DISCRIMINATION OF ADHD CHILDREN BASED ON DEEP BAYESIAN NETWORK

A. Junyu Hao, B. Lianghua He

Tongji University
Department of Computer Science and Technology
Shanghai, China

ABSTRACT

Attention deficit hyperactivity disorder (ADHD) is a threat for the public health all the time, so the effective discrimination of it is significant and meaningful. In current research, Functional Magnetic Resonance Imaging (fMRI) data has become a popular tool for the analysis of ADHD. In this paper, we introduce the Deep Bayesian Network, a combination of Deep Belief Network and Bayesian Network, to classify the ADHD children from the normal. In Deep Bayesian Network, The Deep Belief Network is applied to normalize and reduce dimension of the fMRI data in every brodmann area. And the Bayesian Network is used to extract the feature of relationships between several well-performed brain areas by structure learning. According to the information of structure and probability in Bayesian Network, we predicted the subjects as control,combined ,inattentive or hyperactive using SVM classifier. The final results perform better than using single Deep Belief Network and the best results in ADHD-200 competition.

Index Terms— ADHD, Deep Learning, Bayesian Network, SVM, Deep Belief Network

1. INTRODUCTION

Attention deficit hyperactivity disorder (ADHD) is among the most common psychiatric disorders of childhood that persists into adulthood in the majority of cases[1]. During this period, symptoms such as inattention, hyperactivity(restlessness in adults), disruptive behaviour, and impulsivity are common. Based on the presenting symptom, ADHD can be divided into three subtypes–predominantly inattentive, predominantly hyperactive impulsive or combined if criteria for both types are met. According to American Psychiatric Associations Diagnostic and Statistical Manual, the prevalence of ADHD in the whole world is approximately 5%, especially in the United States the prevalence among 8- to 15-years-olds reaches to 8.7% during the past year. Therefore, the methods of diagnosing the ADHD is in urgent need.

Recently, Functional Magnetic Resonance Imaging(fMRI) has become very popular for brain activity related studies. Researchers use it for identifying the brain regions which are responsible for particular cognitive activities based on the correlation of input stimulus signal and captured brain fMRI signals. Shengfu Liang et al.[2] utilizes the LDA classifier to discriminate children with ADHD by analyzing the resting-state functional magnetic resonance imaging (fMRI) data, and the average accuracy of distinguishing normal and ADHD children reaches 80.08% though 50 times of 2-fold validation. Xunheng Wang[3] applies Kernel Principal Component Analysis (KPCA) method based on connectivity matrix of each functional meaningful brain regions to find the abnormal pattern of ADHD. On this condition, Support Vector Machine (SVM) assifier make the correct classification rate reach about 81% using a leave-one-out cross validation. On the global ADHD-200 competition, Eloyan A et al.[4] from Johns Hopkins University mainly use rs-fc-fMRI based on decomposition of CUR along with gradient boosting. In the end, they achieved the best score. In this paper, three datasets from ADHD-200 competition are applied to discriminate ADHD with typical controls.

As we all know, dimensionality reduction facilitates the classification and storage of high-dimensional data, especially for fMRI data in ADHD children. Compared with other method of reducing dimensionality, such as principal components analysis (PCA), Deep Belief Network not only perform better than them, but also can normalize the different number of voxels in different brodmann areas to the same size, which makes it possible that the data can be exactly dealt with Bayesian network. In addition, Geoffrey E.Hinton et al.[5] derive a fast, greedy algorithm that can learn deep, directed belief networks one layer at a time, provided the top two layers form an undirected associative memory by using complementary prior. In current, Deep Learning method has been applied for many areas including image processing [5], audio classification [6], natural language processing [7] and so on. All of them achieved good performance. Considering its powerful learning ability and advantage of dimensionality reduction and normalization, Deep Learning method will be used to analysis massive fMRI data to dig out the cognitive significance of brain in this paper.

