Discriminative Brain Effective Connectivity Analysis for Alzheimer's Disease: A Kernel Learning Approach upon Sparse Gaussian Bayesian Network

Luping Zhou[1], Lei Wang[1], Lingqiao Liu[2], Philip Ogunbona[1], Dinggang Shen[3]
School of Computer Science and Software Engineering, University of Wollongong, Australia[1]
Research School of Engineering, Australian National University, Australia[2]
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, NC, USA[3]

Abstract

Analyzing brain networks from neuroimages is becoming a promising approach in identifying novel connectivity-based biomarkers for the Alzheimer's disease (AD). In this regard, brain "effective connectivity" analysis, which studies the causal relationship among brain regions, is highly challenging and of many research opportunities. Most of the existing works in this field use generative methods. Despite their success in data representation and other important merits, generative methods are not necessarily discriminative, which may cause the ignorance of subtle but critical disease-induced changes. In this paper, we propose a learning-based approach that integrates the benefits of generative and discriminative methods to recover effective connectivity. In particular, we employ Fisher kernel to bridge the generative models of sparse Bayesian networks (SBN) and the discriminative classifiers of SVMs, and convert the SBN parameter learning to Fisher kernel learning via minimizing a generalization error bound of SVMs. Our method is able to simultaneously boost the discriminative power of both the generative SBN models and the SBN-induced SVM classifiers via Fisher kernel. The proposed method is tested on analyzing brain effective connectivity for AD from ADNI data, and demonstrates significant improvements over the state-of-the-art work.

1. Introduction

As the most common form of dementia, Alzheimer's disease (AD) is a fatal and progressive neurodegenerative disease that has caused serious socioeconomic problems in developed countries. Early diagnosis of AD may benefit the patients with disease-interrupted therapies when the dementia is still mild. Neuroimaging techniques are important in AD study because they may provide more sensitive and consistent measures than traditional cognitive assessment.

Currently, neuroimage analysis has evolved from studying local morphometry to complex relationships and interactions across brain regions. This is because the brain is, by nature, a complex network of many interconnected regions. A brain network is usually modeled by a graph with each node corresponding to a brain region and each edge corresponding to the connectivity between regions. The connectivity could be statistical dependencies (functional connectivity) or causal relationships (effective connectivity) [14], represented by *undirected* or *directed* graph, respectively. This paper focuses on brain *effective connectivity* analysis, an endeavor that has gained research interest due to its ability to analyze the directional effect of one brain region over another. Effective connectivity analysis has been applied to fMRI [5], PET [1], and gray matter morphology in structural MRI [6], and has exhibited promising potential in identifying novel connectivity-based biomarkers for AD.

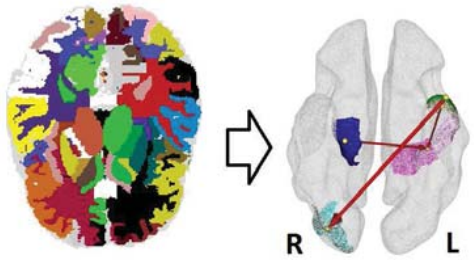

Figure 1. *Left: ROI partitions on MRI. Right: Some identified directional relationships discriminative for AD. (Please refer to Section 4.3. The figure is best viewed on monitor.)*

With sparseness techniques, effective connectivity analysis has been able to handle medium to large scale brain networks. A remarkable recent work is from Huang, et al. [1], where a sparse Gaussian Bayesian network (SGBN) is recovered from more than 40 brain regions in fluorodeoxyglucose PET (FDG-PET) images for AD analysis. That approach learns the Bayesian network (BN) structure and parameters simultaneously in one step, which demonstrates a more accurate network recovery than the conventional two-stage approaches in sparse BN learning (such as LIMB-DAG [13], MMHC [16], TC and TC-bw [9], etc.). Despite the effectiveness in network representation, the above methods (including [1]) are all generative methods. By their

nature, generative methods focus on representing an individual group, thus may not be discriminative. When analyzing brain networks, they are prone to over-emphasizing major structures within an individual group, and neglecting the subtle disease-induced structural changes across different groups. Therefore, generative methods are usually inferior in prediction compared with the discriminative methods that focus on the class boundary. However, discriminative methods are not amenable for interpretative analysis that is critical in exploratory research aimed at both understanding and diagnosing the disease. Therefore, *we aim to integrate the merits of generative and discriminative methods to learn BNs that are not only representative but also discriminative*. Recent progress in [10, 11] for learning discriminative BNs follows the conventional two-stage approach and works for discrete variables. They may not be suitable for brain network analysis where the brain regional measurements are usually continuous variables.

To achieve our goal, we improve the model of the SGBN in [1], and further boost its discriminative power via a kernel learning approach that links the generative SGBN with the SVM classifiers. This paper includes several contributions: 1) We propose an augmented SGBN model (A-SGBN) by revisiting the method in [1]. A-SGBN fits the underlying distribution more precisely, therefore bringing better prediction. 2) By inducing Fisher kernel on our A-SGBN models, we provide a way to obtain subject-specific SGBN-induced feature vectors that can be used by discriminative classifiers such as SVMs. Through this, we integrate the generative and discriminative models. 3) More significantly, we convert the learning of SGBN parameters to the learning of discriminative Fisher kernels, which makes the optimization simple. Specifically, we jointly learn the SGBN parameters and the separating hyperplane of SVMs over Fisher kernel by minimizing a generalization error bound of SVMs. 4) We apply our method on ADNI [1] data to analyze brain effective connectivity for AD from both T1-weighted MRI and FDG-PET images. Our method significantly improves the discriminative power of the generative SGBN and the discriminative SVM classifier simultaneously. 5) By Fisher kernel, we obtain a new kind of features that reflect the changing rate of connection strength, which have not been investigated in conventional approaches.

2. Background and Notation

2.1. Gaussian Bayesian Network

Gaussian Bayesian network (GBN) is the fundamental tool that we use to learn brain effective connectivity in this paper. It is therefore briefly described here, together with the definition of symbols used throughout the paper.

