International Journal of Performability Engineering, Vol.13, No. 2, March 2017, pp. 231-239

Hybrid Model Based Sampling Algorithm to Infer Dynamic Complex Network

JIN GUO[1, 2], SHENGBING ZHANG[1] AND ZHENG QIU[3]

[1]*School of Computing, Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China*
[2]*School of Electronics and Information Engineering, Xi'an Technological University, Xi'an 710032, Shaanxi, China*
[3]*Aeronautical Computing Technique Research Institute, 15th Lab, Xi'an 710000, Shaanxi, China*

(Received on September 03, 2016, revised on October 16, 2016)

Abstract: Inferring dynamic complex network through a small set of samples is a challenging problem in the field of biological network, social network and transportation network, which can help improve understanding of complex network systems. In this letter, a new Hybrid Model based Latent Variables Sampling algorithm is presented to address the problems of high computation complexity and low accuracy faced by traditional approaches. Experimental results on simulated and real data sets show that the presented method possesses better reasoning performance and significantly improves the precision and efficiency of network inference especially when compared with the other three approaches. Under different dimensions, HM-LVS still has higher accuracy (average 80%) and can effectively reverse engineering dynamic complex networks from time series data.

***Keywords*:** *Artificial intelligence, Learning algorithm, Nnetwork inference, Hybrid model, Markov processes*.

1. INTRODUCTION

Inferring dynamic complex network through a small set of samples is a challenging problem in the field of biological network, social network and transportation network, which can help improve understanding of complex network systems. Different from the structure of static network, the structure of dynamic network including gene regulatory network, cell signal conduction network and aviation network changes with time.

Short-term time series data and gene microarray expression data are typically small sets of samples, which are usually used to model the dynamic network [1]. In these models, directed edges of the network represent probability or causal relationship while nodes represent variables. Classic approaches for inferring dynamic network include differential formula, regression, state space model, information theory and so on [2]. Using a differential formula to establish the model of dynamic network system is a popular method. However, estimating parameters through experimental data usually leads to several parameter sets and uncertainty of parameters, which further results in the indetermination of network structure [3]. Moreover, solving the model established by differential formula method requires optimization methods, which is not feasible considering their low robustness. The state space model can lead to loss of information. Thus, the problem cannot be solved easily [4]. Although regression is applicable to a small

*Corresponding author's email: guojin1019@163.com

set of samples and high dimension problems, it contains the assumption of linearity, which is not suitable for modelling the dynamic network. Moreover, regression suffers from over-fitting and high computation cost as the scale of network increases.

In addition, a framework for inferring the complex network based on dynamic Bayesian network is proposed in [5,6]. Approaches under this framework are mainly based on the assumption of homogeneous Markov chain, which is not consistent with most real dynamic process and cannot tackle piece-wise heterogeneous time series.

This paper develops a new Hybrid Model based Latent Variables Sampling algorithm (HM-LVS) to infer dynamic complex network, in which Hidden Markov model and Bayesian network are combined to model dynamic process. Markov Chain Monte Carlo (MCMC) is also applied to sample positions and number of change-points, as well as parameters of network structure. HM-LVS transforms the problem of dynamic network inference to a static network estimation problem in different phases. The algorithm assumes the network structure remains the same in a certain phase and changes smoothly in adjacent phases. Bayesian posterior inference is utilized in the algorithm to search the optimal network structure in all phases.

2. Mathematical Model

This paper adopts Hidden Markov Model (HMM) to modelling successive structure changes and state transitions of the dynamic network. The network in each phase is establishes as homogeneous Bayesian networks (BN), in which edges reflect the probability relationships between different nodes.

2.1. Hidden Markov model

This paper adopts the Hidden Markov model to describe successive network structural changes, state transitions of each phase. Each node may have a different edge connection relation in different phases. The edge connection changes with the switch of adjacent phases and remains the same in a phase. Therefore, the set of hidden states can be defined as $Q=\{q_1,\ldots,\ q_K\}$, where K is the number of phases.