Thanks to National Natural Sciences Foundation of China.(No.61272267, 61270220, 51075306,61273261)

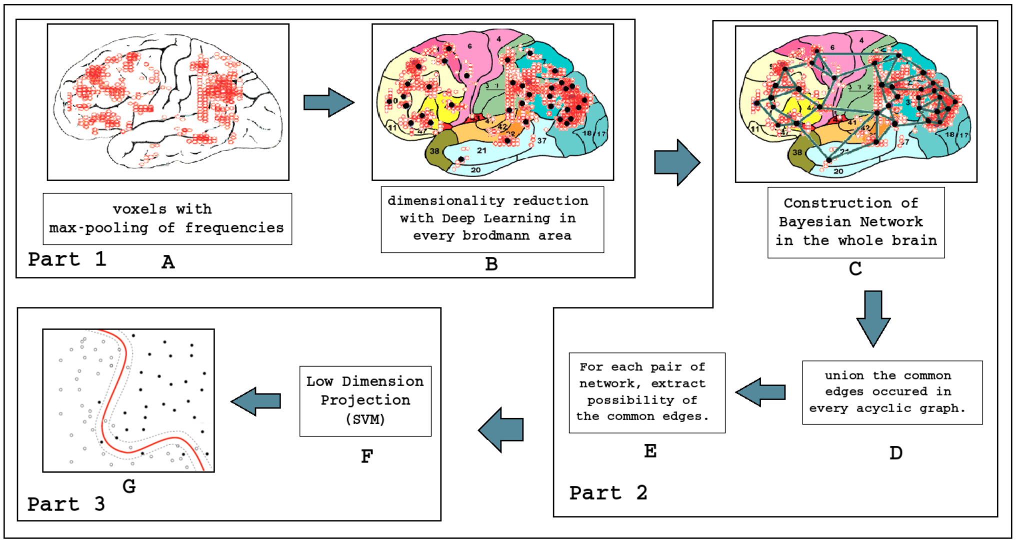

Fig. 1. The procedure of Deep Bayesian Network

To reduce the complexity of computation, the fMRI data, which is a time series of 3D images, is divided into 48 areas according to brodmann template. Due to the cause in ADHD cases is unknown, we should take all the brain areas into consideration rather than analysis the brain area separately, which meets the requirement of Bayesian network. Besides, it is a graphical model that can encodes probabilistic relationships among variables of interest. On one hand, the model can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. For example, in bioinformatics learning Bayesian networks have been used for the interpretation and discovery of gene regulatory pathways [8]. On the other hand, it has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Based on this, Bayesian network has been used in information retrieval[9], natural language processing[10], and for the analysis of a medical services performance for management decisions[11]. In this paper, Bayesian network is applied to retrieval the information between different brain areas, which is shown by the edges in the graph of the Bayesian network.

This paper will introduce the Deep Belief Network as a tool to reduce the dimensionality of data. Before this, the frequency feature on the voxels of the different brain areas(**A** in Fig.1) are used as a vector of feature for the raw input for the Deep Belief Network. Besides, Deep Belief Network can reconstruct the different number of voxels in different areas into the same number of features(**B** in Fig.1). This normalization of fMRI data make it possible that Bayesian network can extract the relationships from different brain areas(**C-E** in Fig.1). Finally, using SVM as a classifier to discriminate the ADHD children from normal(**F-G** in Fig.1). The flow chart of the whole procedure is shown in Figure 1.

Results show that the method proposed in this paper got better performance than the competition and the Deep Belief Network single. Section 2 mainly describes the Deep Belief Network method, Bayesian network and the application of two graph model for ADHD data. Section 3 presents our experiments and results on ADHD dataset.

2. METHOD

The proposed method, Deep Bayesian Network, can be divided into three main parts such as Deep Belief Network for normalization and Dimensionality reduction(Part 1 in Fig.1), Bayesian Network construction for feature extraction(Part 2 in Fig.1), SVM for ADHD subject classification(Part 3 in Fig.1). The following sections describe each of the parts in details.

2.1. Deep Belief Network

Deep Belief Network is composed of a stack of restricted Boltzmann machines (RBM) which is shown in Figure 2. Training the architecture of Deep Belief Network includes two steps, which are pretraining and back propagation. Pretraining each layer of RBM network make sure that the feature can be mapped into different feature space and keep more feature information at the same time. Back propagation spread the error message to every level of RBM from the top to down. This procedure is viewed as a initialization of parameter weights, which overcome the shortcomings of local optimum and long train time due to random initializtion of parameter weights.

