A Hybrid of Deep Network and Hidden Markov Model for MCI Identification with Resting-State fMRI

Heung-Il Suk[1,⋆], Seong-Whan Lee[1], and Dinggang Shen[1,2]

[1] Department of Brain and Cognitive Engineering, Korea University, Republic of Korea
[2] Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, USA
hisuk@korea.ac.kr

Abstract. In this paper, we propose a novel method for modelling functional dynamics in resting-state fMRI (rs-fMRI) for Mild Cognitive Impairment (MCI) identification. Specifically, we devise a hybrid architecture by combining Deep Auto-Encoder (DAE) and Hidden Markov Model (HMM). The roles of DAE and HMM are, respectively, to discover hierarchical non-linear relations among features, by which we transform the original features into a lower dimension space, and to model dynamic characteristics inherent in rs-fMRI, *i.e.*, internal state changes. By building a generative model with HMMs for each class individually, we estimate the data likelihood of a test subject as MCI or normal healthy control, based on which we identify the clinical label. In our experiments, we achieved the maximal accuracy of 81.08% with the proposed method, outperforming state-of-the-art methods in the literature.

1 Introduction

Motivated by Biswal *et al.*'s study [1] that discovered different brain regions still actively interact while a subject lies at rest, *i.e.*, not performing any cognitive task, resting-state fMRI (rs-fMRI) has been widely used as one of the major tools for investigation of brain networks. It provides insights to explore the brain's functional organization and examine the altered functional networks possibly due to brain disorders such as Mild Cognitive Impairment (MCI). In this regard, functional connectivity analysis has played core roles for brain disease diagnosis or prognosis [4, 7, 11, 12, 15, 16].

While many existing methods for MCI diagnosis with rs-fMRI typically assumed stationarity on the functional networks over time [12, 16], recent studies in neuroscience have shown that the functional organization of a brain is dynamic rather than static, changing spontaneously over time [9]. Eavani *et al.* proposed to jointly model sparse dictionary learning within a state-space model framework [2]. Leonardi *et al.* devised a method to reveal hidden patterns of coherent functional connectivity dynamics based on principal component analysis [11]. In this paper, we propose a novel method that discovers non-linear relations among brain regions in a hierarchical manner and explicitly models the dynamic characteristics inherent in rs-fMRI. It is noteworthy that rather than computing correlation matrices and extracting graph-theoretic features [14] such as small-worldness and clustering coefficients as commonly performed in the literature,

⋆ Corresponding author.

N. Navab et al. (Eds.): MICCAI 2015, Part I, LNCS 9349, pp. 573–580, 2015.
DOI: 10.1007/978-3-319-24553-9_70

we directly model functional dynamics from regional mean time series of rs-fMRI. In a testing phase, our model estimates the data likelihood of a test subject as MCI and Normal healthy Control (NC), based on which we make a clinical decision. Although different groups independently devised different types of state-space models to analyze event-related fMRI data [3, 8, 10], due to their use of variables related to external stimulus, *i.e.*, event, those models cannot be applied to rs-fMRI based disease diagnosis.

2 Materials and Preprocessing

We used a cohort[1] of 37 subjects (12 MCI patients and 25 socio-demographically matched NCs) [15]. The subjects were asked to keep their eyes open and to fixate on a crosshair during scanning. The T1-weighted anatomical MRI images were also acquired from the same scanner.

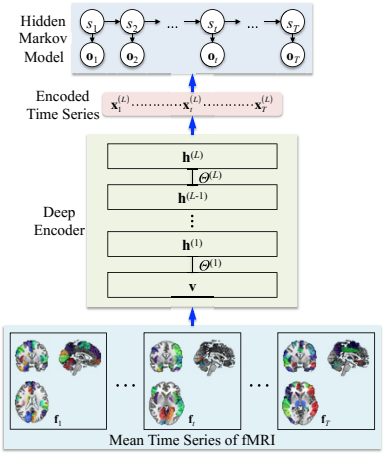
We discarded the first 10 fMRI volume images of each subject for magnetization equilibrium. In order to remove extraneous sources of variation and to isolate the fMRI signals, the remaining 140 fMRI volume images were processed by applying the procedures of slice timing, motion correction, and spatial normalization using SPM8. The images were realigned with TR/2 as a reference time point to minimize the relative errors across TRs. In the motion correction step, we realigned images to the first volume across the subjects. We considered only the signals of gray matter for further processing. The fMRI brain space was then parcellated into 116 Regions-Of-interest (ROIs) based on the Automated Anatomical Labeling (AAL) template.