Let $\mathbf{x} = [x_1, x_2, \cdots, x_m]^\top$ be a sample of m features (variables). Let $\mathbf{D} \in \mathsf{R}^{n\times m}$ be a data matrix of n samples. The i-th row of $\mathbf{D}$ represents a sample $\mathbf{x}_i$. The j-th column of $\mathbf{D}$, denoted as $\mathbf{f}_j$, represents a realization of the j-th random variable x_j on the n samples.

A Bayesian network (BN) $\mathcal{G}$ is a directed acyclic graph (DAG) that expresses the factorization property of a joint distribution $p(\mathbf{x})$. With each variable corresponding to a node in $\mathcal{G}$, the joint distribution is factorized as $p(\mathbf{x}) = \prod_{i=1,\cdots,m} p(x_i|\mathbf{Pa}(x_i))$, where $\mathbf{Pa}(x_i)$ denotes the parent nodes of x_i. A GBN assumes that $p(x_i|\mathbf{Pa}(x_i))$ follows a Gaussian distribution. Each node x_i is regressed over its parent nodes $\mathbf{Pa}(x_i)$: $x_i = \boldsymbol{\theta}_i^\top \mathbf{Pa}(x_i) + \varepsilon_i$, where the vector $\boldsymbol{\theta}_i$ is the regression coefficients, and $\varepsilon_i \ \square \ \mathcal{N}(0, \sigma_i^2)$. The matrix $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_m]$ are called the parameters of a GBN. In this paper, following [1], a $m \times m$ matrix $\mathbf{G}$ is used to represent network structure, in which, if there is a direct *edge* from x_i to x_j, $\mathbf{G}_{ij} = 1$; otherwise, $\mathbf{G}_{ij} = 0$. In addition, another $p \times p$ matrix $\mathbf{P}$ is also kept to record all the directed *paths* in the structure. If there is a directed *path* from x_i to x_j, $\mathbf{P}_{ij} = 1$; otherwise $\mathbf{P}_{ij} = 0$.

2.2. Sparse Gaussian Bayesian Network

The state-of-the-art work for brain causal relationship analysis in [1] underpins our study in this paper. In [1], it is proposed to learn a sparse GBN (SGBN) for brain effective connectivity analysis utilizing FDG-PET images. Compared with the conventional BN methods that learn the network structure and parameters in two steps, SGBN simultaneously learns the structure and parameters by enforcing sparseness constraint on a GBN. This one-step learning approach outperforms the conventional two-step methods with higher accuracies for the network edge recovery. In particular, it is proposed in [1] to solve a constrained least-square fitting problem:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{m} \|\mathbf{f}_i - \boldsymbol{\theta}_i^\top \mathbf{Pa}(\mathbf{x}_i)\|_2^2 + \lambda_1 \|\boldsymbol{\theta}_i\|_1 \qquad (1)$$
$$s.t. \ \ \boldsymbol{\Theta}_{ji} \times \mathbf{P}_{ij} = 0, \forall i, j = 1, \cdots, m, \ \ i \neq j.$$

Here $\mathbf{f}_i$ and $\boldsymbol{\theta}_i$ are defined as above. The i-th row of the matrix $\mathbf{Pa}(\mathbf{x}_i)$ correspond to the parent nodes of x_i, which are initially set as all the nodes other than x_i, and further filtered implicitly by the sparseness constraint over their regression coefficients $\boldsymbol{\theta}_i$. In BN learning, a difficult problem is how to enforce the DAG property to ensure the validity of the resulting BN: there should be no directed cycles in the graph. In [1] it is proved that a sufficient and necessary condition for being a DAG is $\boldsymbol{\Theta}_{ji} \times \mathbf{P}_{ij} = 0$ for all i and j. The $\mathbf{P}_{ij}$ is computed by a Breadth-first search on $\mathbf{G}$ with x_i being the root node. For more details, please read [1].

[1] http://www.adni-info.org/

3. Proposed Method

In this paper, we study brain networks from two sources. The first source is gray matter morphology from T1-weighted MRI. It has been reported that the covariation of gray matter morphology might be related to the anatomical connectivity [15]. Studying brain morphology as a network can take the advantage of statistical tools from graph theory. The second source is FDG-PET images. The retention of tracer in FDG-PET is analogous to the glucose uptake, thus reflecting the tissue metabolic activity.

Building brain networks includes identifying network nodes and reconstructing the connectivity. This paper focuses on the latter. Hence, after briefing how network nodes are defined in our method in Section 3.1, we concentrate on how to infer the effective connectivity that is both representative (Section 3.2) and discriminative (Section 3.3).

3.1. Network Nodes Determination

MRI. This study involves 120 subjects including 50 MCI (mild cognitive impairment, a prodromal AD) patients and 70 NC (normal controls) from the publicly accessible data of ADNI. The T1-weighted MR images are segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) using FAST in the FSL[2] package after intensity correction, skull stripping [17], and cerebellum removal. These tissue-segmented images are spatially normalized into a template space by HAMMER [3], and partitioned into 100 Region of Interest (ROI) via an ROI atlas [3]. We use the GM volumes of each ROI as network nodes, and select 40 ROIs that have the highest correlation with class labels into our study.

PET. This study involves 103 subjects including 51 AD patients and 52 NC whose FDG-PET and MR images are downloaded from ADNI. We first co-register the MR images into a template space and partition them into ROIs as mentioned above. Then the PET images are aligned with their MR images from the same subject by a rigid transformation. The average tracer uptakes within each ROI are used as network nodes. Similarly, we select 40 ROIs that are most discriminative with regards to AD.

3.2. SGBN Model Augmentation

A simple way to use generative BNs for prediction is to train each class a BN individually and classify a new sample $\mathbf{x}_i$ by assigning it to the class with a higher likelihood. The more precisely the BN model reflects the underlying distribution, the more accurate the prediction is. To compare the likelihood for each class in the same space, the data should not be normalized separately for each class as in [1] where a single class is the focus.