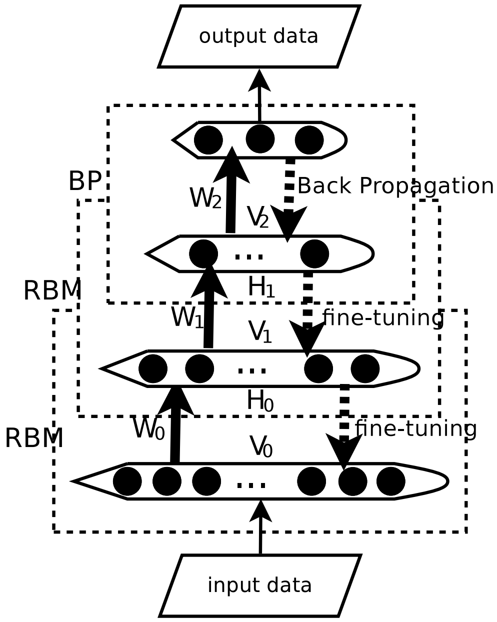

Fig. 2. Architecture of Deep Belief Network

As a component of Deep Belief Network, the restricted Boltzmann machine [12] is a two-layer, undirected, bipartite graphical model where the first layer consists of observed data variables (or visible units), and the second layer consists of latent variables (or hidden units). The visible and hidden layers are fully connected via symmetric undirected weights, and there are no intra-layer connections within either the visible or the hidden layer. A typical RBM model topology is shown in Figure 3.

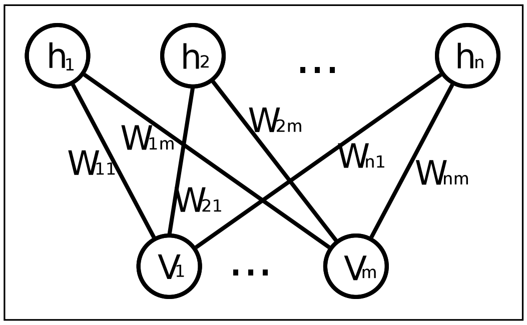

Fig. 3. Struct of RBM

The weights and biases of an RBM determine the energy of a joint configuration of the hidden and visible units $E(v, h)$,

$$E(v,h;\theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} v_i h_j w_{ij} - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j \quad (1)$$

with model parameters $\theta = \{W, b, a\}$ and $v_i, h_j \in \{0, 1\}$. W are the symmetric weight parameters with $V * H$ dimensions, b are the visible unit bias parameters, a are the hidden unit bias parameters.

In general Boltzmann machines, the probability distributions over hidden and visible units are defined in terms of the energy function:

$$p(v,h) = \frac{1}{z}\exp(-E(v,h)) \qquad (2)$$

where z is defined as the sum of $\exp(-E(v, h))$. Based on the joint probability of v and h, the conditional probability of v given h and of h given v is easily obtained. So the individual activation probabilities are given by

$$p(h_j = 1|v) = \alpha(\sum_{i=1}^{m} w_{ij} v_i + a_j) \qquad (3)$$

$$p(v_i = 1|h) = \alpha(\sum_{j=1}^{n} w_{ij} h_j + b_i) \qquad (4)$$

where $\alpha(t) = (1 + \exp(-t))^{-1}$.

A RBM is pre-trained to maximize the log-likelihood $\log P(v)$.Following the gradient of the log likelihood we obtain the update rule for the weights as,

$$\Delta w_{ij} = \epsilon(< v_i, h_j >_{data} - < v_i, h_j >_{recon}) \qquad (5)$$

Where ϵ is the learning rate and the angle brackets manifests the expectations relative to the distribution specified in the subscript. The updating rule makes it the reconstruction of hidden units equals to the data. Only by this way, the hidden unit is approximate to the visible units. Then they can be seen as the exact expression of the data.

