뉴럴 네트워크 모델 구조의 점진적 학습을 위한 랜덤 그래프 방법

저자1 김지섭°, 저자2 장병탁
서울대학교 컴퓨터공학부
{jkim, btzhang}@bi.snu.ac.kr

Random Graph Methods for Incremental Structure Learning of Neural Network Models

Jiseob Kim, Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

Non-parametric Bayesian structure learning approaches such as Indian Buffet Process (IBP) have been studied to resolve the design problem of probabilistic graphical models. However, such methods does not fully reflect the common belief that the correlation between features gets stronger as their geometric distance gets closer. To tackle this problem, we adopt several random graph models such as Barabási-Albert Network model or Erdős-Rényi model as a structure learning algorithm. We compare the properties of these models and test if any can compensate the weakness of the previous models. We experiment our model with MNIST and CIFAR-10 dataset. From the experiment, we conclude Watts-Strogatz is the most efficient model.

1 Introduction

Neural network models have rich representation power and capability of learning complex patterns with a single network. However, due to the structural complexity of the connectionist models, there are too many design parameter options, such as number of layers or average number of connections per node. The selections of these parameters are in many cases very crucial to the performance of the model.

To tackle this problem, non-parametric Bayesian structure-learning approaches such as Indian Buffet Process (IBP) [1] are suggested [2]. In these works, number of edges is automatically decided by posterior inference. However, there are still too many structural candidates in that the rules for connecting nodes are absent. One breakthrough for this problem is to reduce the parameter space by adopting the findings from the network theory.

Recent studies in network theory show that the most of complex networks, regardless of their domain, have three common properties – clustering, scale-freeness and short mean path. Clustering is the property that the network to have small clique size. This is thought to be due to the fact that it is mostly hard to make a connection to the nodes that are far from the current one because of the geometric constraints. Scale-freeness is often recognized as the power law distribution of the degree of nodes (number of edges connecting to the node). In particular, there are few hub nodes that have many connections than other nodes. Finally, the short mean path is the property that the mean number of edges from one node to anoather is short. This essentially means that the communication between nodes is efficient and fast despite sparse connections.

Among these three properties, our interest is particularly the clustering property because it is familiar and intuitive for machine learning problems. For example, in image recognition, edge is the pattern that connects the pixels nearby. And the higher level shapes such as circle, rectangle, or even face are also a local combinations of the edges. Making use of the clustering property is not a novel idea and some of machine learning algorithms already adopts it. Gaussian process regression [3] has assumption in the base that the closer points would have relatively smaller covariance. Another successful usage in connectionist model is Convolutional Deep Belief Network (CDBN) [4]. In comparison with classic Deep Belief Network [5], it is only trained on local features, which gives benefits to the computation time. Even so, CDBN still is a fixed structure model and one should determine number of nodes and connections.

On this account, we try to discover new approach to the structure learning of the neural network models, by investigating three different random graph models – ER, BA, WS. This random graph based approach could achieve efficiency (minimal number of edges) and adaptiveness (number of edges changes as the data comes in) at the same time. However, in this paper, the scope is confined to evaluate the mean of mutual information for each of three random graph models, as a first step of the investigation. We mainly discuss the models in clustering point of view but not limited to it.

In the following sections, first we review the random graph models and their construction algorithms. After, we present the experiments with MNIST and CIFAR-10 dataset.

2 Random Graph Models

2.1 Erdős-Rényi Model

Erdős-Rényi (ER) Model [6] is the very first model developed to explain the small world properties of the networks. The model simply defines the i.i.d. and constant probability for connecting the edge between two nodes i, j as follows.

$$P_{ij} \overset{i.i.d}{=} p \tag{1}$$

p in above equation is arbitrary value in [0,1]. The constructed graph connects any two nodes with substantially short path. However, the ER model contains seirous defect that it does not make clustered structures, which is undesirable.

2.2 Barabási-Albert Model

The distinctive property of Barabási-Albert (BA) Model [7] is scale-freeness. This is due to the preferential attachment aspect of the construction algorithm. The algorithm incrementally connects new node to the existing nodes but the nodes that have many connections are more probable to make edge with the new node. In particular, the probability of new node i connects to one of exsisting node j is

$$P_{ij} = \frac{k_j}{\sum_n k_n} \tag{2}$$

where n is an index for existing nodes.

2.3 Watts-Strogatz Model

The construction algorithm for Watts-Strogatz (WS) Model [8] starts with clustered graph; each of the nodes is connected to the k nearest neighbors. It then iteratively visits the nodes and randomly rewire the edge with probability β. The WS model shows strong clustering property but does not exhibit the scale-freeness.

3 Experiments

We analyze the mutual informations between dimensions that are selected pairs by edges of each model. If the model is efficient despite the sparse connections, the connecting edges should be informative and therefore, the mutual information presented by the edge should be high. To compare this efficiency between models, we calculate Mean Mutual Information (MMI) of the edges in the network. Additionally, to exhibit the comparative performance to the fully connected network, we calculate the Total Mutual Information (TMI), which is the sum of all mutual information of the edges in the network.

For the data, we use MNIST and CIFAR-10 dataset in the experiments. The former consists of handwritten digit data (28x28 pixels, 10 classes) with 60000 instances and the latter consists of natural images of the objects (32x32 pixels, 10 classes) with 50000 instances. Note that for CIFAR-10 dataset, we use resized (16x16 pixels) and grayscaled images for convenience.