As for change-points, the starting position of a phase, discrete hidden variable ξ is introduced, the dimension of which equals K. Thus, we have $\xi_i=(\xi_1,\ldots,\ \xi_K)$, $1\le i\le p$ for each node and $\xi_1=1$. During the phase $h(1\le h\le K)$, $\forall t\in[\xi_h,\ldots,\xi_{h+1})$, $\mathbf{G}(t)=\mathbf{G}_h$. The hidden variable ξ and K determine positions and numbers of phases respectively, which can be found easily using posteriori probability sampling. Detailed discussion will be presented later.

As for the observation sequence $\boldsymbol{X}_i=\{\boldsymbol{X}_i(1),\ldots,\boldsymbol{X}_i(n)\}$ of node i, there exists the corresponding hidden state sequence $q^i=\{q_1^i,\ldots$ and directed acyclic sub-graph $P(\mathbb{G}^i,X|bnHMM_i)=P(\mathbb{G}^i|bnHMM_i)\cdot P(X|\mathbb{G}^i,bnHMM_i)$, $\mathbb{G}^i=\{\mathbb{G}_1^i,\ldots$.

According to the hybrid model bnHMM established in this paper, the sub-network output probability of node i can be denoted as

$$= \pi_{q_1}^{i}\left(\prod_{h=2}^{K} a_{q_{(h-1)}q_h}^{i}\right)\left(\prod_{h=1}^{K} P\left(X\middle|\mathbb{G}_{q_h}^{i}\right)\right) \tag{1}$$

where bnHMM is the hybrid model established in this paper, X is the observation sequence, $\mathbf{G}^i$ is the directed acyclic sub-graph of node i in related time interval $1,\cdots$, $\pi_{q_1}^{i} = P\left(\mathbf{G}_{q_h}^{i}\right)$ is the initially sub-graph state probability of node i. $A^i = \left\{a_{q_{h-1},q_h}^{i}\right\}$ is the probability matrix of state transition.

Theorem1: hidden variable sampling theorem. Based on the state transition probability of hybrid model bnHMM and formula (2-3), using ξ and K, sampling hidden variables of time series data as well as Bayesian network structure **G** in different phases, the joint posterior probability can be decomposed as

$$P(K)P\left(\xi|K\right)P\left(\mathbf{G}_{q_1}\right)\prod_{i=1}^{p}\left(\prod_{h=2}^{K} a_{q_{(h-1)}q_h}^{i}\right)\prod_{h=1}^{K} P\left(X\middle|\mathbf{G}_{q_h}\right) \tag{2}$$

ξ and K are sampling hidden variables of time series data, $P(K)$ is the prior probability of hidden variable K.

Proof:

$$\begin{aligned}
P\left(\mathbf{G},\xi,K,bnHMM\middle|X\right) &\propto P\left(\mathbf{G},\xi,K,X,bnHMM\right)\\
&\propto P(K)P\left(\xi|K\right)P\left(\mathbf{G}|bnHMM\right)\cdot\\
&\quad P\left(X|\mathbf{G},\xi,K,bnHMM\right)\\
&\propto P(K)P\left(\xi|K\right)P\left(\mathbf{G},X|bnHMM\right)\\
&= P(K)P\left(\xi|K\right)\cdot\prod_{i=1}^{p}P\left(\mathbf{G}^i\middle|bnHMM_i\right)\cdot P\left(X\middle|\mathbf{G}^i,bnHMM_i\right)\\
&= P(K)P\left(\xi|K\right)\cdot\mathbf{G}\\
&\quad \prod_{i=1}^{p}\left\{P\left(\mathbf{G}_{q_1}^{i}\right)\prod_{h=2}^{K}a_{q_{(h-1)}q_h}^{i}\prod_{h=1}^{K}P\left(X\middle|\mathbf{G}_{q_h}^{i}\right)\right\}\\
&= P(K)P\left(\xi|K\right)\cdot\\
&\quad P\left(\mathbf{G}_{q_1}\right)\prod_{i=1}^{p}\left(\prod_{h=2}^{K}a_{q_{(h-1)}q_h}^{i}\right)\prod_{h=1}^{K}P\left(X\middle|\mathbf{G}_{q_h}\right)
\end{aligned}$$

Through the above reasoning, we have completed the proof of theorem 1.