2.2. Bayesian Network

Bayesian Network includes a directed acyclic graph (DAG) and a condition probability table related to every node (CPT). The fronter is called network structure learning, and the other is called parameters learning. This paper mainly uses the structure to extract the information, so we will pay more attention on structure learning, which can find the best Bayesian network which is most suitable to the brain areas. The directed acyclic graph (DAG) can expresses the factorization property of a joint distribution p(x). With each variable corresponding to a node in $\mathcal{G}$, the joint distribution is factorized as

$$p(x) = \prod_{i=1}^{m} p(x_i|\boldsymbol{Pa}(x_i)) \qquad (6)$$

where $\boldsymbol{Pa}(x_i)$ denotes the parent $i = 1, \cdots, m$ nodes of x_i .

Currently, there are three methods to learn the structure of Bayesian network, which are method based on searching and scoring, method based on testing independence and the mix of two. The third method performs well, especially for Max-Min Hill Climbing(MMHC) algorithm. MMHC algorithm can be divided into two stages. Firstly, probable parent nodes of each node have to be determined, which can determine the initial Bayesian network architecture. Secondly, use search-and-score method to determine the relationship between each node. For the variable x, Max-Min Parents and Children (MMPC) algorithm will determine the candidate set of parent nodes in the first stage, which is call $CPC(x)$, If another variable a meet this condition $a \in CPC(x)$, what we can determine is that there is a undirected edge between x and a, but the direction is uncertain. In this way, MMPC algorithm can determine the preliminary framework of undirected network. In order to determine the exact relationship between each node, search-and-score method performs the operations of adding, subtracting and reversing the edge to change the structure of the network, and scores each network to find the best one. The difference between MMHC and Greedy searching algorithm is that searching space is just limited to the case when variable a belongs to $CPC(x)$, MMHC algorithm will add an edge from a to x. By this, MMHC can reduce the searching spaces and improve the learning efficiency of the whole algorithm. In the end, we can get a DAG and the condition probability of each edge, which is shown in Figure

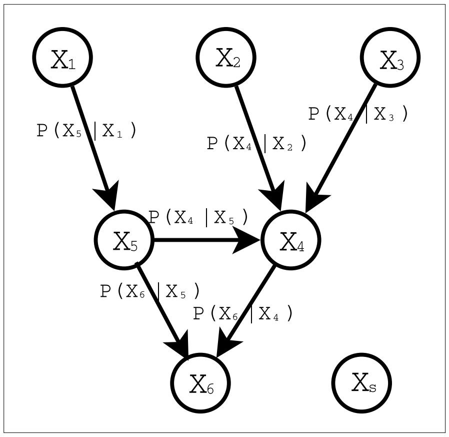

Fig. 4. Struct of Bayesian Network

2.3. The Construction of Deep Bayesian Network

2.3.1. Deep Feature Extraction

To extract and normalize the deep feature, Deep Belief Network is applied to this. But before DBN, some preprocessing are needed for the fMRI data, such as realign, slice time, co-register, normalize, smooth which are described in this paper[13] in detail. First, according to brodmann template, the 3D images of fMRI datas X are divided into 48 areas, where divided data X_d is a 1*48 cell array. In every cell, X_d have the same data structure as the orginal data X. Second, transform the divided data X_d from time domain to frequency domain by the fast Fourier transform algorithm (FFT).

$$X_{fft} = fft(X_d) \tag{7}$$

third, execute max-pooling of frequencies in each voxel to select the frequency which has the maximum value of voxels in some areas may be different during the scanning procedure and the frequency is higher when the voxel is more active. At this time, each voxel has only one property and the number of properties in each brain areas depends on the number of voxels, which is present as X_{max}. That is to say, after preprocessing, the DBN architecture with three hidden layers convert the datas from $X_{max}(48 * n)$ to $X_{dbn}(48 * 50)$, where n is changeable and represents the number of voxels in different brain areas. The pseudo-code for application of Deep Belief Network on ADHD data can be seen as following Algorithm 1 .

