Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 111 (2017) 160–166

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Conference on Advances in Information Technology, IAIT2016, 19-22 December 2016, Macau, China

Estimating bayesian networks parameters using EM and Gibbs sampling

Huimin Chai*, Jiangnan Lei, Min Fang

School of Computer Science and Technology, Xidian University, Xi'an, China

Abstract

A method based on Expectation Maximization (EM) algorithm and Gibbs sampling is proposed to estimate Bayesian networks (BNs) parameters. We employ the Gibbs sampling to approximate the E-step of EM algorithm. According to transition probability, Gibbs sampling is utilized in data completion of E-step, which can reduce the computational complexity of EM algorithm. The experiments for comparison between the proposed method and EM algorithm are made. For the proposed method, the consumed time and the number of iterations are all less than those of EM algorithm. However, the KL divergence is higher than that of EM algorithm, which is a limitation for the proposed method.

Peer-review under responsibility of the organizing committee of the 8th International Conference on Advances in Information Technology

Keywords: BNs parameters estimation; EM algorithm; Gibbs sampling

1. Introduction

Bayesian networks are a way to represent uncertainty that is consistent with the axioms of probability theory[1]. BNs are directed, acyclic graphs that encode the cause effect and conditional independence relationships among variables in the probabilistic reasoning system, where nodes of the graphs are the variables of the domain one wants to model[2]. BNs can model complex systems which have been successfully applied in some domains, such as fault diagnosis[3], image processing[4], speech recognition[5] and situation assessment[6]. However, estimating BNs parameters

* Corresponding author. Tel.: 15802919895.
E-mail address: chaihm@mail.xidian.edu.cn

1877-0509
Peer-review under responsibility of the organizing committee of the 8th International Conference on Advances in Information Technology
10.1016/j.procs.2017.06.023

is one of the most important things for BNs application. In most cases, data set that is composed of data samples for BNs parameters estimation is often incomplete. Therefore, a method is needed to resolve the BNs parameters estimation with incomplete data.

EM algorithm which is proposed by Dempster et al. in 1977 provides an iterative procedure for maximum posteriori estimation in the case of incomplete data[7]. EM algorithm consists of two steps: Expectation step (E-step) and Maximization step (M-step) [8,9]. However, EM algorithm may not find the globally optimal solution and converge on local optimum. It is also sensitive to initialization. Some researchers want to overcome the limitation. In MCEM algorithm[10], the calculation of E-step is replaced by a Monte Carlo approximation. SEM algorithm[11,12] applied stochastic imputation principle to simulate the unobserved data based on the observed data at E-step. The two algorithms are effectively replacing E-step with a stochastic step.

We utilize EM algorithm to learn BNs parameters under the condition of data missing. In order to reduce the computational complexity of EM algorithm, we employ the Gibbs sampling to approximate the calculation of E-step. The Markov chain Monte Carlo sampling algorithm can directly obtain data sample from probability distribution. Gibbs sampling is an MCMC algorithm[13,14] and it is very suitable for BNs.

The rest of the paper is organized as follows. EM algorithm for BNs parameters estimation is described in section II. Gibbs sampling for E-step approximation is discussed in section III. In section IV, experiments for comparison between EM algorithm and the method proposed in the paper are done and the results are analyzed. Finally, section V concludes the paper and presents the prospect for future work.

2. EM algorithm for BNs parameters estimation

2.1. Definition of Bayesian networks

Position figures and tables at the tops and bottoms of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be centered below the figures; table captions should be centered above. Avoid placing figures and tables before their first mention in the text. Use the abbreviation “Fig. 1,” even at the beginning of a sentence.

A Bayesian network is a probabilistic graphical model[1] that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). A directed link between two BNs nodes represents the probabilistic causal dependency. BNs can be expressed as $B(G,P)$:

(1) G denotes the topology of BNs which is a directed acyclic graph. A set of random variables constitutes the nodes of the directed graph. If there is a link from node X_i to node X_j, then X_j is called the parent of X_i .