By following studies in the literature, we utilized the low frequency fluctuation features in rs-fMRI with a frequency band of 0.025∼0.1Hz. The representative mean time series of each ROI was computed by averaging the intensity of all voxels in an ROI. Lastly, we had a set of mean time series $\mathbf{F} \in \left\{ F^{(n)} = \left[\mathbf{f}_1^{(n)}, \ldots, \mathbf{f}_T^{(n)} \right] \in \mathbb{R}^{R \times T} \right\}_{n=1}^{N}$ of the number N(=37) of subjects, the number R(=116) of ROIs, and the number T(=140) of volumes.

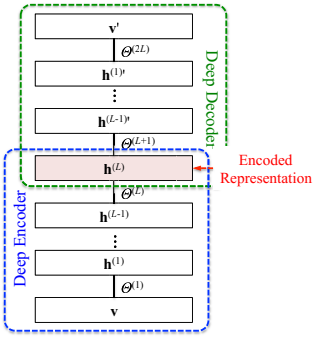
3 Proposed Method

Unlike many existing methods that mostly assumed stationarity of a rs-fMRI time series and explicitly constructed a functional connectivity map, in this paper, we propose a novel probabilistic method that models functional dynamics inherent in rs-fMRI and estimates the data likelihood of a test subject as NC and MCI to make a clinical decision. Specifically, we devise a hybrid architecture by combining Deep Auto-Encoder (DAE) and Hidden Markov Model (HMM) as illustrated in Fig. 1. The roles of DAE and HMM are, respectively, to identify intrinsic networks in a hierarchical manner, from which we extract low-dimensional feature representations, and to model dynamic functional characteristics, *i.e.*, internal functional state changes. It should be noted that HMMs are trained for the classes of NC and MCI separately, while the DAE is shared between classes, and thus it is guaranteed for features of the two classes to lie in the same space.

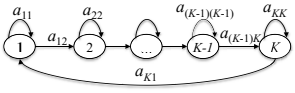
[1] 150 volumes, TR=2,000ms; TE=32ms; flip angle=77°; acquisition matrix size=64×64; field of view=$256 \times 256\text{mm}^2$; 34 axial slices parallel to the anterior commissure-posterior commissure plane; voxel size=$4 \times 4 \times 4\text{mm}^3$.



(a) A hybrid architecture of the proposed method



(b) Deep auto-encoder



(c) Circular topology in HMM

Fig. 1. Illustration of (a) the proposed method for modelling dynamics in rs-fMRI, (b) graphical representation of a deep auto-encoder used to find internal networks and to reduce dimensionality, and (c) the state topology in HMM, where the hidden state variables $[s_1, \ldots, s_t, \ldots, s_T]$ change over time.

3.1 Deep Auto-Encoder

Recently, Hjelm *et al.* [7] demonstrated that Restricted Boltzmann Machines (RBMs) can be used to identify functional networks from fMRI and supported its use as a building block for deeper network models in neuroimaging research. Justified by their work, we design a DAE, structured by stacking multiple RBMs, to discover an embedded representation of functional patterns in a volume of rs-fMRI.

An RBM is a two-layer undirected graphical model with a number D of units in a visible layer and a number F of units in a hidden layer. It assumes symmetric inter-layer connections, but no intra-layer connections. An RBM can be specified with a parameter set Θ of inter-layer connections $\mathbf{W} = [W_{ij}] \in \mathbb{R}^{D \times F}$, a visible layer's bias $\mathbf{z} = [z_i] \in \mathbb{R}^D$, and a hidden layer's bias $\mathbf{q} = [q_j] \in \mathbb{R}^F$, *i.e.*, $\Theta = \{\mathbf{W}, \mathbf{z}, \mathbf{q}\}$, which are learned by minimizing an energy function. In this work, we consider two different energy functions, according to the value types of the visible units $\mathbf{v}$, while using a binary hidden units $\mathbf{h}$. Specifically, when the visible layer has real continuous values, we use a Gaussian-Bernoulli energy function defined as $E(\mathbf{v}, \mathbf{h}; \Theta) = \sum_{i=1}^{D} \frac{(v_i - z_i)^2}{2\sigma_i^2} - \sum_{i=1}^{D} \sum_{j=1}^{F} \frac{v_i}{\sigma_i} W_{ij} h_j - \sum_{j=1}^{F} q_j h_j$, where σ_i denotes a standard deviation of the i-th visible variable that should be learned from data. Meanwhile, for an RBM of binary visible units, it is simplified to a Bernoulli-Bernoulli energy function as $E(\mathbf{v}, \mathbf{h}; \Theta) = -\sum_{i=1}^{D} \sum_{j=1}^{F} v_i W_{ij} h_j - \sum_{i=1}^{D} z_i v_i - \sum_{j=1}^{F} q_j h_j$.

