Max-Margin Based Learning for Discriminative Bayesian Network from Neuroimaging Data

Luping Zhou[1], Lei Wang[1], Lingqiao Liu[2], Philip Ogunbona[1], and Dinggang Shen[3]

[1] University of Wollongong, Australia
[2] University of Adelaide, Australia
[3] University of North Carolina at Chapel Hill, USA

Abstract. Recently, neuroimaging data have been increasingly used to study the causal relationship among brain regions for the understanding and diagnosis of brain diseases. Recent work on sparse Gaussian Bayesian network (SGBN) has shown it as an efficient tool to learn large scale directional brain networks from neuroimaging data. In this paper, we propose a learning approach to constructing SGBNs that are both representative and discriminative for groups in comparison. A max-margin criterion built directly upon the SGBN models is proposed to effectively optimize the classification performance of the SGBNs. The proposed method shows significant improvements over the state-of-the-art works in the discriminative power of SGBNs.

1 Introduction

Neuroimaging techniques have been widely adopted in brain research for analyzing mental diseases, such as the Alzheimer's disease (AD). They could provide more sensitive and consistent assessments for the early diagnosis of disease. Recently, neuroimage analysis is shifting its emphasis from local brain regions to regional interactions (known as brain network) using graph theory [1]. Such analysis is important because brain network change is often a response to damages like mental diseases. Generally a brain network is constructed as follows (Fig. 1). After aligning to a common stereotaxic space, brain images are partitioned into regions of interest (ROI). A brain network is then modeled by a graph with each node corresponding to a brain region and each edge corresponding to the connectivity between regions. Brain "effective connectivity" analysis focuses on the causal relationships between brain regions [1]. The *directionality* is often of interest, because it may disclose the pathways of how one brain region affects the other. Evidence of causal relationship changes has been found in many mental diseases including AD from multiple imaging modalities [2,3], shedding light on discovering novel connectivity-based biomarkers for disease diagnosis.

Early research works in this regard usually require a prior model of connectivity and study only a small number (≤ 10) of brain regions, such as Structural Equation Modeling [4] and Dynamic Causal Modeling [5]. This situation has been improved recently by [2], where a completely data-driven method, denoted as

P. Golland et al. (Eds.): MICCAI 2014, Part III, LNCS 8675, pp. 321–328, 2014.

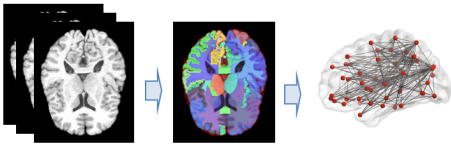


Fig. 1. Illustration of brain network construction from neuroimaging data

H-SGBN in this paper, is proposed to recover sparse Gaussian Bayesian network (SGBN) from more than 40 brain regions in fluorodeoxyglucose PET (FDG-PET) images. It employs the strategy of sparsity constraint to handle large scale Bayesian Network (BN) construction, and circumvents the traditional two-stage procedure for parent set identification in many sparse BN learning methods, achieving a more accurate network recovery [2].

As most BN methods in the literature, H-SGBN is a generative method, which, as pointed out in [6], may ignore the subtle but critical brain structural changes induced by mental diseases. Therefore, a learning approach is proposed in [6], denoted as DL-SGBN, to introduce class discrimination into the SGBN models. DL-SGBN employs Fisher kernel to extract sample-based features from SGBNs, and minimizes a generalization error bound for SVM classifiers with these SGBN-induced features. In that work, the class discrimination is learned by optimizing the classification performance of SVMs, which does not guarantee the equivalent improvement on SGBNs. However, SGBN models are the ultimate goal in such research since they represent the brain connectivity.

In this paper, we propose a new method to learn discriminative SGBN models from neuroimaging data, which overcomes the drawbacks of the state-of-the-art works mentioned above. We propose a max-margin framework to jointly learn two SGBNs, one for each class, for both discrimination and representation. Unlike DL-SGBN in [6], our framework optimizes a criterion directly built upon the classification performance of SGBNs, thus further improves the discriminative power of the models from DL-SGBN (and H-SGBN). Our method is different from the literature of BN classifiers where a *single* BN is learned to represent the *differences* of two classes (in either structure or parameter but not in both) [7,8]. These methods work on discrete variables, while the brain ROI measurements are usually continuous variables whose discretization is often hard to decide. Our experiment shows significant improvement of our proposed method over the state-of-the-art works of H-SGBN and DL-SGBN in terms of the discriminative power of SGBNs. The notations of symbols frequently appearing in this paper are summarized in Table 1.