(2) P denotes the conditional probability parameters of BNs. Each node in BNs stores its conditional probabilities given its direct parents, which constitutes numerical parameters of BNs. In discrete case, the parameters of BNs correspond to the conditional probability tables (ab. CPTs).The CPT of node X_i is expressed as $\{P_i = P_i(X_i \mid pa(X_i))\}$, which pictures the mutual relationship between X_i and its parent nodes.

For BNs obeying the usual conditional independence assumptions, the function of BNs can be characterized by the joint probability distribution function:

$$p(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} p(X_i / pa(X_i)) \tag{1}$$

where $pa(X_i)$ denotes the parents of X_i .

The random variable discussed in the paper is discrete. We will use uppercase letters for random variables and lowercase letters for the instances of random variables in the following.

2.2. EM algorithm

EM algorithm provides a general approach to maximum likelihood parameter estimation with incomplete data. Each iteration of EM algorithm consists of E-step and M-step.

Let θ^t be the current estimation of parameter vector θ, the E-step computes the expected log-likelihood function of θ:

$$Q(\theta \mid \theta^t) = \int P(x \mid y, \theta^t) \log P(\theta \mid y, x) dx \tag{2}$$

where x is the random variable which data is missing in a piece of incomplete data sample and y is the random variable which data is not missing.

In the M-step, θ^{t+1} is obtained which can make $Q(\theta \mid \theta^t)$ maximized:

$$\theta^{t+1} = arg \sup_{\theta} Q(\theta \mid \theta^t) \tag{3}$$

Then, this process is iterated until convergence. EM algorithm can be utilized to BNs parameters estimation with incomplete data. According to the definition of BNs, E-step and M-step can be described as follows.

Suppose that $D = \{D_1, D_2, ..., D_m\}$ is a data set for BNs parameters learning and D belongs to incomplete data. $D_l (0 < l \le m)$ denotes a piece of data sample. If D_l is incomplete, X_l denotes the set of random variables which data is missing in D_l.

(1) E-step

On the basis of (2), E-step is computing as:

$$Q(\theta \mid \theta^t) = \sum_{l=1}^{m} \sum_{x_l \in \Omega X_l} P(X_l = x_l \mid D_l, \theta^t) \log P(D_l, X_l = x_l \mid \theta) \tag{4}$$

$Q(\theta \mid \theta^t)$ is the expected log-likelihood function of θ. If X_l is empty, then $P(X_l = x_l \mid D_l, \theta^t) = 1$ or else the value of $P(X_l = x_l \mid D_l, \theta^t)$ should be calculated. In order to obtain $P(X_l = x_l \mid D_l, \theta^t)$, data completion for each incomplete sample D_l should be made. After $X_l = x_l$ is added to D_l, a piece of complete sample $\{X_l = x_l, D_l\}$ is obtained. Then we can compute $P(X_l = x_l \mid D_l, \theta^t)$ according to BNs inference method. It is obvious that the process of computing $P(X_l = x_l \mid D_l, \theta^t)$ is equal to BNs inference with a high-computational complexity. And then, the value of $P(X_l = x_l \mid D_l, \theta^t)$ is set to be the weight of data sample $\{X_l = x_l, D_l\}$.

Next, define the characteristic function $\chi(i, j, k : D_l, X_l = x_l)$ for $\{X_l = x_l, D_l\}$:

$$\chi(i, j, k : D_l, X_l = x_l) = \begin{cases} 1 & if\ X_i = k \quad and\ par(X_i = j) \\ 0 & else \end{cases} \tag{5}$$

Suppose that the value of $P(X_l = x_l \mid D_l, \theta^t)$ is obtained, we will have:

$$w_{ijk}^t = \sum_{l=1}^{m} \sum_{x_l \in \Omega x_l} P(X_l = x_l \mid D_l, \theta^t) \chi(i, j, k : D_l, X_l = x_l) \tag{6}$$

For $\log P(D_l, X_l = x_x \mid \theta)$ in (4), it can be calculated as:

$$\log P(D_l, X_l = x_x \mid \theta) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \chi(i, j, k : D_l, X_l = x_l) \log \theta_{ijk} \tag{7}$$

where n is the number of nodes for a BN, r_i is the number of state for node X_i, q_i is the number of state for the parents of X_i, θ_{ijk} is the parameter of X_i under the condition of $X_i = k$ and $par(X_i = j)$.