3.1 Mean Mutual Information and Number of Edges

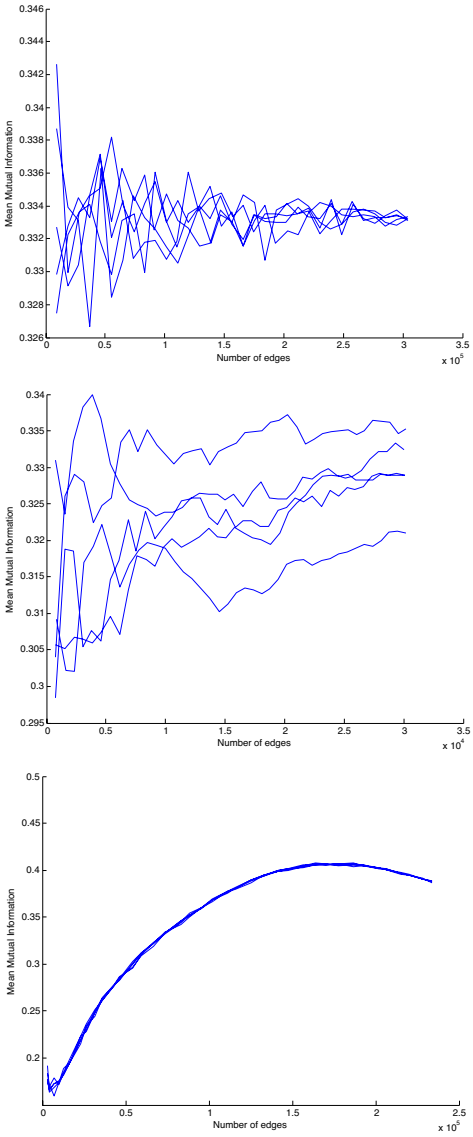

그림 1: Relation between MMI and number of edges. The case of MNIST dataset is shown. 5 trials are presented. (Top: ER, Middle: BA, Bottom: WS)

3.2 Total Mutual Information and Number of Edges

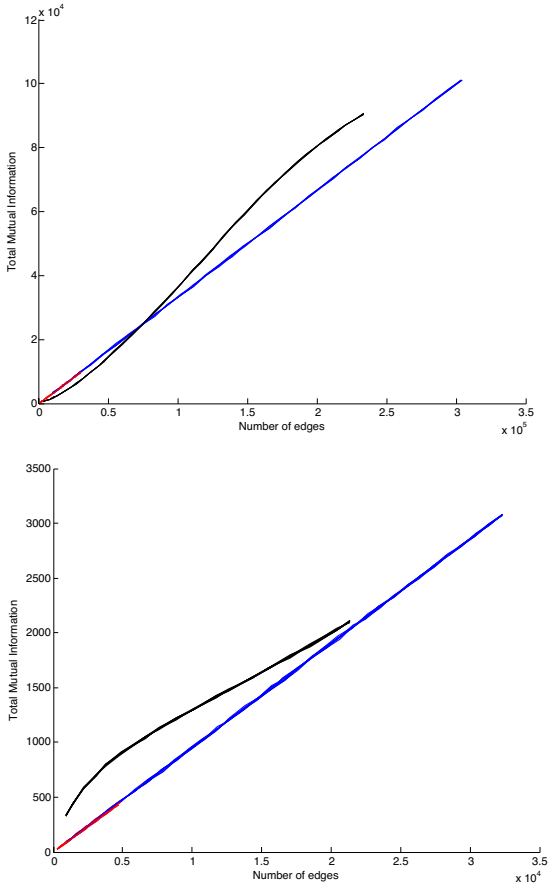

그림 2: Relation between TMI and number of edges (Top: MNIST, Bottom: CIFAR-10). The colors represent each different model (Blue: ER, Red: BA, Black: WS). 10 trials are shown.

4 Discussion

Figure 1 presents MMI according to number of edges of each model. One can notice that in ER model and BA model, the MMI values are retained relatively constant as the number of edges increases (although, the fluctuations decrease). This is due to the scale-freeness of two models. On the other hand, in WS model, MMI value changes as the number of edges increases. That is, in WS model, the efficiency of each edge changes. This phenomenon affects the graph of TMI also. In Figure 2, we can see that the relation between TMI and number of edges in ER and BA models is presented with clean lines. On the other hand, WS model shows curve passes above the lines of the other two models. This leads up to an important conclusion that in the situation of sparsely connected network (small number of edges), WS model efficiently contains the mutual information between dimensions.

5 Conclusion

We have compared the structural efficiency of three random graph models. Mean and total mutual information values for each model are presented with simulation experiments. Among others, Watts-Strogatz model is shown to be the most efficient model in that it contains more mutual information in the same restricted number of edges. Based on our analysis, structure learning algorithms for neural network models that use random graph priors models can be developed in future works.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF-2010-0017734-Videome)

참고 문헌

[1] T. L. Griffiths and Z. Ghahramani, "The indian buffet process: An introduction and review.," *Journal of Machine Learning Research*, vol. 12, pp. 1185–1224, 2011.

[2] R. P. Adams, H. M. Wallach, and Z. Ghahramani, "Learning the structure of deep sparse graphical models," *arXiv preprint arXiv:1001.0160*, 2009.

[3] M. Ebden, "Gaussian processes for regression: A quick introduction," *Available on Internet" http://www. robots. ox. ac. uk/mebden/reports/GPtutorial. pdf*, 2008.

[4] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, (New York, NY, USA), pp. 609–616, ACM, 2009.

[5] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[6] P. Erdos and A. Rényi, "{On the evolution of random graphs}," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, pp. 17–61, 1960.

[7] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, pp. 47–97, Jan. 2002.

[8] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.