2 Background

Because this paper is based on sparse Gaussian Bayesian Network (SGBN) model, in the following, we review the fundamentals of SGBN in the original paper [2]. For DL-SGBN, the discriminative learning of SGBN, please refer to [6] for the technical details. We compare with both methods experimentally.

Table 1. Notation

x_i	a random variable
$\mathbf{x}$	a sample of m variables: $\mathbf{x} = [x_1, x_2, \cdots, x_m]^\top$
$\mathbf{X}$	the data matrix of n samples, $\mathbf{X} \in \mathbb{R}^{n \times m}$
$\mathbf{x}_{i,:}$	the i-th row of $\mathbf{X}$, representing a sample
$\mathbf{x}_{:,i}$	the i-th column of $\mathbf{X}$, representing the realization of the random variable x_i on n samples
$\mathbf{W}$	the parameters of a Gaussian Bayesian Network: $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_m]$, $\mathbf{W} \in \mathbb{R}^{m \times m}$
$\mathbf{Pa}_i$	a vector containing the parents of x_i
$\mathbf{PA}_i$	a matrix whose j-th column represents a realization of $\mathbf{Pa}_i$ on the j-th sample.
$\mathbf{G}$	an $m \times m$ matrix for BN: if there is a directed *edge* from x_i to x_j, $\mathbf{G}_{ij} = 1$, otherwise $\mathbf{G}_{ij} = 0$
$\mathbf{P}$	an $m \times m$ matrix for BN: if there is a directed *path* from x_i to x_j, $\mathbf{P}_{ij} = 1$, otherwise $\mathbf{P}_{ij} = 0$

A graph of BN $\mathcal{G}$ expresses the factorization property of a joint distribution $p(\mathbf{x}) = \prod_{i=1,\cdots,m} p(x_i|\mathbf{Pa}_i)$. The conditional probability $p(x_i|\mathbf{Pa}_i)$ is assumed to follow a Gaussian distribution in Gaussian BN (GBN). Each node x_i is regressed over its parent nodes $\mathbf{Pa}_i$: $x_i = \mathbf{w}_i^\top \mathbf{Pa}_i + \varepsilon_i$, where the vector $\mathbf{w}_i$ is the regression coefficients, and $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. A BN is a directed acyclic graph (DAG), i.e., there is no closed path within the graph. Identifying parent sets is critical for BN learning. Traditional methods often consist of two stages: determine candidate parent sets and further prune them by some criteria. A drawback rises that a missing true parent in the first stage will never be recovered. The work in [2] proposed a different approach (H-SGBN) based on sparse GBN (SGBN). In H-SGBN, each node x_i is regressed over all the other nodes, and its parent set is implicitly selected by the regression coefficients $\mathbf{w}_i$ that are estimated by:

$$\min_{\mathbf{W}} \sum_{i=1}^{m} \|\mathbf{x}_{:,i} - \mathbf{PA}_i^\top \mathbf{w}_i\|_2^2 + \lambda_1 \|\mathbf{w}_i\|_1 \tag{1}$$
$$s.t. \quad \mathbf{W}_{ji} \times \mathbf{P}_{ij} = 0, \forall i, j = 1, \cdots, m, \quad i \neq j.$$

All the symbols are defined as in Table 1. A challenge for BN learning is how to enforce the DAG property, i.e., avoiding directed cycles in the graph. A sufficient and necessary condition for being a DAG is proposed in [2], which requires $\mathbf{W}_{ji} \times \mathbf{P}_{ij} = 0$ for all i and j. Note that $\mathbf{P}_{ij}$ is an implicit function of $\mathbf{W}_{ji}$. H-SGBN has been shown to outperform the conventional two-stage methods with higher accuracy for the network edge recovery in [2].

3 Our Proposed Method

As a generative model, BN models the density of the data, revealing how the data could be generated through an underlying process. This is desirable in

the exploratory research of brain, where discovering new knowledges about the brain and the mental diseases is critical. When used for classification, a BN is trained for each class independently and a sample is categorized to the class that produces the higher probability. However, the BNs individually trained by each class may ignore some subtle but critical network differences that distinguish two classes. Since we usually have access to both classes in comparison (e.g., AD and normal control), it is argued in [6] that the parameters of the two SGBNs, one for each class, should be learned from the two classes jointly in order to retain the essential discrimination. Therefore, a joint learning method DL-SGBN is proposed in [6], which introduces group discrimination into SGBNs by optimizing the performance of SVM classifiers with SGBN-induced features. Although this leads to a relatively simple optimization problem, optimizing the performance of SVMs does not necessarily equal to optimizing the discrimination of SGBNs that represent the brain networks. We believe that, the discrimination of SGBNs can be further improved if we *directly* optimize their (instead of SVMs') classification performance. Therefore we propose a new learning framework based on max-margin formulation directly built on SGBNs. We call our method MM-SGBN.