According to (4), (6) and (7), we will have:

$$Q(\theta \mid \theta^t) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} w_{ijk}^t \log \theta_{ijk} \tag{8}$$

(2) M-step

In order to make $Q(\theta \mid \theta^t)$ maximized, θ should have the value as:

$$\theta_{ijk}^{t+1} = \begin{cases} \dfrac{w_{ijk}^t}{\sum_{k=1}^{r_i} w_{ijk}^t} & if \quad \sum_{k=1}^{r_i} w_{ijk}^t > 0 \\ \dfrac{1}{r_i} & else \end{cases} \tag{9}$$

3. Gibbs sampling for approximation

Gibbs sampling is one of the simplest Monte Carlo sampling procedures. It starts with a random setting of hidden states and then updates each hidden state according to the probability distribution conditioned on all the other states and the fixed parameters. The procedure of Gibbs sampling can be described as follows:

- Suppose that $X = \{X_i \mid 0 < i \le n_1\}$ are the unobserved random variables which states are hidden and $Y = \{Y_j \mid 0 < j \le n_2\}$ are the observed variables which states are fixed. Then the variable vector V is $V = \{X, Y\}$.
- Start a random state for vector V in which the state of Y is fixed and the state of X is random. Then, we can make the data sampling for one of variable in X, for example, X_i, according to the probability distribution $P(X_i \mid \overline{X_i}, Y)$. In $P(X_i \mid \overline{X_i}, Y)$, $\overline{X_i} \in X \setminus \{X_i\}$. In Gibbs sampling, $P(X_i \mid \overline{X_i}, Y)$ is also called as transition probability for X_i.
- Next, depending on the result of sampling $X_i = x_i$, we make the data sampling for another variable X_j according to the transition probability for X_j: $P(X_j \mid \overline{X_j, X_i}, X_i = x_x, Y)$. This step is iterated until the unobserved random variables in X are all sampled. Then a piece of complete data sample for V is obtained.

If $V = \{X, Y\}$ is the random variable vector for a BN, it should obey the conditional independence assumptions according to (1). For node X_i, it is conditionally independent to any other nodes given the value of Markov blanket of X_i. Then, we can simplify the calculation of transition probability, for example, $P(X_i \mid \overline{X_i}, Y)$:

$$P(X_i \mid \overline{X_i}, Y) = P(X_i \mid mb(X_i)) \tag{10}$$

where $mb(X_i)$ denotes the value of Markov blanket of X_i. It is obvious that Gibbs sampling is very suitable for BNs. We can utilize Gibbs sampling to improve EM algorithm for BNs parameters estimation.

As mentioned above, data completion for each incomplete sample D_l is made in E-step. During data completion, each state value of X_l will be added to D_l respectively and the corresponding $P(X_l = x_l \mid D_l, \theta^t)$ is needed to be computed. If Gibbs sampling is employed in data completion, the state of X_l is sampled according to transition probability. The value of $P(X_l = x_l \mid D_l, \theta^t)$ is not needed to be computed. Then, the computational complexity of EM algorithm will be reduced. With Gibbs sampling, EM algorithm for BNs parameters estimation is modified as follows:

(1) E-step

According to Gibbs sampling, start a random setting of hidden states in D_l. For each random variable X_i in X_l which data is missing in D_l, we can get its state value by sampling from $P(X_i \mid mb(X_i))$. Then a piece of complete

sample $\{X_l = x_l, D_l\}$ can be obtained. Data completion for each incomplete sample in data set D is made. Equation (6) is modified as:

$$w_{ijk}^t = \sum_{l=1}^{m} \chi(i, j, k : D_l, X_l = x_l) \tag{11}$$

According to (8), the expected log-likelihood function of θ can be computed as: $Q(\theta \mid \theta^t) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} w_{ijk}^t \log \theta_{ijk}$.

