Bayesian Network's Parameter Update Method Based on Maximum Likelihood Estimates

Jiang Guo-Ping, Lin Ming-Chi, Zhang Cheng-Yi
Department of Management Engineering and Equipment Economy
Naval Engineering University
Wuhan Hubei, China, 430033
e-mail: gpjiang1029@163.com

***Abstract*—A probability parameter updating method for Bayesain network with fixed structure and initial probability parameters is discussed, when new data set is available. It should reserve the original probability parameters as much as possible and also reflect the probability distribution contained in data set furthest. Principle of parameters updating based on MLE is discussed, algorithms for complete data set and incomplete data set are put forward. A case study is carried out to show how the algorithm works.**

Keywords-Bayesian networks; maximum likelihood estimates; parameter update

I. INTRODUCTION

After a BN's topological structure is constructed, valuing its probability parameters is called as parameter learning. Bayesian arithmetic, Maximum Likelihood Estimates (MLE) and Adaptive Probabilistic Networks(APN) are the most familiar parameter learning algorithm. Bayesian algorithm [1] calculates all the possible value of the parameters in the given topological structure and seek the parameters' value with maximum posterior probability on the condition of topological structure and dataset are both known. MLE is put forward by Spiegelhalter [2], and it is considered as the especial example of Bayesian arithmetic. When parameters' prior information is neglected entirely, Bayesian arithmetic is converted into MLE. Russell [3]1995 has brought forward a parameter learning method based gradient degression, namely APN. This method takes the samples in the date set as evidence and propagate the evidence to all the topological structure. In order to obtain the parameters of the DAG with maximal probability contained in the dataset, it updates the parameters adopting the method of gradient degression.

This paper, probability parameters updating method for Bayesain network with fixed structure and initial probability parameters is discussed, when new data set is available. It should reserve the original probability parameters as much as possible and also reflect the probability distribution contained in data set furthest.

An outline of the remainder of the paper is as follows. In Section 2, we discuss the principle of parameters updating method based on MLE. In Section 3, BN's parameters updating algorithm for complete data set is discussed. In Section 4, we discuss parameter updating algorithm for incomplete data set. In Section 5, a case study is put forward to show how the algorithm works. Finally, in Section 6, we make a summarizer for this paper.

II. PROBABILITY PARAMETER UPDATING PRINCIPLE

The update of BN parameters needs to meet two requirements: One is to improve the parameters to fit the new data set, and the other is to keep the probability parameter information contained in the original network as much as possible. Therefore, there is a need to strike a balance between these two requirements.

The BN is known as $BN=(V,E)$, and $V=\{V_1,...,V_i,...,V_n\}$ is the set of all nodes in the network, and E is the set of all the arcs in the network. Suppose the set of probability parameters of the conditional probability of each node in the network is $\bar{\theta}=\{\bar{\theta}_{ijk}\}$. The updating of BN parameters is based on the new data set D, updating the probability parameter $\bar{\theta}$ to $\theta=\{\theta_{ijk}\}$, so that $\theta=\{\theta_{ijk}\}$ can be consistent with the original network parameters as much as possible while reflecting the potential probability distribution in the new data set.

According to the Maximum Likelihood estimation method (MLE), in order to meet the first requirement of parameter updating, we take the Likelihood function $L_D(\theta)=\frac{1}{N}\ln\prod_{l=1}^{N}P_\theta(y_l)$, and take the natural log of both sides, and get the following formula:

$$L_D(\theta)=\frac{1}{N}\ln(\prod_{l=1}^{N}P_\theta(y_l))=\frac{1}{N}\sum_{l=1}^{N}\ln P_\theta(y_l) \qquad (1)$$

Solving $\frac{\partial L_D(\theta)}{\partial\theta_{ijk}}=0$ we can obtain the only set of probability parameters suitable for the new data sets. However, the second requirement for updating the BN parameter is also required, namely, the new probability parameter can retain information in the original network as much as possible. Assume the distance between θ and $\bar{\theta}$ is:

$$d(\theta,\bar{\theta})=\frac{1}{2}\sum_{i,j,k}(\theta_{ijk}-\bar{\theta}_{ijk})^2 \qquad (2)$$

Let

978-1-5386-6987-7/18/$31.00 ©2018 IEEE

$$F(\theta) = \eta L_D(\theta) - d(\theta, \bar{\theta}) \tag{3}$$

η reflects the learning speed. The new probability parameter estimation can be obtained by maximizing formula (3).