3. Hybrid Model Based Latent Variables Sampling Algorithm

3.1. Bayesian posterior inferring

This paper uses hidden variables ξ and K to control dynamic network structural changes. Considering that the hidden variables are very flexible, dynamic properties of the network structure can be incorporated to reflect network structure transition and the

homogeneous feature in each phase. Therefore, incorporating the information of hidden variables into the posterior probability can reflect the dynamic property of the network, which is of great help for further inference.

According to theorem 1 and formula(2), suppose $P(K)$ and $P(\xi|K)$ denote the prior probabilities of change-points number and positions, based on [7] and multivariable Gaussian distribution as well as the Normal Wishart distribution, the likelihood probability distribution of network structure **G** with regard to observation data X in phase h can be depicted as

$$P\left(X\middle|G_{q_h}\right)=(2\pi)^{-\frac{n\cdot p}{2}}\cdot\left\{\frac{v}{v+n}\right\}^{p/2}\cdot\frac{c(p,\alpha)}{c(p,\alpha+n)}\cdot det(T_0)^{\alpha/2}\cdot det(T_D)^{-\frac{\alpha+n}{2}} \tag{3}$$

$$T_D=T_0+\left\|\boldsymbol{X}_j-\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{X}_j\right\|_2^2+\frac{vn}{v+n}\left(\mu_0-\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{X}_j\right)\left(\mu_0-\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{X}_j\right)^T \tag{4}$$

n is the number of observing time interval, p is the node number of network, v and α both are the probability distribution parameters of hybrid model, $c()$ and $det()$ are the computation function using in hybrid model, T_0 and T_D are the precision matrix of corresponding status. μ_0 is probability distribution average value of hybrid model, X_j is the observation data sequence of node j.

Substituting formula (5) into formula (4) and applying MCMC posterior inferring, probability of accepting a new network structure when network structure changes in phase h is

$$\Phi\left(\mathbf{G}_h^*|\mathbf{G}_h\right)=min\left\{1,\frac{P(\boldsymbol{X}|\mathbf{G}_h)P(K^*)P(\xi^*|K)}{P(\boldsymbol{X}|\mathbf{G}_h^*)P(K)P(\xi|K)}\cdot\frac{Q(\xi^*|\xi)}{Q(\xi|\xi^*)}\right\} \tag{5}$$

$\mathbf{G}_h$ is the network structure graph in phase h, ξ is the hidden variable of hybrid model, K is the dimension of ξ, and $Q(\xi|\xi^*)$ is the conditional probability distribution.

When applying the MCMC technique to sample from the posterior probability, three types of moving step are constructed to carry out the searching process in the state space $\{\mathbf{G}, \xi, K\}$. Type M_1 is the operation on a single edge in network **G** when hidden variables ξ and K remain constant, which will change the state of the edge including generation, deletion or inversion. Afterwards, judgement is made to decide whether to accept or reject the suggested network structure $\{\mathbf{G}^*, \xi^*, K^*\}$. The other two moving

types are the operations on ξ or K respectively. M_2 represents moving the position of ξ while keeping K unchanged, that is, increasing or decreasing neighbouring time points for a certain phase. M_3 represents increasing or decreasing the value of K, that is increasing or decreasing a phase and reassign ξ. As a result, a transition from one dynamic network structure to another is achieved through MCMC sampling. The proposed algorithm then searches the structure space step-by-step via acceptance probability until it converges or the maximum number of iteration is reached.