[2]http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/

[3]http://www.med.unc.edu/bric/ideagroup/tools/projects-1/brain/pages-1/hammer

To handle this, we introduce a bias term x_0 in the regression, i.e., $x_i = \boldsymbol{\theta}_i^\top [\mathbf{Pa}(x_i), x_0] + \varepsilon_i$, and demonstrate it not a trivial improvement over the case when directly applying the SGBN in [1]. Accordingly, in the graph $\mathcal{G}$, a bias node is added. It has no parent but is the parent of all the other nodes. If originally $\mathcal{G}$ is a DAG, adding x_0 in this way does not cause the violation of DAG. To be distinguished from SGBN in [1], we call ours A-SGBN. Intuitively, there could be two reasons to include such a bias node into a brain network: i) there possibly exist some latent variables related to the disease, which are not included in the current study, and their influences may be absorbed by the bias node; or ii) the state of a node may depend not only on the interactions with other nodes, but also on the prior of itself. Our experiment in Section 4 demonstrates that A-SGBN is a more precise model than SGBN (smaller fitting errors for both training and test data) for our case, and effectively improves the classification. In addition to the advantages of A-SGBN over SGBN, in the following we show that actively learning the discrimination can further boost the classification performance.

3.3. Discriminative SBN Learning via Fisher Kernel

Both SGBN and A-SGBN learn the brain networks for AD or NC separately. This may ignore some subtle but critical network differences that distinguish the two classes. We argue that the parameters of the generative model should be learned from the two classes jointly to keep the essential discrimination. This can be achieved by maximizing the posterior probability $p(y|\mathbf{x})$, where y is the class label of $\mathbf{x}$. Although conceptually direct, this approach often leads to complicated optimization problems. This paper takes another approach. Specifically, *we employ Fisher kernel to extract feature vectors from the SGBN models of two classes, and then convert the model parameter learning to Fisher kernel learning with SVMs*. We find that the SGBN-induced Fisher vector (see below) is a linear function of parameters $\boldsymbol{\Theta}$, which well simplifies the optimization.

3.3.1 Induction of Fisher vectors from SGBN

Below we introduce how to use Fisher kernel on SGBNs to obtain feature vectors used for kernel learning.

Fisher kernel provides a way to compare samples induced by a generative model. It maps a sample to a feature vector in the gradient space of the model parameters. The intuition is that similar objects induce similar log-likelihood gradients of the model parameters. Fisher kernel is computed as $K(\mathbf{x}, \mathbf{x}') = \mathbf{g}_\mathbf{x}^\top \mathbf{U}^{-1} \mathbf{g}_{\mathbf{x}'}$, where the Fisher vector $\mathbf{g}_\mathbf{x} = \nabla_{\boldsymbol{\theta}} \log(p(\mathbf{x}|\boldsymbol{\theta}))$ describes the changing direction of parameters to better fit the model. The Fisher information metric $\mathbf{U}$ weights the similarity measure, but is often set as an identity matrix in practice [2].

Fisher kernel has recently witnessed successful applica-

tions in image categorization [12, 4] for inducing feature vectors from Gaussian Mixture Model (GMM) of a visual vocabulary. Despite its success, to the best of our knowledge, Fisher kernel has not been applied to BN for brain connectivity analysis. More importantly, in the applications above, there is no discriminative learning for Fisher kernel as in this paper. The advantage of discriminative Fisher kernel has also been confirmed by a very recent study that uses a different learning criterion within a different context [8].

Following [1], we only consider $\boldsymbol{\Theta}$ as parameters and predefine σ. Let $\mathcal{L}(\mathbf{x}|\boldsymbol{\Theta}) = \log(p(\mathbf{x}|\boldsymbol{\Theta}))$ denote the log-likelihood. Our Fisher vector for each sample $\mathbf{x}$ is $\Phi_{\boldsymbol{\Theta}}(\mathbf{x}) = [\nabla_{\boldsymbol{\Theta}_1}\mathcal{L}(\mathbf{x}|\boldsymbol{\Theta}_1)^\top,\ \nabla_{\boldsymbol{\Theta}_2}\mathcal{L}(\mathbf{x}|\boldsymbol{\Theta}_2)^\top]^\top$, where $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are the parameters of the SGBNs for the two classes ($y = 1, 2$), respectively. Recall that, using a BN, the probability $p(\mathbf{x}|\boldsymbol{\Theta})$ can be factorized as $p(\mathbf{x}|\boldsymbol{\Theta}) = \prod_{i=1,\cdots,m} p(x_i|\mathbf{Pa}(x_i), \boldsymbol{\theta}_i)$. Therefore, it holds that

$$\begin{aligned}\mathcal{L}(\mathbf{x}|\boldsymbol{\Theta}) &= \sum_{i=1}^{m} \log p(x_i|\mathbf{Pa}(x_i), \boldsymbol{\theta}_i) \\ &= \sum_{i=1}^{m} \frac{-(x_i - \boldsymbol{\theta}_i^\top \mathbf{Pa}(x_i))^2}{2\sigma_i^2} - \log(2\pi\sqrt{\sigma_i}).\end{aligned} \tag{2}$$

Taking partial derivative over $\boldsymbol{\theta}_i$, we have

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{x}|\boldsymbol{\Theta})}{\partial \boldsymbol{\theta}_i} &= -\frac{\mathbf{Pa}(x_i)\mathbf{Pa}(x_i)^\top}{\sigma_i^2}\boldsymbol{\theta}_i - \frac{x_i \mathbf{Pa}(x_i)}{\sigma_i^2} \\ &\triangleq \mathbf{S}(x_i)\boldsymbol{\theta}_i + \mathbf{s}_0(x_i),\end{aligned} \tag{3}$$

where $\mathbf{S}(x_i)$ is a matrix and $\mathbf{s}_0(x_i)$ is a vector. As shown, $\Phi_{\boldsymbol{\Theta}}(\mathbf{x})$ is a *linear* function of $\boldsymbol{\Theta}$. This simple form of $\Phi_{\boldsymbol{\Theta}}(\mathbf{x})$ significantly facilitates our further kernel learning.