Algorithm 1 Deep Feature Extraction

```
Input: $X$
$initial();$                                      ▷ $maxepoch, vhn, tn$
$X_{max} \leftarrow preprocess(X)$                ▷ preprocess data
for $k \leftarrow 1, 3$ do                        ▷ train the weights of rbm
    $[\theta, X_{tmp}] \leftarrow RBN(X_{max})$   ▷ train $\theta = [w, vb, hb]$
    $X_{max} \leftarrow X_{tmp}$                  ▷ update data
end for
$backprop(\theta, X_{max})$                       ▷ back propagation
$X_{dbn} \leftarrow DBN(\theta, X_{max})$         ▷ reduce dimensionality
Output: $X_{dbn}$
```

2.3.2. Structure Learning of Deep Bayesian Network

To extract the feature of relationships between different brain areas, the structure learning of Deep Bayesian network is applied to solve this. But before constructing Deep Bayesian Network, some preprocessing are needed for the data from Deep Bayesian Network.

First, the input data X_{dbn} need to be filtered so that it can be wipe out the noise. Second, the MMHC algorithm can speed up the structure of Bayesian network, but it needs the limited parent nodes of each node. So the dependent nodes K are treated as the limited parent nodes computing by conditional independence testing in filtered data X_{filter}. Third, we union the limited parent nodes and use it K_{un} to learn the structure of all data including training data and testing data. By this, we can get a DAG G and a table P including the information of probability of edge between the two nodes. Forth, the probabilities of each edge in Bayesian network will be exacted out and viewed as feature F of ADHD children and normal children. Finally, SVM will be acted as classifier to train the training sample, which will get the structure parameter of the mode S_{mod}, and classify the testing sample, which will get the predicted label of testing sample L. The accuracy T can be computed by predicted labels L and true labels Y. The pseudo-code for application of BN on ADHD data can be seen as following Algorithm 2.

Algorithm 2 Structure Learning of Deep Bayesian Network

```
Input: $X_{dbn}, Y$
  $X_{filter} \leftarrow preprocess(X_{dbn})$
  $K \leftarrow indep(X_{filter})$    ▷ Conditional independence test
  $K_{un} \leftarrow union(K)$    ▷ union the parent's limitation
  $[G, P] \leftarrow BN(K_{un}, X_{filter})$    ▷ structure learning
  $F \leftarrow extract(G, P)$    ▷ extract feature
  $S_{mod} \leftarrow svmtrain(F, Y)$    ▷ svm training
  $L \leftarrow svmpre(S_{mod}, P)$    ▷ svm predicting
  $T \leftarrow com(Y, L)$    ▷ compute the accuracy
Output: T
```

3. EXPERIMENTS AND RESULTS

3.1. Data

The data used in this paper can be downloaded from the ADHD-200 Global Competition website. DBN and BN model is built upon the ADHD dataset for NYU, Peking-1 and KKI respectively.

For NYU, the training subjects are 216, and testing subjects are 41; for Peking-1 dataset, the training subjects are 85, and testing subjects are 50; and for KKI the training subjects and testing subjects are 83 and 11 respectively. The detail information for the subjecets is shown in Table 3.1.

site	NYU		Peking-1		KKI	
	train 216	test 41	train 85	test 50	train 83	test 11
control	98	7	61	27	61	8
combined	73	22	7	9	16	3
inattentive	2	0	0	1	5	0
hyperactive	43	7	17	13	1	0

Table 1. Demographic Information of three Datasets

3.2. The Experiment of Parameter Designing

To construct a Deep Bayesian Network, we must set up the number of selected brain areas. If the parameter of number is small, the Bayesian network will can't extract the enough information of relationships between brain areas from the structure. If the parameter of number is too large, the probability of that unrelated brain areas will be chose as a node will get higher, which has a side effect on the final classification. In addition, with the increase of the variables, the search space of network structure will presents exponentially. So the experiment of parameter designing is significant.

3.2.1. *Accuracy in Single Brain Area*

As NYU dataset in the ADHD-200 competition achieved the lowest discrimination results, Deep Belief Network algorithm is particularly tested on the NYU dataset of 48 areas and softmax as a classifier generates the accuracies of different brain areas,which is convenient for the choice of node in the Bayesian network. The results of 48 regions are shown in Figure 5.