We construct a DAE by using RBMs as building blocks and taking the probability of the lower layer as the inputs to the neighbouring upper layer. The conditional probability of units in the l-th layer given the values of units in the $(l-1)$-th layer is computed as $P\left(\mathbf{h}^{(l)}|\mathbf{h}^{(l-1)}, \Theta^{(l)}\right) = sigm\left(q_j^{(l)} + \sum_i W_{ij}^{(l)} h_i^{(l-1)}/\sigma_i^{(l-1)}\right)$, where $sigm(\cdot)$ denotes a sigmoid function, $\mathbf{h}^{(0)} = \mathbf{v}$, and $\sigma_i^{(l-1)} = 1$ for a binary random vector $\mathbf{h}^{(l-1)}$. Our DAE structurally consists of two parts, namely, 'encoder' and 'decoder' as shown in Fig. 1(b), similar to Hinton and Salakhutdinov's work [6]. Let L denote the number of hidden layers in the encoder, thus the decoder also has L hidden layers. It should be noted that, in our work, the units of bottom input layer, *i.e.*, $\mathbf{v}$ in Fig. 1(b), is modelled with a Gaussian function, while the units of hidden layers remain binary except for those of the middle hidden layer $\mathbf{h}^{(L)}$, for which we use a linear continuous units with Gaussian noises[2].

To learn the parameter sets $\{\Theta^{(1)}, \ldots, \Theta^{(2L)}\}$, we perform the following three steps sequentially with a set of mean time series $\mathbf{F}$ as training samples:

1. Pretrain the parameters $\{\Theta^{(l)}\}_{l=1}^{L}$ of a deep encoder, *i.e.*, network in a blue box in Fig. 1(b), in a greedy layer-wise manner via contrastive divergence algorithm [5]. Note that the mean ROI intensities of the t-th fMRI volume of a subject n, *i.e.*, $\mathbf{f}_t^{(n)}$, becomes the input to $\mathbf{v}$.
2. Unfold the pretrained deep encoder to build a deeper network of encoder and decoder, which we call 'DAE', as shown in Fig. 1(b). For the decoder, *i.e.*, network in the green box in Fig. 1(b), we initially use the same weights of the encoder, pretrained in the first step, *i.e.*, $\Theta^{(L+k)} \leftarrow \Theta^{(L-k+1)}$, $k = 1, \ldots, L$.
3. Fine-tune the parameter sets $\{\Theta^{(1)}, \ldots, \Theta^{(2L)}\}$ of the whole deep neural network, *i.e.*, DAE, jointly by using a back-propagation algorithm [6] with the inputs $\mathbf{v}$ and target outputs $\mathbf{v}'$ kept identical.

Hereafter, we omit the subject index $^{(n)}$ for uncluttered. After completing our DAE training, we use the lower half of our DAE, *i.e.*, deep encoder, to transform the rs-fMRI feature vectors $F = [\mathbf{f}_1, \cdots, \mathbf{f}_t, \cdots, \mathbf{f}_T]$ into encoded representations $X = [\mathbf{x}_1, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_T]$, which are further fed into HMMs to identify clinical status between NC and MCI. It is remarkable that by setting the number of hidden units in the top hidden layer $\mathbf{h}^{(L)}$ of a deep encoder smaller than the dimension of the input, *i.e.*, R, it naturally has the effect of reducing dimensionality of the input vector $\mathbf{f}_t$ but still has the rich information necessary to reproduce the input in a non-linear way.