3.3.2 Discriminative Fisher kernel learning via SVM

As each Fisher vector is a function of the SGBN parameters, discriminatively learning these parameters can thus be converted to learning discriminative Fisher kernels. We require that the learned SGBN models possess the following properties. Firstly, the Fisher vectors induced by the learned SGBN model should be well separated between classes. Secondly, the learned SGBN models should maintain reasonable capacity of representation. Thirdly, the learned SGBN models should not violate DAG.

We use the following strategies to achieve our goal. Firstly, to obtain a discriminative Fisher kernel, we jointly learn the parameters of SGBN and the separating hyperplane of SVMs with Fisher kernel. Radius-margin bound, the upper bound of the Leave-One-Out error, is minimized to keep good generalization of the SVMs. Secondly, to maintain reasonable representation, we explicitly control the fitting errors of the learned model during optimization. Thirdly, we enforce the DAG constraint in [1] to ensure the validity of the graph. For convenience, we call our method DL-A-SGBN. More details are given below.

In order to use radius-margin bound, $\mathcal{L}_2$-SVM with soft margin has to be employed, which optimizes

$$\begin{aligned}&\min_{\mathbf{w},\boldsymbol{\xi}}\ \frac{1}{2}\|\mathbf{w}\|_2^2 + C\boldsymbol{\xi}^\top\boldsymbol{\xi} \\ &s.t.\ \ y_i(\mathbf{w}^\top\Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,\ \ \xi_i \geq 0,\ \ \forall i\end{aligned} \tag{4}$$

Following the convention in SVMs, $\mathbf{x}_i$ is the i-th sample with class label y_i, $\mathbf{w}$ the normal of separating plane, b the bias term, $\boldsymbol{\xi}$ the slack variables and C the regularization parameter. $\mathcal{L}_2$-SVM can be rewritten as SVM with hard margin by slightly modifying the kernel $\mathbf{K} := \mathbf{K} + \mathbf{I}/C$, where $\mathbf{I}$ is identity matrix. For convenience, in the following, we redefine $\Phi(\mathbf{x}_i) := [\Phi^\top(\mathbf{x}_i)\ \ \mathbf{e}_i^\top/\sqrt{C}]^\top$. The vector $\mathbf{e}_i$ has the value of 1 at the i-th element, and 0 elsewhere.

Incorporating radius information leads to solving

$$\begin{aligned}&\min_{\mathbf{w}}\ \frac{1}{2}R^2\|\mathbf{w}\|_2^2 \\ &s.t.\ \ y_i(\mathbf{w}^\top\Phi(\mathbf{x}_i) + b) \geq 1,\ \ \forall i,\end{aligned} \tag{5}$$

where R^2 denotes the radius of Minimal Enclosing Ball (MEB). It has been observed that when the sample size is small, the estimation of R^2 may become noisy and unstable. Therefore, it has been proposed to use trace-based scatter matrix instead for such cases [7]. We optimize

$$\begin{aligned}&\min_{\boldsymbol{\theta},\mathbf{w}}\ \frac{1}{2}\mathrm{tr}(\mathbf{S}_T)\|\mathbf{w}\|_2^2 \\ &s.t.\ \ y_i(\mathbf{w}^\top\Phi_{\boldsymbol{\Theta}}(\mathbf{x}_i) + b) \geq 1,\ \ \forall i \\ &\quad h(\mathbf{D}_1, \boldsymbol{\Theta}_1) \leq T_1,\ \ h(\mathbf{D}_2, \boldsymbol{\Theta}_2) \leq T_2, \\ &\quad \boldsymbol{\Theta}_1 \in DAG,\ \ \boldsymbol{\Theta}_2 \in DAG.\end{aligned} \tag{6}$$

Here $\mathrm{tr}(\mathbf{S}_T)$ is the trace of the total scatter matrix $\mathbf{S}_T$, where $\mathbf{S}_T = \sum_{i=1}^{n}(\Phi(\mathbf{x}_i) - \mathbf{m})(\Phi(\mathbf{x}_i) - \mathbf{m})^\top$, and $\mathbf{m}$ is the mean of total n samples in the kernel-induced space. It can be shown that $\mathrm{tr}(\mathbf{S}_T) = \mathrm{tr}(\mathbf{K}) - \mathbf{1}^\top\mathbf{K}\mathbf{1}/n$, where $\mathbf{1}$ denotes a vector whose elements are all 1, and $\mathbf{K}$ the kernel matrix. Fisher vector $\Phi_{\boldsymbol{\Theta}}(\mathbf{x}_i)$ is obtained as mentioned in Section 3.3.1. The function $h(\cdot)$ measures the squared fitting errors of the corresponding SGBNs for the data $\mathbf{D}_1$ and $\mathbf{D}_2$ from the two classes. It is defined as

$$h(\mathbf{D}, \boldsymbol{\Theta}) = \sum_{i=1}^{m} \|\mathbf{f}_i - \boldsymbol{\theta}_i^\top \mathbf{Pa}(\mathbf{x}_i)\|_2^2,$$

where all the symbols are defined as in Eqn. (1). The two user-defined parameters T_1 and T_2 explicitly control the degree of fitting during the learning process (Section 4.2). The DAG constraints here are the same to that used in Eqn.(1). Recall that the DAG constraint is $\boldsymbol{\Theta}_{ji} \times \mathbf{P}_{ij} = 0$, where

$\mathbf{P}_{ij} = \{0, 1\}$, reflecting the structure of $\mathbf{\Theta}$. It is observed that enforcing DAG in this way has somewhat enforced the graph sparsity. Therefore, to avoid complicating our optimization we do not impose additional sparseness constraints on $\mathbf{\Theta}$ here. Our A-SGBN could serve as a good initial solution for this problem.