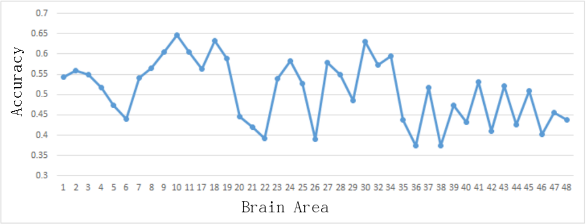

Fig. 5. Accuracy of NYU in 48 Areas

From the figure above, we can see that different brain areas have different performance of discriminate the ADHD children from the normal. The area of 9, 10 , 11, 18, 30 perform well. According to brodmann definition, area 9, 10, 11 stands for prefrontal cortex, area 18 plays a role in the visual cortex and area 30 is part of cingulate cortex.

3.2.2. *Accuracy of Several Brain Area*

As we all know, Bayesian network can extract the feature of relationships between brain areas. Obviously, selecting the well-performed brain areas first in current is a good strategy. According to Figure 5, the performance in ares of 10, 18, 30, 9, 11, 19, 34, 32, 8, 17, 2, 28, 3, 1, 7, 23, 25, 41 decreases step by step. The result of different number of areas by BN is shown in Figure 6.

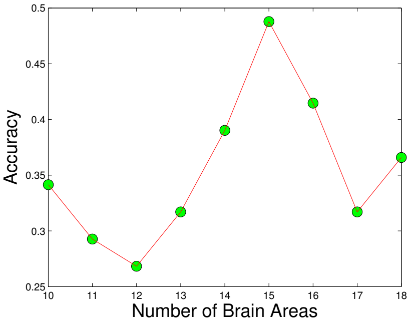

Fig. 6. Accuracy in Different Number of Brain Areas

From the chart we can see that the accuracy of discriminate fluctuate a little from 10 to 18, but it reaches the peak when the number of brain areas is equal 15. So we will choose

15 brain areas in the front to construct the Bayesian network. Besides, using the relationships between brain areas to discriminate ADHD is better than using information in single brain area and the best result in ADHD-200 competition.

3.3. Performances on NYU, Peking-1 and KKI dataset

The experiments are also excuted on the NYU, Peking-1 and KKI dataset. The results released by ADHD-200 competition are 35.19% for NYU, 51.05% for Peking-1 and 61.90% for KKI respectively. The prediction accuracies of Attributed graph distance[14] (AGD) are none, 58.82%,54.55%. The Deep Bayesian Network gain a higher prediction accuracies than Deep Belief Network single, which are 48.78% for NYU, 54.00% for Peking-1 and 72.72% for KKI. The details show in Figure 7.

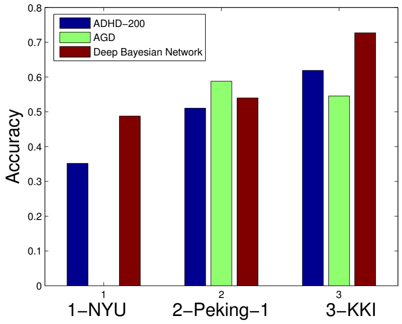

Fig. 7. Performance in Different Datasets

From this chart, we can see that Deep Bayesian Network improves the prediction accuracies in these three datasets compared with the results of ADHD-200 competition. Besides, the increase of accuracy is the highest in the NYU dataset than other two dataset. Table 3.1 shows us that the number of training samples in NYU dataset is far more than Peking-1 and KKI. So Bayesian network can extract more useful feature of relationships between brain areas. The different number of training samples can have a big effects on the accuracy. This effect is shown in Figure 8.

Considering that discriminating the ADHD is important and meaningful, here we take the prediction accuracy of the method along with the specificity and sensitivity values into consideration. The detail information is show in Tab 3.3.