However, note that a DAE is utilized to find the highly non-linear relations among different regions at one time without considering the temporal information, which is important to discriminate MCI from NC. We handle such temporal or dynamic information with an HMM described below.

3.2 Hidden Markov Models

Based on recent studies in [4, 11], it is reasonable to assume that the groups of NC and MCI exhibit different functional characteristics, depending on the unobservable

[2] The rationale of using linear units with Gaussian noises is to obtain continuous values for better representational power of the coded representations [6].

functional states that spontaneously change over time. In this paper, we model such dynamics inherent in rs-fMRI by the first-order Markov chain in HMMs [13] for NC and MCI, separately.

An HMM is a doubly stochastic process of 1) hidden process $\{s_1, \cdots, s_t, \cdots, s_T\}$ that is latent but can be estimated by 2) observable process $\{\mathbf{o}_1, \cdots, \mathbf{o}_t, \cdots, \mathbf{o}_T\}$, which produces a sequence of observations, where s_t and $\mathbf{o}_t$ denote random variables of hidden state and observation at time t, respectively. A hidden process is represented by two probability distributions, namely, state transition probability $A = [a_{ij}]_{i,j=\{1,\ldots,K\}}$ and initial state probability $\mathit{\Pi} = [\pi_i]_{i=\{1,\ldots,K\}}$, where K denotes the number of hidden states, $a_{ij} = P(s_t = j | s_{t-1} = i)$, and $\pi_i = P(s_1 = i)$. Meanwhile, the observable process is depicted by emission probability density function (*pdf*) $B = \{b_i\}_{i=\{1,\ldots,K\}}$, where $b_i = p(\mathbf{o}_t = \mathbf{x}_t | s_t = i)$. In this work, we use a mixture of Gaussians for an emission *pdf* b_i. Thus, an HMM is completely defined by the parameter set of $\lambda = (A, B, \mathit{\Pi})$. For simplicity, we denote, hereafter, HMMs for NC and MCI with λ_{NC} and λ_{MCI}, respectively.

Note that a functional pattern of rs-fMRI at a time-point belongs to one of a finite number K of states, which is represented by an observation probability B. Meanwhile, the changes of the unobservable states in rs-fMRI are denoted by the state transition probability A along with the initial state probability $\mathit{\Pi}$. By training HMMs with a Baum-Welch algorithm [13] for NC and MCI individually, they can be used as a way to represent the functional dynamic characteristics of the respective groups in a probabilistic manner. In other words, given a sequence of functional features, *i.e.*, encoded representations $X = [\mathbf{x}_1, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_T]$ in our work, we infer that how likely the sequence of functional features X is generated from HMMs of NC (λ_{NC}) and MCI (λ_{MCI}), respectively, as follows:

$$p(X|\lambda_c) = \sum_S p(X|S, \lambda_c) P(S|\lambda_c) \tag{1}$$

where $c \in \{\text{NC}, \text{MCI}\}$, $S = [s_1, \cdots, s_t, \cdots, s_T]$, and $s_t \in \{1, \ldots, K\}$. Eq. (1) can be efficiently computed by the forward algorithm [13]. We identify the clinical label of the rs-fMRI of a test subject to the class of the higher data likelihood.

4 Experiments and Discussion

4.1 Experimental Settings

With regard to the structure of our DAE, we considered four hidden layers, *i.e.*, $L = 4$, to encode the input functional features by setting the number of hidden units as 200($\mathbf{h}^{(1)}$)-100($\mathbf{h}^{(2)}$)-50($\mathbf{h}^{(3)}$)-2($\mathbf{h}^{(4)}$). Thus, the decoder was structured as 50($\mathbf{h}^{(5)}$)-100($\mathbf{h}^{(6)}$)-200($\mathbf{h}^{(7)}$)[3]. A Gaussian-Bernoulli energy function was used for the input visible units, *i.e.*, $\mathbf{v}$ in Fig. 1, while a Bernoulli-Bernoulli energy function was exploited for the hidden layers by taking the outputs of the lower layer as inputs. But the units of the top hidden layer in the encoder, *i.e.*, $\mathbf{h}^{(4)}$ in Fig. 1, had stochastic real-values, allowing the low-dimensional codes to distribute in a continuous feature space [6].