One possible approach for solving Eqn. (6) is to alternately optimize the separating hyperplane $\mathbf{w}$ and the parameter $\mathbf{\Theta}$. That is,

$$\begin{aligned} &\min_{\theta} \ J(\mathbf{\Theta}) \\ &s.t. \ h(\mathbf{D}_1, \mathbf{\Theta}_1) \leq T_1, \ \ h(\mathbf{D}_2, \mathbf{\Theta}_2) \leq T_2, \\ &\quad \mathbf{\Theta}_1 \in DAG, \ \ \mathbf{\Theta}_2 \in DAG. \end{aligned} \tag{7}$$

where

$$\begin{aligned} J(\mathbf{\Theta}) = \min_{\mathbf{w}} \ &\frac{1}{2}\mathrm{tr}(\mathbf{S}_T)\|\mathbf{w}\|_2^2 \\ s.t. \ &y_i(\mathbf{w}^\top \Phi_{\mathbf{\Theta}}(\mathbf{x}_i) + b) \geq 1, \ \ \forall i. \end{aligned} \tag{8}$$

Note that for a given $\mathbf{\Theta}$, the term $\mathrm{tr}(\mathbf{S}_T)$ is constant to Eqn. (8). Due to the strong duality in SVM optimization, we solve the term $\|\mathbf{w}\|_2^2$ by

$$\begin{aligned} J_0(\mathbf{\Theta}) = \max_{\alpha} &\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K_{\mathbf{\Theta}}(\mathbf{x}_i, \mathbf{x}_j) \\ s.t. &\sum_{i=1}^{n} \alpha_i y_i = 0, \ \ \alpha_i \geq 0 \ \ \forall i, \end{aligned} \tag{9}$$

where α_i is the Lagrangian multiplier. Many quadratic programming packages could be used to solve Eqn. (7). We use fmincon-SQP (sequential quadratic programming) in Matlab. Our learning process is summarized in Table 1.

Table 1. **Discriminatively Learning $\mathbf{\Theta}$**

Input: $\mathbf{\Theta}^{(0)}$ estimated by A-SGBN
Output: $\mathbf{\Theta}^\star$ learned by Eqn. (7)
1. Let $\mathbf{\Theta}^{(t-1)} = \mathbf{\Theta}^{(0)}$
2. Compute $\Phi_{\mathbf{\Theta}}^{(t-1)}$ and $\mathbf{K}_{\mathbf{\Theta}}^{(t-1)}$ by Eqn. (3)
3. Compute $\mathrm{tr}(\mathbf{S}_T)^{(t-1)} = \mathrm{tr}(\mathbf{K}_{\mathbf{\Theta}}^{(t-1)}) - \mathbf{1}^\top \mathbf{K}_{\mathbf{\Theta}}^{(t-1)} \mathbf{1}/n$
4. Solve $J_0(\mathbf{\Theta}^{(t-1)})$ and $\alpha^\star$ by Eqn. (9)
5. $J(\mathbf{\Theta}^{(t-1)}) = J_0(\mathbf{\Theta}^{(t-1)}) \times \mathrm{tr}(\mathbf{S}_T)^{(t-1)}$
6. Compute $\nabla_{\mathbf{\Theta}^{(t-1)}} J$ by Eqn. (10)
7. For a given $\alpha^\star$, minimize Eqn. (7) using $J(\mathbf{\Theta}^{(t-1)})$ and $\nabla_{\mathbf{\Theta}^{(t-1)}} J$; Obtain the optimal $\mathbf{\Theta}^{(t)}$
8. Let $\mathbf{\Theta}^{(t-1)} = \mathbf{\Theta}^{(t)}$
9. Repeat Step 2-8 until convergence, let $\mathbf{\Theta}^\star = \mathbf{\Theta}^{(t)}$

3.3.3 Discussion

Gradient of Eqn. (7). Gradient information is required by many optimization algorithms (including fmincon-SQP) to speed up the line search. In our case, the gradient of the objective function in Eqn. (7) can be simply calculated as

$$\begin{aligned} \nabla_{\mathbf{\Theta}} J = &-\frac{1}{2}\mathrm{tr}(\mathbf{S}_T)\sum_{ij} \alpha_i^\star \alpha_j^\star y_i y_j \nabla_{\mathbf{\Theta}} K_{\mathbf{\Theta}}(\mathbf{x}_i, \mathbf{x}_j) \\ &+ J_0(\mathbf{\Theta})\left(\mathbf{I} + \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\nabla_{\mathbf{\Theta}} K_{\mathbf{\Theta}}(\mathbf{x}_i, \mathbf{x}_j), \end{aligned} \tag{10}$$

where $\alpha^\star$ maximizes Eqn. (9). The symbols $\mathbf{I}$ and $\mathbf{1}$ are defined as before. The terms $\mathrm{tr}(\mathbf{S}_T)$ and $J_0(\mathbf{\Theta})$ have been computed when evaluating the objective function $J(\mathbf{\Theta})$ in Eqn.(7), thus introducing no additional computational cost. $\nabla_{\mathbf{\Theta}} K_{\mathbf{\Theta}}(\mathbf{x}_i, \mathbf{x}_j)$ is just a linear function of $\mathbf{\Theta}$:

$$\begin{aligned} \frac{\partial K_{\mathbf{\Theta}}(\mathbf{x}_i, \mathbf{x}_j)}{\partial \theta_l} = &[\mathbf{S}(x_{il})^\top \mathbf{S}(x_{jl}) + \mathbf{S}(x_{jl})^\top \mathbf{S}(x_{il})]\theta_l \\ &+ (\mathbf{S}(x_{jl})^\top \mathbf{s}_0(x_{il}) + \mathbf{S}(x_{il})^\top \mathbf{s}_0(x_{jl})), \end{aligned} \tag{11}$$

where x_{il} denotes the l-th feature of the i-th sample, and $\mathbf{S}$ and $\mathbf{s}_0$ are defined in Eqn. (3).

