers. The preliminary results showed

that the MBK algorithm outperformed the state-of-the-art classification-based methods and had a great potential in computer aided AD diagnosis.

References

1. Jack, C.R., Bernstein, M.A., et al.: The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. JMRI 127(4), 685–691 (2008)
2. Park, H.: ISOMAP induced manifold embedding and its application to Alzheimer's disease and mild cognitive impairment. Neurosci. Letters 513(2), 141–145 (2012)
3. Risacher, S.L., Saykin, A.J., et al.: Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. Curr. Alz. Res. 6(4), 347–361 (2009)
4. Singh, N., Wang, A.Y., Sankaranarayanan, P., Fletcher, P.T., Joshi, S.: Genetic, structural and functional imaging biomarkers for early detection of conversion from MCI to AD. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part I. LNCS, vol. 7510, pp. 132–140. Springer, Heidelberg (2012)
5. Liu, S., et al.: Neuroimaging biomarker based prediction of Alzheimer's disease severity with optimized graph construction. In: ISBI 2013, pp. 1324–1327. IEEE (2013)
6. Ye, J., Farnum, M., et al.: Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurology 12(1), 46 (2012)
7. Shen, L., et al.: Identifying neuroimaging and proteomic biomarkers for MCI and AD via the elastic net. In: Liu, T., Shen, D., Ibanez, L., Tao, X. (eds.) MBIA 2011. LNCS, vol. 7012, pp. 27–34. Springer, Heidelberg (2011)
8. Zhang, D., Wang, et al.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. NeuroImage 55(3), 856–867 (2011)
9. Liu, S., Cai, W., et al.: Multi-channel brain atrophy pattern analysis in neuroimaging retrieval. In: ISBI 2013, pp. 206–209. IEEE (2013)
10. Lazebnik, S., Raginsky, M.: Supervised learning of quantizer codebooks by information loss minimization. PAMI 31(7), 1294–1309 (2009)
11. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science 315(5814), 972–976 (2007)
12. Jagust, W.J., Bandy, D., et al.: The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core. Alzheimer's & Dementia 6(3), 221–229 (2010)
13. Jenkinson, M., Bannister, P., et al.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17(2), 825–841 (2002)
14. Mazziotta, J., Toga, A., et al.: A probabilistic atlas and reference system for the human brain: international consortium for brain mapping (ICBM). Phil. Trans. Royal Soc. B Biol. Sci. 356(1412), 1293–1322 (2001)
15. Schnabel, J.A., et al.: A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In: Niessen, W.J., Viergever, M.A. (eds.) MICCAI 2001. LNCS, vol. 2208, pp. 573–581. Springer, Heidelberg (2001)
16. Heckemann, R.A., Keihaninejad, S., et al.: Automatic morphometry in Alzheimer's disease and mild cognitive impairment. NeuroImage 56(4), 2024–2037 (2011)
17. Pieper, S., Lorensen, B., et al.: The NA-MIC kit: ITK, VTK, pipelines, grids and 3D Slicer as an open platform for the medical image computing community. In: ISBI 2006, pp. 698–701. IEEE (2006)
18. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM TIST 2(3), 27 (2011)
19. Delong, A., Osokin, A., et al.: Fast approximate energy minimization with label costs. IJCV 96(1), 1–27 (2012)