[3] Therefore, the complete structure of our DAE was 116-200-100-50-2-50-100-200-116.

For the HMMs of NC and MCI classes, we varied the number of hidden states K from 2 to 6 with different number of Gaussians for emission *pdf*s, varying between 1 and 4. Due to a small data set, a circular state topology in Fig. 1(c) was used for both classes. In order to learn the parameters (A, B, Π), we used a BNT toolbox[4].

To validate the effectiveness of the proposed method, we compared with four competing methods in the literature, namely, group Independent Component Analysis (gICA) [12], group Sparse Representation (gSR) [16], Principal Component of Functional Connectivity (PCFC) [11], and a joint framework of HMM and Sparse Dictionary Learning (HMM+SDL) [2]. We also compared with a method of combining kernel Principal Component Analysis (kPCA) with HMM to validate the effectiveness of DAE-based dimension reduction.

- gICA: We applied a fastICA algorithm using a GIFT toolbox[5]. The number of independent components was set to 30 by following Li *et al.*'s work [12]. After performing group ICA, for each subject, we computed the correlation coefficients of every pair of time courses and used them as features.
- gSR: For the regularization control parameter, we applied a grid search technique in the space of $\{0.01, 0.05, 0.1, 0.15, 0.2, 0.5\}$. We used clustering coefficients obtained from a functional connectivity map as features.
- PCFC[6]: We used a sliding window-based Functional Network (FN) modeling with a window size of 30 time points and a stride of 5 time points between consecutive windows. The estimated FNs were then projected into eigen-networks, the number of which was determined based on eigenvalues such that the transformed features hold more than 85% of the total variance. The features from each FN were then concatenated into a long vector.
- HMM+SDL[7]: We set the weighting parameters to the priors of the covariance matrices to 1 by following the original work.
- kPCA+HMM: We used a Gaussian kernel. The dimensionality of a new space was determined based on the eigenvalues so as to reflect more than 85% of the total variance.

For the competing methods of gICA, gSR, and PCFC, we further applied feature selection based on the paired t-test and used a linear support vector machine as classifier. For both HMM+SBL and kPCA+HMM, the number of hidden states was varied between 2 and 6 with a circular topology as the proposed method. To evaluate the performance, we conducted a leave-one-subject-out cross-valuation technique due to small sample sizes.

4.2 Performance Comparison

We considered five different metrics to compare performance among the competing methods and showed the results in Table 1. In a nutshell, the proposed method achieved the best accuracy of 81.08% with a sensitivity of 85.71% and a specificity of 80%.

[4] Available at '`https://github.com/bayesnet/bnt`'
[5] Available at '`http://www.nitrc.org/projects/gift`'
[6] The codes are available at '`http://miplab.epfl.ch/leonardi/`'
[7] The source codes were provided by the author of the original paper [2].

Table 1. A summary of the performances of the competing methods. The boldface denotes the best performance in each metric. (PPV: Positive Predictive Value; NPV: Negative Predictive Value)

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
gICA [12]	72.97	62.51	88.00	41.67	75.85
gSR [16]	75.68	80.00	75.00	33.33	96.00
PCFC [11]	75.68	71.43	**92.00**	41.67	76.67
HMM+SDL [2]	70.27	**100.0**	69.44	8.33	**100.0**
kPCA+HMM	70.27	60.00	**92.00**	25.00	71.88
Proposed method	**81.08**	85.71	80.00	**50.00**	96.00

Note that compared to HMM+SDL and kPCA+HMM, our method enhanced the classification accuracy by 10.81%. From a clinical perspective, since it is important to consider the prevalence of the disease, we also presented Positive Predictive Values (PPVs) and Negative Predictive Values (NPVs). Statistically, PPV and NPV measure, respectively, the proportion of subjects with MCI who are correctly diagnosed as patients and the proportion of subjects without MCI who are correctly diagnosed as cognitive normal. Our method achieved the PPV of 50% and the NPV of 96%, outperforming gICA by 8.33% (PPV) and 20.15% (NPV), gSR by 16.67% (PPV), PCFC by 8.33% (PPV) and 19.33% (NPV), and kPCA+HMM by 25% (PPV) and 24.12% (NPV). While HMM+SDL achieved a high NPV of 100%, its PPV was significantly lower than that of our method.