3.2. Prior probability distribution

The Hidden Markov Model in this paper is a first order heterogeneous Markov process with K states. Thus, the probability transition matrix can be denoted as

$$A=\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & \\ 0 & a_{2,2} & \ddots & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & a_{k,k} \end{pmatrix}$$

It is assumed that elements on the primary diagonal of the initial probability transition matrix A are $a_{i,i}^{d}=1-s/\pi_{q_1}^{d}$, where s is a random value satisfying the uniform distribution. The elements on the minor diagonal are assumed to be $a_{i,i+1}^{d}=s/\left(2\pi_{q_1}^{d}\right)$, while all other elements are assumed to be zero.

In order to reduce the searching space and take advantage of the similarity with respect to the functionalities and structures of nodes in the same class, the algorithm in this paper first clusters all the nodes, then constructs a Markov Chain and assigns initial state probability for all clusters. This is beneficial for information sharing between nodes and can avoid assigning initial state distribution for all nodes, which effectively decreases the computation complexity. Let C be the number of clusters, then the prior probability distribution can be denoted as $P\left(\mathbb{G}_{q_1}\right)=\prod_{i=1}^{C}P\left(\mathbb{G}_{q_1}^{i}\right)=\prod_{i=1}^{C}\pi_{q_1}^{i}$, where $\pi_{q_1}\sim Dir\left(\alpha_1,\ldots\right.$ and Dir(.) is the Dirichlet distribution. Supper parameters $\alpha_1,\ldots$ are positive integers. Initial state probability distribution is given as $\pi_{q_1}=\left(\pi_{q_1}^{1},\ldots\right.$, where $\sum_{i=1}^{C}\pi_{q_1}^{i}=1$, $\pi_{q_1}\geq 0$.

In addition, it is assumed that $P(K)$ satisfies truncated Poisson distribution (λ=1 and $2\leq K\leq K_{\max}$) and K_{max}=10. It is also supposed that $P\left(\xi\middle|K\right)$ satisfies Bernoulli distribution ($p=\exp(\lambda)/(1+\exp(\lambda))$, $\lambda=K$).

3.3. Algorithm procedure

The procedure of the proposed HM-LVS algorithm is as follows:

Input: time series data sets $X, v, \alpha, T_0, \pi_{q1}, A$.

Output: adjacent matrix **G** of the network structure in different phases.

Step 1: Initialize parameters.（ I is p order unit matrix） $v = 1, \alpha = p + 2, T_0 = 0.5 \cdot I$, $\mu_0 = (0, \ldots, 0)^T$ 。 The maximum number of iteration is L=1,000,000. Iteration number m is zero (initial value).

Step 2: Cluster nodes by k-means algorithm and assign initial state probability π_{q_1} and state transition probability A for each cluster. Initialize bnHMM^0=$\{\mathbf{G}^0, \xi^0, K^0\}$.

Step 3: Execute one kind of MCMC sampling randomly, calculate $P(K^m)$ and $P(\xi^m | K^m)$.

Meanwhile, using formula (5-6) to calculate the likelihood probability of observation data X for each class in different phases. Calculate the posterior probability according to formula (4) and increase the iteration number m.

Step 4：If the iteration number reaches the maximum iteration number L or all the Markov chains converge, generate dynamic complex network based on the adjacent matrix of the network structure G in different phases and the algorithm ends. Otherwise, calculate the acceptance probability Φ according to formula (7). If Φ is larger than the uniform distributed random number, then accept new bnHMM^m=$\{\mathbf{G}^m, \xi^m, K^m\}$ and go to step 3. Otherwise, go directly step 3.