Variable selection. Learning the whole set of SGBN parameters may encounter the "curse of dimensionality" when the training samples are insufficient. For example, we have less than 100 training samples, but 3600 parameters (from two classes) to learn. This may cause overfitting and make the estimation unstable. To handle this issue, we hypothesize that, learning only a selected subset of parameters may mitigate the overfitting and improve the discrimination. For this purpose, $\mathbf{\Theta}$ is partitioned into two parts: $\mathbf{\Theta} = \{\mathbf{\Theta}_{sel}, \mathbf{\Theta}_{nosel}\}$. We keep using the whole $\mathbf{\Theta}$ for computing $\mathbf{K}_{\mathbf{\Theta}}$, but optimize Eqn. (7) only over $\mathbf{\Theta}_{sel}$. There are many options to determine $\mathbf{\Theta}_{sel}$. We initially compute the Pearson correlation between each component of $\Phi_{\mathbf{\Theta}}$ and the class labels on the training data, and select the top θ_i with the highest correlations. To keep our problem simple, only the parameters associated with edges present in the graph are optimized. In this way, the optimization may only eliminate but never add edges in the graph, which avoids the violation of DAG, as well as maintaining the sparsity of the initial A-SGBN. It is remarkable that even this simple selection process has been able to greatly improve the discrimination experimentally.

Extension. Although focusing on each node corresponding to a scalar ROI feature, our method is readily extendable to handle multiple features (feature vector) of an ROI. In this case, the conditional distribution for node i becomes $p(\mathbf{x}_i|\mathbf{PA}(\mathbf{x}_i)) = \mathcal{N}(\mathbf{x}_i|\sum_{\mathbf{x}_j \in \mathbf{PA}(\mathbf{x}_i)} \mathbf{M}_{ij}\mathbf{x}_j, \mathbf{\Sigma}_i)$, where $\mathbf{PA}$ and $\mathbf{M}$ are both matrices. Our learning remains the same. In our future work, we will apply this extension to analyze fMRI where each ROI is associated with a vector of temporal signal.

4. Experiment

We evaluate our proposed A-SGBN and DL-A-SGBN against the baseline SGBN (B-SGBN) from [1] (without normalizing the data) in three aspects: i) model fitting, ii) discrimination, and iii) connectivity. Three data sets are used in our experiment: the MRI and FDG-PET data mentioned in Section 3.1, and another MRI-II data that uses the MR images from the same subjects as MRI, but involves 40 different ROIs. Although not as discriminative as that in MRI, the ROIs in MRI-II are more spread across the frontal, parietal, occipital and frontal lobes, thus specially used for a detailed lobe-to-lobe comparison on connectivity. We randomly partition each data set into 30 groups of training-test pairs. Each group includes 80 training and 40 test samples in MRI and MRI-II, or 60 training and 43 test samples in PET.

4.1. Comparison of Fitting

Our DL-A-SGBN targets to become discriminative without sacrificing too much power of data representation compared with B-SGBN. Since the change of data fitting from A-SGBN to DL-A-SGBN has been explicitly controlled by the user-defined parameters T_1 and T_2, we simply compare the model fitting between A-SGBN and B-SGBN. The fitting errors are tested on both training and test data for each class in all three data sets. The root of mean squared fitting errors (RMS) are summarized in Table 2. In order to test if the fitting errors of A-SGBN are statistically different from that of B-SGBN, a paired t-test (two-tailed) is conducted on the fitting errors over the 30 groups for each data set, respectively. The resulting p-value is also given in the last column in Table 2.

Table 2. ***Fitting Error (RMS) Averaged over 30 Training-Test Groups***

MRI		B-SGBN	A-SGBN	p-value
Training	MCI	0.6344	**0.6192**	0
	NC	0.5962	**0.5896**	0
Test	MCI	0.7385	**0.7301**	0
	NC	0.6801	**0.6763**	3.2e-4
PET		**B-SGBN**	**A-SGBN**	**p-value**
Training	NC	0.5466	**0.5334**	0
	AD	0.6291	**0.6195**	0
Test	NC	0.6171	**0.6100**	5e-8
	AD	0.7508	**0.7467**	2.2e-6
MRI-II		**B-SGBN**	**A-SGBN**	**p-value**
Training	MCI	0.6756	**0.6675**	0
	NC	0.6441	**0.6382**	0
Test	MCI	0.8047	**0.8033**	0.055
	NC	0.7381	**0.7366**	9.1e-4

As shown, on all three data sets, our A-SGBN fits the data consistently better than B-SGBN. Such improvement is significant as indicated by the small p-values (except for MCI group in MRI-II). This finding indicates that our A-SGBN might better reflect the underlying distribution of the data, which makes it perform well on both the training and the test data. Another interesting finding is that the generative models explain the NC better than the MCI (in MRI data set) or the AD (in PET data set) patients. This may reflect the common impression that compared with the healthy population, the AD population might be more heterogeneous and therefore more difficult to be represented by a single Gaussian model.

4.2. Comparison of Discrimination

Our proposed learning process results in two kinds of models: two DL-A-SGBN models with one for each class, and one SBN-induced SVM classifier that considers only the boundary of the two classes. We test whether our learning can improve the discriminative power on both kinds. The A-SGBN models estimated separately for each class are used as the initial solution. In order to keep reasonable interpretation, we allow maximal 1% additional squared fitting errors (that is, $T_i = 1.01 \times T_{i0},\ (i = 1, 2)$, where T_{i0} is the squared fitting error of the initial solution) to be introduced during the learning of DL-A-SGBN. We test both the SVM classifier and the DL-A-SGBNs. For the SVM classifier, we use $\mathcal{L}_2$-SVM with Fisher kernels . For DL-A-SGBNs, as mentioned before, we simply compare the values of likelihood, and assign the sample to the class with a higher likelihood. We also conduct a paired t-test (two-tailed) to examine the statistical significance of the improvement over the 30 groups for all three data sets. The results are summarized in Table 3.