5 Conclusion

In this paper, we proposed a novel method to model functional dynamics in rs-fMRI for MCI identification. Specifically, we designed a deep network, by which we could discover the non-linear relationships among ROIs in a hierarchical manner and effectively reduce feature dimensionality. Meanwhile, by building generative models with HMMs for each class individually, we could estimate the feature likelihood of a test subject as MCI and NC, based on which we identified the clinical label. In our experiments, we achieved the highest performance with the proposed method, outperforming state-of-the-art methods in the literature. It is noteworthy that although it is not performed in this paper because of the limited space, by decoding the state sequence for the rs-fMRI data of a testing subject via Viterbi algorithm [13], we can construct functional connectivities, one for each hidden state, based on which further neurophysiological investigation can be conducted.

Acknowledgement. This work was supported by ICT R&D program of MSIP/IITP. [B0101-15-0307, Basic Software Research in Human-level Lifelong Machine Learning (Machine Learning Center)].

References

1. Biswal, B., Yetkin, F.Z., Haughton, V.M., Hyde, J.S.: Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magnetic Resonance in Medicine 34(4), 537–541 (1995)
2. Eavani, H., Satterthwaite, T.D., Gur, R.E., Gur, R.C., Davatzikos, C.: Unsupervised learning of functional network dynamics in resting state fMRI. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 426–437. Springer, Heidelberg (2013)
3. Faisan, S., Thoraval, L., Armspach, J.P., Heitz, F.: Hidden Markov multiple event sequence models: A paradigm for the spatio-temporal analysis of fMRI data. Medical Image Analysis 11(1), 1–20 (2007)
4. Handwerker, D.A., Roopchansingh, V., Gonzalez-Castillo, J., Bandettini, P.A.: Periodic changes in fMRI connectivity. NeuroImage 63(3), 1712–1719 (2012)
5. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation 18(7), 1527–1554 (2006)
6. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (2006)
7. Hjelm, R.D., Calhoun, V.D., Salakhutdinov, R., Allen, E.A., Adali, T., Plis, S.M.: Restricted Boltzmann machines for neuroimaging: An application in identifying intrinsic networks. NeuroImage 96, 245–260 (2014)
8. Hutchinson, R.A., Niculescu, R.S., Keller, T.A., Rustandi, I., Mitchell, T.M.: Modeling fMRI data generated by overlapping cognitive processes with unknown onsets using hidden process models. NeuroImage 46(1), 87–104 (2009)
9. Hutchison, R.M., Womelsdorf, T., Allen, E.A., Bandettini, P.A., Calhoun, V.D., Corbetta, M., Penna, S.D., Duyn, J.H., Glover, G.H., Gonzalez-Castillo, J., Handwerker, D.A., Keilholz, S., Kiviniemi, V., Leopold, D.A., de Pasquale, F., Sporns, O., Walter, M., Chang, C.: Dynamic functional connectivity: Promise, issues, and interpretations. NeuroImage 80, 360–378 (2013)
10. Janoos, F., Machiraju, R., Singh, S., Morocz, I.: Spatio-temporal models of mental processes from fMRI. NeuroImage 57(2), 362–377 (2011)
11. Leonardi, N., Richiardi, J., Gschwind, M., Simioni, S., Annoni, J.M., Schluep, M., Vuilleumier, P., Ville, D.V.D.: Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest. NeuroImage 83, 937–950 (2013)
12. Li, S., Eloyan, A., Joel, S., Mostofsky, S., Pekar, J., Bassett, S.S., Caffo, B.: Analysis of group ICA-based connectivity measures from fMRI: Application to Alzheimer's disease. PLoS One 7(11), e49340 (2012)
13. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
14. Rubinov, M., Sporns, O.: Complex networks measures of brain connectivity: Uses and interpretations. NeuroImage 52(3), 1059–1069 (2010)
15. Suk, H.I., Wee, C.Y., Lee, S.W., Shen, D.: Supervised discriminative group sparse representation for mild cognitive impairment diagnosis. Neuroinformatics, 1–19 (2014)
16. Wee, C.Y., Yap, P.T., Zhang, D., Wang, L., Shen, D.: Group-constrained sparse fMRI connectivity modeling for mild cognitive impairment identification. Brain Structure and Function 219(2), 641–656 (2014)