Dataset	Accuracy	Specificity	Sensitivity
NYU	64.7	68.8	43.9
Peking-1	66.3	87.7	22.9
KKI	59.0	83.0	55.6

Table 2. The Detail information in Different Datasets

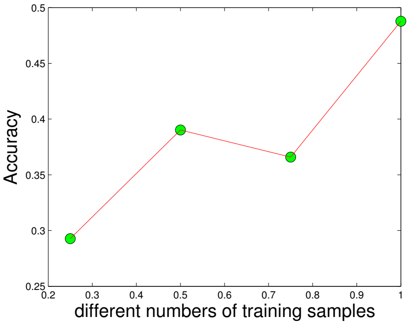

Fig. 8. Performance in Different Number of training samples

4. CONCLUSION

In this paper, Deep Bayesian Network, which is a mix of Deep Belief Network and Bayesian Network, is introduced to discriminate the ADHD children from the normal. Deep Belief Network as a one of the Deep Learning models can be better applied for reducing dimensionality and normalizing the fMRI data. In addition, Bayesian Network is used to extract the feature of relationships between the selected brain areas. During the procedure, the number of perform-well brain areas is discussed in order to find the best value of this parameter.

In the future, we could find more effective method to choose the different brain areas as the input of Bayesian Network. In addition, the distance between different brain areas is different, so they have different influence over each other. we could take the area's neighborhood into consideration manually to strengthen the structure learning.

5. REFERENCES

[1] Sandra JJ Kooij, Susanne Bejerot, Andrew Blackwell, Herve Caci, Miquel Casas-Brugué, Pieter J Carpentier, Dan Edvinsson, John Fayyad, Karin Foeken, Michael Fitzgerald, et al., "European consensus statement on diagnosis and treatment of adult adhd: The european network adult adhd," *BMC psychiatry*, vol. 10, no. 1, pp. 67, 2010.

[2] Sheng-Fu Liang, Tsung-Hao Hsieh, Pin-Tzu Chen, Ming-Long Wu, Chun-Chia Kung, Chun-Yu Lin, and Fu-Zen Shaw, "Differentiation between resting-state fmri data from adhd and normal subjects: Based on functional connectivity and machine learning," in *Fuzzy Theory and it's Applications (iFUZZY), 2012 International Conference on*. IEEE, 2012, pp. 294–298.

[3] Xunheng Wang, Yun Jiao, and Zuhong Lu, "Discriminative analysis of resting-state brain functional connectivity patterns of attention-deficit hyperactivity disorder using kernel principal component analysis," in *Fuzzy Sys-*

tems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on, vol. 3.

[4] Ani Eloyan, John Muschelli, Mary Beth Nebel, Han Liu, Fang Han, Tuo Zhao, Anita D Barber, Suresh Joel, James J Pekar, Stewart H Mostofsky, et al., "Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging," *Frontiers in systems neuroscience*, vol. 6, 2012.

[5] Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[6] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in neural information processing systems*, 2009, pp. 1096–1104.

[7] Ruhi Sarikaya, Geoffrey E Hinton, and Anoop Deoras, "Application of deep belief networks for natural language understanding.," *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 22, no. 4, pp. 778–784, 2014.

[8] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er, "Using bayesian networks to analyze expression data," *Journal of computational biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[9] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al., *Modern information retrieval*, vol. 463, ACM press New York, 1999.

[10] Wendy Webber Chapman, Marcelo Fizman, Brian E Chapman, and Peter J Haug, "A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia," *Journal of biomedical informatics*, vol. 34, no. 1, pp. 4–14, 2001.

[11] Silvia Acid, Luis M de Campos, Juan M Fernández-Luna, Susana Rodrıguez, José Marı?a Rodrı?guez, and José Luis Salcedo, "A comparison of learning algorithms for bayesian networks: a case study based on data from an emergency medical service," *Artificial intelligence in medicine*, vol. 30, no. 3, pp. 215–232, 2004.

[12] Geoffrey E Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[13] Scott A Huettel, Allen W Song, and Gregory McCarthy, *Functional magnetic resonance imaging*, vol. 1, Sinauer Associates Sunderland, MA, 2004.

[14] Shah M Dey S, Rao A R, *Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects*, vol. 8: 64., Frontiers in Neural Circuits, 2014.