Table 3. ***Test Classification Accuracy ($\%$) Averaged over 30 Training-Test Groups***

	SGBN-induced SVM classifier		
	A-SGBN ($\%$)	DL-A-SGBN ($\%$)	p-value
MRI	71.42	**74.50**	6.2e-5
PET	57.75	**65.43**	0
MRI-II	57.25	**61.83**	1.4e-6
	SGBN classifier		
	A-SGBN ($\%$)	DL-A-SGBN ($\%$)	p-value
MRI	71.08	**74.83**	6.8e-5
PET	67.36	**71.47**	4.7e-7
MRI-II	59.75	**65.42**	1.2e-6

It can be seen that, as expected, optimizing Eqn. (6) significantly improves the discriminative power of SVM classifiers by 3.08% for MRI, 7.68% for PET, and 4.58% for MRI-II. More importantly, by learning a discriminative SVM classifier, we also simultaneously improve the discriminative power of the generative models DL-A-SGBN

by 3.75% for MRI, 4.11% for PET, and 5.67% for MRI-II. Such improvements are statistically significant as indicated by the small p-values. Moreover, when cross-referencing the third columns in Table 3, it is noticed that our SVM classifiers perform just comparably (for MRI) or even worse (for PET and MRI-II) than our generative DL-A-SGBNs. This may be because our Fisher vectors have very high dimensionality, which causes the serious overfitting of data in SVM classifiers. Such situation might be somewhat improved for DL-A-SGBN since the simple Gaussian model may "regularize" the fitting. Based on this assumption, we further select a number of leading features from Fisher vectors by computing the Pearson correlation of the features and the labels, and use the selected features to construct the Fisher kernel for the SVM classifiers. As shown in the last column in Table 4, the simple feature selection step can further significantly improve the classification performance of the Fisher-kernel based SVM: from 74.5% to 80.08% for MRI, from 65.43% to 77.83% for PET, and from 61.83% to 73.5% for MRI-II.

In Table 4, the improvement from our proposed learning method is scrutinized at each processing step. Compared with the B-SGBN induced from [1], introducing a bias node (A-SGBN) better fits the population, therefore improves the prediction on test data by 9% for MRI, 6.04% for PET, and 6.67% for MRI-II. The discriminative power of A-SGBN is further improved by $3 \sim 6\%$ via our discriminative parameter learning. This leads to generative models DL-A-SGBN achieving a classification accuracy above 70%, with no more than 1% increase of the squared fitting error. Moreover, by selecting leading features in the SGBN-induced Fisher vectors, we can construct more discriminative SVM classifiers with additional 6% or more improvement from our DL-A-SGBN to differentiate both MCI vs NC groups in MRI or MRI-II and AD vs NC groups in PET.

In sum, compared with the baseline B-SGBN, our proposed method can increase the prediction accuracy by as high as **18%** for MRI, **16%** for PET, and **20%** for MRI-II, using Fisher kernel induced SVM classifiers with feature selection. Meanwhile, these SVM classifiers are linked to the learned generative model DL-A-SGBN whose discriminative powers have also been increased by about 10% from B-SGBN on all three data sets. Our DL-A-SGBN models are not only discriminative, but also descriptive with only a slight increase in the squared fitting errors (at most 1% increase, controlled by the optimization parameter).

Table 4. ***Test Classification Accuracy (%) Averaged over 30 Training-Test Groups***

	B-SGBN	A-SGBN	DL-A-SBN	SVM (sel)
MRI	62.08	71.08	74.83	**80.08**
PET	61.32	67.36	71.47	**77.83**
MRI-II	53.08	59.75	65.42	**73.50**

4.3. Comparison of Connectivity

In order to gain more insight into the results, we also conduct a lobe-to-lobe comparison on the connectivity derived by our methods and B-SGBN. It is found that, although the 40 ROIs used in MRI and PET are individually discriminative, they do not necessarily cover the representative regions across the whole brain. For example, the 40 nodes used in the MRI data set are mostly located in the temporal lobe and the subcortical region. Therefore, we specially design the MRI-II data set by selecting 40 regions that cover the frontal, parietal, occipital and temporal (including the subcortical region) lobes from MR images of the same subjects involved in MRI data. Although MRI-II (with the best test accuracy of 73.5%) is less discriminative than MRI (with the best test accuracy of 80.08%) as shown in Table 4, we consistently observe significant improvements of our method over B-SGBN.

The structures of the brain networks recovered from NC and MCI groups are displayed in Fig. 2 by using B-SGBN and DL-A-SGBN, respectively. The network structure is obtained by binarizing the edges $\mathbf{\Theta}$ with a threshold of 0.01. Each row i represents the effective connections (dark dots) starting from the node i, and each column j represents the effective connections ending at the node j.

With similar parameter settings, the B-SGBN produces 273 edges for NC, and 224 edges for MCI, while our DL-A-SGBN produces 285 edges for NC, and 236 edges for MCI. Note that DL-A-SGBN has an additional bias node corresponding to the last row and column. Because the bias node has no parent node, the last column is all zero. We check the edge difference between the two methods lobe by lobe, and give the result in Table 5. As shown, the two methods produce similar network structures both visually and quantitatively in most brain regions. There are in total 36 different edges (less than 15%) for NC network, and 11 different edges (around 5%) for MCI network. About half different connections are identified within the temporal lobe (15 for NC, 5 for MCI), for which we also include subcortical structures such as hippocampus and amygdala. It is known that temporal lobe (and some subcortical structures) plays a very important role in the progression of AD. Such a structural difference in this lobe may potentially reflect the different capacity of prediction between our DL-A-SGBN and the B-SGBN.