(2) M-step

For M-step, it is not needed to be modified. We calculate the value of θ according to (9).

4. Experiment and results

The lawn wet BN shown in Fig. 1 is chosen in the paper. It is widely used to test method of BNs parameter estimation. The parameters of the BN are given in the following tables.

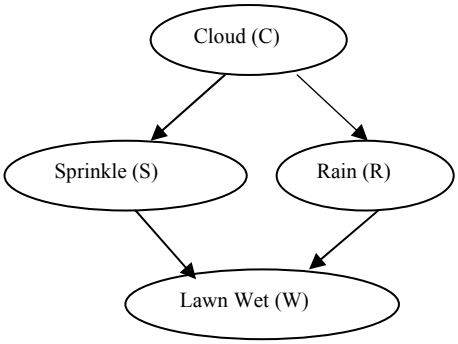

Fig. 1. Structure of lawn wet Bayesian network.

Table 1. Parameters of node C

C	True	False
	0.5	0.5

Table 2. Parameters of node S

C	C:true	C:false
S: true	0.2	0.6
S: false	0.8	0.4

Table 3. Parameters of node R

C:	C:true	C:false
R: true	0.8	0.2
R: false	0.2	0.8

Table 4. Parameters of node W.

S	S:true	S:false

R	R:true	R:false	R:true	R:false
W:true	1.0	0.8	0.6	0.2
W:false	0	0.2	0.4	0.8

According to the parameters and structure of lawn wet BN, random sampling algorithm is utilized to generate data samples for parameter estimation. In order to form incomplete sample, we will delete the data randomly in data sample. Six incomplete data set are generated which is represented by the form of sample number (missing data rate): 300(10%), 300(30%), 300(50%), 600(10%), 600(30%), 600(50%).

We realized EM algorithm and the method proposed in the paper by C++ programming. The experiments are performed under Windows 7 and Intel CPU 2.60GHz.

The results of experiments are shown in the follows. In Fig. 2, the proposed method in the paper which utilized Gibbs sampling in EM consumes less time than EM algorithm, especially in the case of high missing data rate. Because the proposed method does not converge, we choose the number of iterations while the value of expected log-likelihood function is set to stable. In Fig. 3, while the data sample is becoming bigger and the missing data rate is getting higher, the number of iterations increases in EM algorithm. However, the number of iterations in the proposed method is stable. It is possible that the number of iterations for the proposed method has a close relation with BNs structure. This will be studied in future work.

In Fig. 4, the KL (Kullback-Leibler) divergence of EM algorithm is lower than our method, which presents that the precision of parameters estimation for the proposed method is not good. But if the parameters estimated from the proposed method do not greatly affect the inference results of BNs, the method can be applied. It also needs further experiment for the degree of impact of KL divergence on inference results of BNs.

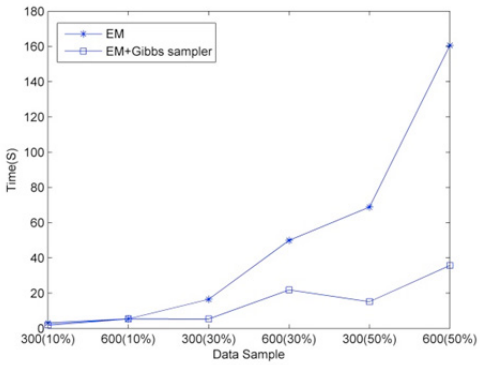

Fig. 2. Timeliness analysis of two algorithms.