For binary classification, maximizing the minimum margin between two classes can be obtained by maximizing the minimum conditional likelihood ratio (MCLR):

$$\mathrm{MCLR}(\mathbf{W}) = \min_{i=1}^{n} \frac{P(y_i|\mathbf{x}_i, \mathbf{W}_{y_i})}{P(\bar{y}_i|\mathbf{x}_i, \mathbf{W}_{\bar{y}_i})},$$

where n is the number of samples. Without loss of generality, y_i and $\bar{y}_i \in \{-1, 1\}$, representing the true and false labels for the i-th sample, respectively. The parameter $\mathbf{W}_{y_i} = \mathbf{W}_1$ if $y_i = 1$, or $\mathbf{W}_{y_i} = \mathbf{W}_2$ if $y_i = -1$. We can see that MCLR identifies the most confusing sample whose probability of the true class assignment is close to or even less than that of the false class assignment. Hence, maximizing MCLR targets the maximal separation of the most confusing samples in the two classes. It is not difficult to see that MCLR can naturally handle multi-class case when replacing the denominator by the maximal probability induced by all false class assignments. Taking log-likelihood of MCLR, we have

$$\log\ \mathrm{MCLR}(\mathbf{W}) = \min_{i=1}^{n} (\log\ p(\mathbf{x}_i|y_i, \mathbf{W}_{y_i}) - \log\ p(\mathbf{x}_i|\bar{y}_i, \mathbf{W}_{\bar{y}_i})) + const, \quad (2)$$

which can be shown as a quadratic function of $\mathbf{W}$ in the case of SGBN. In order to maximize MCLR, we require the log-likelihood difference in Eqn. (2) larger than a margin for all samples and maximize the margin. To deal with hard separations, we employ a soft margin formulation as follows.

$$\min_{\mathbf{W}_1, \mathbf{W}_2, \xi_i, r} \quad \lambda \sum_{i=1}^{n} \xi_i - r \quad (3)$$

$$s.t. \quad y_i \left(\mathcal{L}(\mathbf{W}_1, \mathbf{x}_i) - \mathcal{L}(\mathbf{W}_2, \mathbf{x}_i)\right) \geq r - \xi_i, \quad \forall i \quad (3a)$$

$$\xi_i \geq 0, \quad r \geq 0, \quad (3b)$$

$$f(\mathbf{X}_1, \mathbf{W}_1) \leq T_1, \quad f(\mathbf{X}_2, \mathbf{W}_2) \leq T_2 \quad (3c)$$

$$\mathbf{W}_1 \in DAG, \quad \mathbf{W}_2 \in DAG \quad (3d)$$

Algorithm 1. MM-SGBN: Discriminative Learning

Input: data $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{n\times m}$, label $\mathbf{y} \in \mathbb{R}^{n\times 1}$

1. Obtain the initial solution for Eqn. (3):
   Get initial $\mathbf{W}^{(0)} = [\mathbf{W}_1^{(0)}, \mathbf{W}_2^{(0)}]$ by Eqn. (1);
   Get initial $r^{(0)}$ and $\epsilon_i^{(0)}$ by solving Eqn. (3) with only the two constraints (3a) and (3b) and a fixed $\mathbf{W} = \mathbf{W}^{(0)}$.
2. Select a subsets of parameters $(\mathbf{W}_{i,j})$ that satisfy:
   i) the gradient (change) of SGBN model at these parameters are highly correlated with the class label, and ii) the corresponding edges present in the graph.
3. Optimize the parameters of the selected nodes by Eqn.(3).

Eqn. (3) has three components addressing class separation (3a), model representation (3c) and DAG property (3d), respectively.

The constraints in (3a) enforce the likelihood of $\mathbf{x}_i$ to its true class larger than that to its false class by a margin r. The variable ξ_i is the slack variables indicating the intrusion of the margin. The function $\mathcal{L}(\cdot)$ denotes the log-likelihood:

$$\mathcal{L}(\mathbf{W}, \mathbf{x}) = \sum_{i=1}^{m} \frac{-(x_i - \mathbf{P}\mathbf{a}_i^\top \mathbf{w}_i)^2}{2\sigma_i^2} - \log(2\pi\sqrt{\sigma_i}).$$

The constraints in (3c) control the fitting errors to maintain reasonable representation. Adding these constraints also avoids the scaling problem of $\mathbf{W}$. The function $f(\cdot)$ measures the squared fitting errors of the corresponding SGBNs for the data $\mathbf{X}_1$ and $\mathbf{X}_2$ from the two classes. It is defined as

$$f(\mathbf{X}, \mathbf{W}) = \sum_{i=1}^{m} \|\mathbf{x}_{:,i} - \mathbf{P}\mathbf{A}_i^\top \mathbf{w}_i\|_2^2.$$

The parameters of T_1 and T_2 are application dependent and predefined by users to control how much representation could be sacrificed for discrimination.