Table 5. ***Number of edge difference between the baseline B-SGBN and the proposed DL-A-SGBN in two groups, respectively: NC (MCI)***

	Frontal [1:8]	Parietal [9:16]	Occipital [17:24]	Temporal [25:40]
Frontal	1 (0)	1 (1)	3 (0)	0 (0)
Parietal	0 (1)	0 (0)	3 (0)	4 (0)
Occipital	0 (0)	1 (0)	2 (1)	5 (1)
Temporal	0 (0)	0 (1)	1 (1)	**15 (5)**

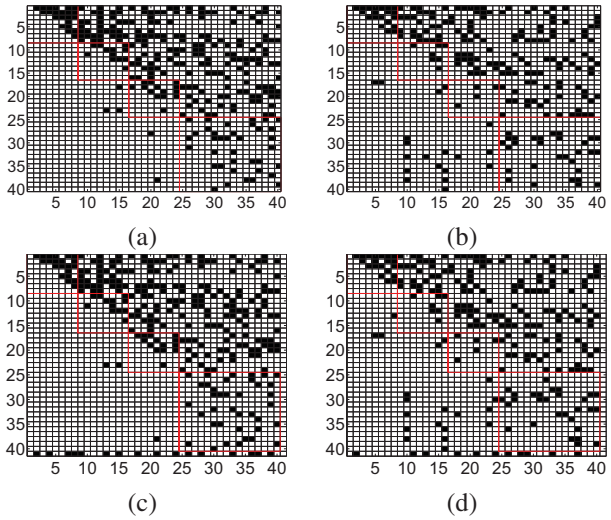

Figure 2. *Structure of Connectivity: (a) NC by B-SGBN, (b) MCI by B-SGBN, (c) NC by the proposed DL-A-SGBN, (d) MCI by the proposed DL-A-SGBN.*

Traditional brain connectivity analysis focuses on the analysis of brain structure which is a binarized connectivity. For example, the network structures from both the B-SGBN and our DL-A-SGBN indicate the loss of effective connections (around 17%) in MCI group in almost all lobes (slightly in the frontal lobe), which agrees well with documented studies [1, 6]. However, binarizing connectivity depends on the selection of threshold. If some connection strength has been weakened by the disease but not reduced below the threshold, this change will be unnecessarily ignored when merely studying the brain structure. This observation is affirmed by our learning process that promotes the discrimination of A-SGBN. Simply optimizing the connection strength across a subset of selected nodes has already significantly improved the prediction with only a minimum (or mostly no) change of brain structure.

Moreover, using SGBN-induced Fisher kernels, we are able to produce a new kind of features to analyze brain connectivity: the subject specific change of connection strength between nodes. We investigate the selected features of MRI-II used in our SBN-induced SVM classifier and visualize three most discriminative connection changes (Fig. 1 right) happening at "middle temporal gyrus left" (in brown) →"superior parietal lobe left" (in purple), "hippocampus right" (in blue) →"superior parietal lobe left", and "inferior temporal gyrus left" (in green) →"middle occipital gyrus right" (in cyan). Also discriminative are the connections from the bias node to "middle occipital gyrus right" and to "precuneus left".

5. Conclusion

In this paper, we present an approach to model brain effective connectivity encoded with essential discriminative information. With the link of Fisher kernel, our approach is able to simultaneously produce generative SGBN models and its associated SVM classifier, both of which possess sufficient discriminative power for brain network analysis of AD. In addition, by considering the changing rate of connection strength, our method also provides a new perspective for brain connectivity analysis.

References

[1] S. Huang, J. Li, J. Ye, A. Fleisher, K. Chen, T. Wu, and E. Reiman. A sparse structure learning algorithm for gaussian bayesian network identification from high-dimensional data. *IEEE TPAMI*, 2012.

[2] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, 1998.

[3] N. Kabani, J. MacDonald, C. Holmes, and A. Evans. A 3d atlas of the human brain. *Neuroimage*, 7:S7–S17, 1998.

[4] J. Krapac, J. Verbeek, and F. Jurie. Modeling spatial layout with fisher vectors for image categorization. In *ICCV*, 2011.

[5] R. Li, X. Wu, K. Chen, and A. e. Fleisher. Alterations of directional connectivity among resting-state networks in alzheimer disease. *Am J Neuroradiol*, 2012.

[6] X. Li, D. Coyle, L. Maguire, D. Watson, and T. McGinnity. Gray matter concentration and effective connectivity changes in alzheimer's disease: A longitudinal structural mri study. *Neuroradiology*, 53(10):733–748, 2011.

[7] X. Liu, L. Wang, J. Yin, E. Zhu, and J. Zhang. An efficient approach to integrating radius information into multiple kernel learning. *IEEE. TSMC-B*, 2012.

[8] L. Maaten. Learning discriminative fisher kernels. In *ICML*, pages 217–224, 2011.

[9] J. Pellet and A. Elisseeff. Using markov blankets for causal structure learning. *JMLR*, 9:1295–1342, 2008.

[10] F. Pernkopf and J. Bilmes. Efficient heuristics for discriminative structure learning of bayesian network classifiers. *JMLR*, 11:2323–2360, 2010.

[11] F. Pernkopf, M. Wohlmayr, and S. Tschiatschek. Maximum margin bayesian network classifiers. *IEEE TPAMI*, 34(3):521–532, 2012.

[12] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[13] M. Schmidt, A. Niculescu-Mizil, and K. Murphy. Learning graphical model structures using l1-regularization paths. In *AAAI*, 2007.

[14] O. Sporns. Brain connectivity. *Scholarpedia*, 2(10):4695, 2007.

[15] B. Tijms, P. Seris, D. Willshaw, and S. Lawrie. Similarity-based extraction of individual networks from gray matter mri scans. *Cereb Cortex*, 22(7):1530–1541, 2012.

[16] I. Tsamardinos, L. Brown, and C. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

[17] Y. Wang, J. Nie, P. Yap, F. Shi, L. Guo, and D. Shen. Robust deformable-surface-based skull-stripping for large-scale studies. In *MICCAI*, 2011.