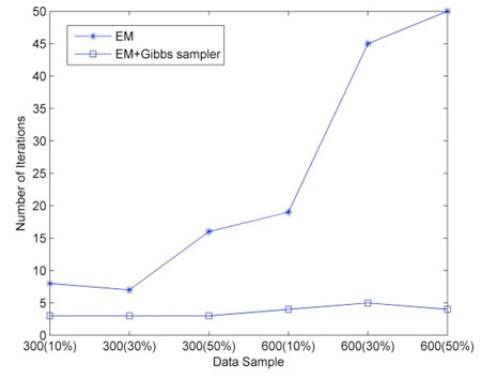

Fig. 3. Number of iterations analysis of two algorithms.

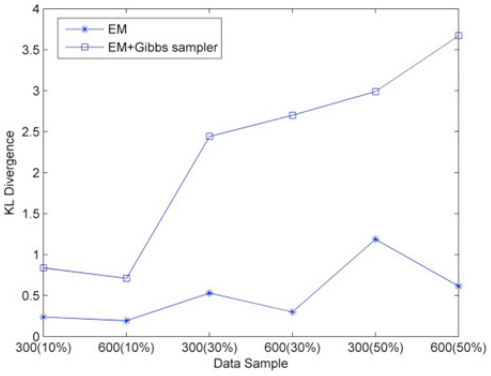

Fig. 4. KL divergence analysis of two algorithms.

4. Conclusion

A method for Bayesian networks parameters estimation which combines Gibbs sampling with EM algorithm is proposed in this paper. It can reduce the computational complexity effectively. The results of experiments show that the consumed time and the number of iterations of the proposed method are all less than those of EM algorithm. However, the KL divergence is higher than that of EM algorithm, which is a limitation for the proposed method.

In the future, we will make further experiments to study the impact of KL divergence on inference results of BNs and improve the method to overcome its limitation.

Acknowledgements

The research is supported by Key Science and Technology Program of Shaanxi Province, China (No. 2016GY-112), the Fundamental Research Funds for the Central Universities (No.KS7215406701) and National Natural Science Foundation of China (Grant No. 61472305).

References

1. J. Pearl. Probabilistic Reasoning in Intelligent System: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann, 1988.
2. Heckerman, D.: A Tutorial on Learning with Bayesian Network. Technical Report MSDTR-95-06, Microsoft Research (March 1995).
3. Bickson D, Baron D, Ihler A, et al. Fault Identification Via Nonparametric Belief Propagation [J]. IEEE Transactions on Signal Processing, 2011, 59(6): 2602-2613.
4. Zhang Lei, Ji Qiang. Bayesian Network Model for Automatic and Interactive Image Segmentation [J]. IEEE Transactions on Image Processing, 2011, 20(9): 2582-2593.
5. Fernandez R, Picard R. Recognizing affect from speech prosody using hierarchical graphical models [J]. Speech Communication, 2011, 53(9/10): 1088-1103.
6. K.Laskey,P.Costa,T.Janssen. Probabilistic ontologies for knowledge fusion. Proceedings of the 11th International Conference on Information Fusion, 2008,pp: 1-8.
7. A.P. Dempster, N.M. Laird and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B, 39: 1-38, 1977.
8. G. McLachlan and T. Krishnan. The EM Algorithm and Extensions. New York, NY, USA: Wiley, 1997.
9. J. Choi. Adaptive and Iterative Signal Processing in Communications. Cambridge, U.K.: Cambridge Univ. Press, 2006.
10. G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. Journal of the American Statistical Association, 85: 699-704, 1990.
11. G. Celeux and J. Diebolt. The SEM Algorithm: a Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem. Computational Statistics Quarterly, 2: 73-82, 1985.
12. M. Lavielle. A stochastic algorithm for parametric and non-parametric estimation in the case of incomplete data. IEEE Transactions on Signal Processing, 42: 3-17, 1995.
13. S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.
14. D. J. C. MacKay. Information Theory, Inference and Learning Algorithms. Cambridge, U.K.: Cambridge Univ. Press, 2003.