The constraints in (3d) are the DAG constraint proposed in Eqn. (1), i.e., $\mathbf{W}_{1\{ji\}} \times \mathbf{P}_{1\{ij\}} = 0,\ \mathbf{W}_{2\{ji\}} \times \mathbf{P}_{2\{ij\}} = 0, \forall i, j = 1, \cdots, m,\quad i \neq j$. By these constraints, we enforce the validity of both graphs.

The optimization in Eqn. (3) is quadratic programming, which can be solved iteratively by fmincon-SQP (sequential quadratic programming) in Matlab. The details are given in Algorithm 1.

Our method differs from the conventional BN classifiers [7,8] that solely focus on classification. In those methods, only one BN is learned to merely represent the "difference" of the two classes. They no longer model the individual class as our method does, and hence are less interpretative. Moreover, they cannot handle the continuous variables of brain imaging measures, and inherit the drawbacks of the traditional two-stage methods. In practice, learning the whole set of SGBN parameters could become unreliable when the training samples are insufficient. Therefore, we follow the line in [6] to optimize only a selected subset of parameters. Note that this does not introduce the same problem as the

traditional two-stage methods. It is just an engineering trick to handle small sample size problem and becomes unnecessary when sufficient training data are available. In contrast, identifying the candidate-parent sets is an indispensable step in two-stage methods to obtain computationally tractable solutions.

4 Experiment

We evaluate our proposed MM-SGBN against the single class method H-SGBN from [2] and the discriminative learning method DL-SGBN from [6]. For comparison, following [6], we apply all methods on the publicly accessible ADNI[1] database to analyze brain effective connectivity for AD. Three data sets are used from two imaging modalities of MRI and FDG-PET downloaded from ADNI.

MRI data set includes 120 T1-weighted MR images belonging to 50 mild cognitive impairment (MCI) patients and 70 normal controls (NC). These images are preprocessed by the typical procedure of intensity correction, skull stripping, and cerebellum removal. They are segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) using the standard FSL[2] package, and parcellate them into Regions of Interest (ROI) based on an ROI atlas after spatial normalization. The GM volumes of each ROI are used as network nodes. Forty ROIs similar to [6] are used[3], mainly in the temporal lobe and around.
PET data set includes 103 FDG-PET images (and their corresponding MR images) of 51 AD patients and 52 NC. The MR images belonging to different subjects are co-registered and partitioned into ROIs as mentioned above. The ROI partitions are copied onto their corresponding PET images by a rigid transformation. The average tracer uptakes within each ROI in PET images are used as network nodes. Forty discriminative ROIs to AD are used.
MRI-II data set is similar to the MRI data set but using 40 different ROIs covering the typical brain regions spread over the frontal, parietal, occipital and temporal lobes.

We test how the learning process improves the discriminative power of the individual SGBNs estimated by each class. The individual SGBNs are obtained by H-SGBN. We test two methods for discriminative learning: our max-margin-based method MM-SGBN and DL-SGBN in [6]. In order to maintain representation capability, 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 process. To classify a test sample, we compare the values of its likelihood and assign the sample to the class with a higher likelihood. The test accuracies are averaged over the 50 randomly partitioned training-test groups and presented in Table. 2. Paired t-tests (two-tailed) are also conducted to examine the statistical significance of the results.

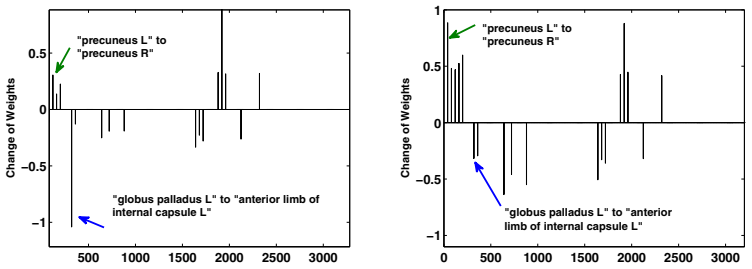
[1] `http://www.adni-info.org/`
[2] `http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/`
[3] Forty ROIs are used to be comparable to that in [2,6].