4. Experiment Results

4.1. Simulation data experiment

In order to guarantee the reliability of the results, SynTReN[8] simulation software is used to generate ten sets of data, each of which includes 30 time points and 3 phases configured at time points 0, 10 and 20 respectively. Noises are set as 0.1, 0.5 and 1 correspondingly. Each network is consisted of 50 nodes. Every set of data is tested 10 times and the inference results are compared with the basic network in each phase, results are presented in Figure 1., Figure 3. and Table 1. Inputs of the algorithm include simulation data and 5 parameters while the output is the dynamic network structure.

Fig. 1. illustrates the Receiver Operating Characteristic (ROC) of HM-LVS, GlobalMIT[6], KELLER[3] and ARTIVA[7] under the above mentioned simulation data sets. Figure 2. shows the Positive Predictive Value (PPV) curves of four algorithms. From Figure 1. and Fig. 2., the proposed algorithm shows obvious superiority over the other three algorithms since it can achieve higher learning precision and more accurate results. Fig. 3. conducts further comparison in inference accuracy of four algorithms from the perspective of Area under curve (AUC). As illustrated in Figure 3., the value of AUC equals 0.5 suggests that the performance is similar to that of stochastic algorithm, <0.7 means the performance is ordinary, >0.8 means the performance is rather good. The mean value of AUC in HM-LVS is 0.81. The AUC of ARTIVA (~0.62) is slightly higher than KELLER (~0.6). The AUC of GlobalMIT is the smallest and is close to 0.5. Therefore, it

can be concluded that the algorithm proposed in this paper has a higher inferring accuracy and has value of application.

Table 1: Comparison of performance on simulated datasets

Algorithm	True positive rate（False Positive Rate）/%		
	1 Phase	2 Phase	3 Phase
HM-LVS	0.798(0.19)	0.814(0.21)	0.802(0.20)
GlobalMIT	0.474(0.23)	0.513(0.24)	0.501(0.25)
KELLER	0.642(0.23)	0.673(0.22)	0.594(0.20)
ARTIVA	0.652(0.21)	0.637(0.23)	0.601(0.26)

Table 2: Comparison of average runtime on simulated datasets

Nodes number	Execution time /Minute			
	HM-LVS	Global MIT	KELLER	ARTIVA
30	2.08	2.95	3.36	3.05
50	5.14	6.91	8.07	7.92
100	10.02	13.64	17.48	15.08
150	17.18	26.55	28.32	22.53
200	26.89	36.07	45.63	33.43

Table 1 displays experiment results of four algorithms in true positive rate and false negative rate under the 10 sets of simulation data mentioned above. Regarding different network structure in 3 phases, Table 1 presents accuracy performance verified with the basic network structure in each phase. We can observe that the performance of HM-LVS achieves higher accuracy than GlobalMIT. Meanwhile, the performance of HM-LVS and GlobalMIT is better than that of ARTIVA and KELLER, whose performance is similar to each other. Table 2 illustrates mean execution time of four algorithms on the simulation data set generated by SynTReN when the sampling number is 30 and nodes number is 30, 50, 100, 150 and 200 respectively. It can be observed from Table 2 that HM-LVS requires a shorter execution time than the other three algorithms.