Table 2. Test Classification Accuracy Averaged over 50 Training-Test Groups (left) and p-values of Paired t-tests (right)

	Accuracy			p-value		
	H-SGBN (%)	DL-SGBN (%)	MM-SGBN (%)	H-SGBN vs. DL-SGBN	H-SGBN vs. MM-SGBN	DL-SGBN vs. MM-SGBN
MRI	66.08	72.92	**76.25**	7e-7	0	1e-4
PET	61.47	66.74	**69.92**	4e-4	0	5e-6
MRI-II	57.08	63.92	**67.17**	7e-6	0	3e-3

From the results we observe that: **i)** Both DL-SGBN and MM-SGBN can greatly improve the discriminative power of the SGBNs estimated from individual classes by H-SGBN. DL-SGBN increases the test accuracy by 6.8% for MRI, 5.3% for PET and 6.8% for MRI-II. MM-SGBN increases the test accuracy by 10.2% for MRI, 8.5% for PET and 10.1% for MRI-II. These improvements are all statistically significant as shown by the very small p-values. This indicates the effectiveness of jointly learning two classes. **ii)** Our proposed MM-SGBN generates the best classification accuracies over all the data sets, which also further improves the classification accuracy of the DL-SGBN by 3.4% for MRI, 3.2% for PET and 3.3% for MRI-II. These improvements are all statistically significant. The advantages of MM-SGBN over DL-SGBN come from directly optimizing the discriminative power of SGBNs, instead of getting indirect help from optimizing the performance of SVM on SGBN-induced features. **iii)** Remind that these improvements on discrimination are achieved with no more than 1% increase of squared fitting errors, as explicitly controlled via the user-defined parameters T_1 and T_2. Note that the rate of 1% is application dependent. More tolerance of fitting errors can potentially bring more discrimination. When we relax fitting error to 10%, another 3% increase of test accuracy could be further achieved.

An example of 18 edge weight changes learned by DL-SGBN and MM-SGBN on PET data is given in Fig. 2, where the SGBN networks from two classes are vectorized and concatenated as x-axis. As shown, both methods learn similar



(a)DL-SGBN (b) MM-SGBN

Fig. 2. Change of edge weights learned by DL-SGBN and MM-SGBN

discriminative patterns despite of using different learning criteria. However, MM-SGBN significantly increases the positive weight of the edge from "precuneus L" to "precuneus R", and reduces the negative weight from "globus palladus L to "anterior limb of internal capsule L". Such differences may lead to the superior performance of MM-SGBN on this dataset and are worthy of further research.

5 Conclusion

In this paper, we propose a max-margin framework directly built on SGBN models to learn causal relationship of brain regions from neuroimaging data. Compared with the state-of-the-art, our method significantly improves the discrimination of the obtained SGBNs, as well as maintaining good representation capacity of the SGBN models.

References

1. Bressler, S., Menon, V.: Large-scale brain networks in cognition: emerging methods and principles. Trends Cogn. Sci. 14(6), 227–290 (2010)
2. Huang, S., Li, J., Ye, J., Fleisher, A., Chen, K., Wu, T., Reiman, E.: A sparse structure learning algorithm for gaussian bayesian network identification from high-dimensional data. IEEE TPAMI 35(6), 1328–1342 (2013)
3. Li, X., Coyle, D., Maguire, L., Watson, D., McGinnity, T.: Gray matter concentration and effective connectivity changes in alzheimer's disease: A longitudinal structural mri study. Neuroradiology 53(10), 733–748 (2011)
4. Kim, J., Zhu, W., Chang, L., Bentler, P., Ernst, T.: Unified structural equation modeling approach for the analysis of multisubject, multivariate functional mri data. Human Brain Mapping 28, 85–93 (2007)
5. Friston, K., Harrison, L., Penney, W.: Dynamic causal modeling. Neuroimage 19, 1273–1302 (2003)
6. Zhou, L., Wang, L., Liu, L., Ogunbona, P., Shen, D.: Discriminative brain effective connectivity analysis for alzheimers disease: A kernel learning approach upon sparse gaussian bayesian network. In: CVPR, pp. 2243–2250 (2013)
7. Pernkopf, F., Bilmes, J.: Efficient heuristics for discriminative structure learning of bayesian network classifiers. J. Mach. Learn. Res. 11, 2323–2360 (2010)
8. Guo, Y., Wilkinson, D., Schuurmans, D.: Maximum margin bayesian networks. In: UAI, pp. 233–242. AUAI Press (2005)