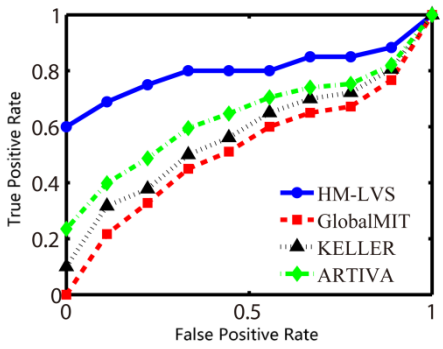

Fig. 1: Comparison of ROC curves

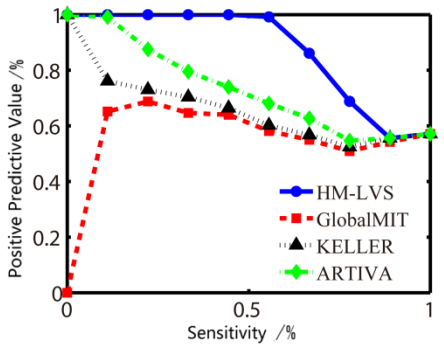

Fig. 2: Comparison of PPV curves

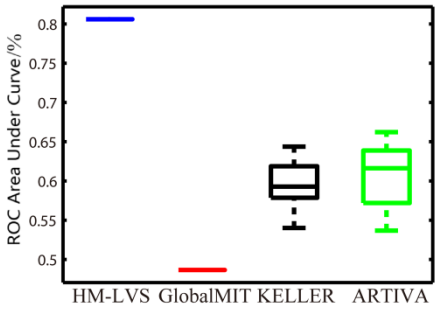

Fig. 3: Comparison of AUC curves

5. Conclusions

Traditional inference methods of dynamic network assume that samples are independent identically distributed. As a consequence, the average estimation of network topology structure is inferred based on all time points, which ignores the dynamic changing property of the network structure. To address the problems of high computation complexity and low accuracy faced by traditional approaches, this paper presents a new Hybrid Model based Latent Variables Sampling algorithm, which uses Markov Chain Monte Carlo to infer the number of phases of different network structures and positions of change-points. In the algorithm, the parameter searching space is reduced and computation complexity is decreased. Simulations and experiments verify both the effectiveness and reasonability of the method. Moreover, the proposed method overcomes drawbacks of traditional algorithms such as discretization of data, prohibition of dynamic structure change and over-fitting, thus providing an effective tool for dynamic network modeling with small set of samples.

References

[1]. Lu T, Liang H, Li H, et al. *High-dimensional odes coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification*. Journal of the American Statistical Association, 2011; 106: 1242-1258.

[2]. Hanshan Li. *Research on target information optics communications transmission characteristic and performance in multi-screens testing system.* Optics communications. 2016; 364: 139-144.

[3]. Zidong W. *An extended kalman filtering approach to modeling nonlinear dynamic gene regulatory networks via short gene expression time series,* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2009; 6(3): 410-419.

[4]. Nguyen X V, Chetty M, Coppel R, et al,Learning globally optimal dynamic bayesian network with the mutual information test criterion, Bioinformatics, 2011; 27(19): 2765-2766.

[5]. Lebre S, Becq J, Devaux F, et al. *Statistical inference of the time-varying structure of gene-regulation networks,* BMC Systems Biology, 2010; 4(1), 130.

[6]. Wu, B., Feng, Y., Zheng, H. *Posterior Belief Clustering Algorithm for Energy-efficient Tracking in Wireless Sensor Networks,* International Journal on Smart Sensing and Intelligent Systems, 2014; 7(3) : 925-941.

[7]. Kawaguchi J, Ninomiya T, Miyazawa Y. *Stochastic approach to robust flight control design using hierarchy-structured dynamic inversion,* Journal of Guidance, Control, and Dynamics, 2011;

34(5) :1573-1577.
[8]. Li Hanshan. *Limited Magnitude Calculation Method and Optics Detection Performance in a Photoelectric Tracking System,* Applied Optics, 2015; 54(7), pp.1612-1617.
[9]. Zhang, J., Yu, J., Chi, N. *Multi-Modulus Blind Equalizations for Coherent Quadrature Duobinary Spectrum Shaped PM-QPSK Digital Signal Processing,* Journal of Lightwave Technology, 2013; 31(7) : 1073-1078.

Jin Guo is a lecturer at Xi'an Technological University, also a PhD Candidate of Northwestern Polytechnical University, China. Her major research interests include network reliability and ddynamic ccomplex network. She is invited as a reviewer by editors of some international journals. She has published many papers in related journals.

Shengbing Zhang is the professor at Northwestern Polytechnical University, China. His major research interests include work reliability and ddynamic ccomplex network. He is invited as a reviewer by the editors of some international journals. He has also published many papers in related journals.