For ease of solution, simplify (3) linearly. $L_D(\theta)$ is expanded at $\bar{\theta}$ into the Taylor formula with the Lagrange remainder, getting

$$L_D(\theta) = L_D(\bar{\theta}) + \nabla L_D(\bar{\theta})(\theta - \bar{\theta}) \tag{4}$$

Where $\nabla L_D(\bar{\theta})$ is the gradient vector of $L_D(\bar{\theta})$, and $\nabla_{ilk} L_D(\bar{\theta})$ corresponds to parameter $\bar{\theta}_{ijk}$ in the gradient vector. So, we substitute (4) into (3), and get

$$F(\theta) = \eta(L_D(\bar{\theta}) + \nabla L_D(\bar{\theta})(\theta - \bar{\theta})) - d(\theta, \bar{\theta}) \tag{5}$$

Because of the structural decomposition of BN, $\nabla_{ijk} L_D(\bar{\theta}) = \frac{1}{\bar{\theta}_{ijk}} \frac{\sum_{l=1}^{N} P_{\bar{\theta}}(v_{ik}, pa_{ij} \mid y_l)}{N}$,where $P_{\bar{\theta}}(V_{ik}, pa_{ij} \mid y_l)$ is the probability of V_i 's parent node $Pa(V_i) = pa_{ij}$, when the probability parameter is $\bar{\theta}$ and $V_i = v_{ik}$ in sample y_l. $P_{\bar{\theta}}(v_{ik}, pa_{ij} \mid y_l) = \begin{cases} 1 & if \;\; V_i = v_{ik}, \;\; Pa(V_i) = pa_{ij} \\ 0 & otherwise \end{cases}$ Thus, $\nabla_{ijk} L_D(\bar{\theta}) = \frac{1}{\bar{\theta}_{ijk}} \cdot \frac{\sum_{l=1}^{N} P_{\bar{\theta}}(v_{ik}, pa_{ij} \mid y_l)}{N}$. And $\sum_{l=1}^{N} P_{\bar{\theta}}(v_{ik}, pa_{ij}, y_l)$ is the number of sample points of the parent node set $Pa(V_i) = pa_{ij}$ of node $V_i = v_{ik}$ in the new data set D. Suppose $\sum_{l=1}^{N} P_{\bar{\theta}}(v_{ik}, pa_{ij}, y_l) = N_{ijk}$, then

$$\nabla_{ijk} L_D(\bar{\theta}) = \frac{N_{ijk}}{N \cdot \bar{\theta}_{ijk}} \tag{6}$$

To get the maximum value of (5) is a conditional extremum problem that satisfies the constraint $\sum_{k}' \theta_{ijk} = 1$. Using Lagrange multipliers to construct $\Psi = F(\theta) + \sum_{i,j}' \gamma_{ij} (\sum_{k}' \theta_{ijk} - 1)$, let the partial derivative of Ψ with respect to θ_{ijk} be 0. We get:

$$\eta \nabla_{ijk} L_D(\bar{\theta}) - \frac{\partial}{\partial \theta_{ijk}} \left(\frac{1}{2} \sum_{i,j,k} (\theta_{ijk} - \bar{\theta}_{ijk})^2\right) + \gamma_{ij} = 0 \tag{7}$$

Substitute (6) into the above formula,

$$\eta \cdot \frac{N_{ijk}}{N \cdot \bar{\theta}_{ijk}} - (\theta_{ijk} - \bar{\theta}_{ijk}) + \gamma_{ij} = 0 \tag{8}$$

Each θ_{ijk} corresponds to an equation (8). Union all equations and constraint equations. Then we get,

$$\begin{cases} \theta_{ijk} = \frac{N_{ijk}}{N} \cdot \frac{\eta}{\bar{\theta}_{ijk}} + \gamma_{ij} + \bar{\theta}_{ijk} & * \\ \sum_k \theta_{ijk} - 1 = 0 & ** \end{cases} \tag{9}$$

From (9), the probabilistic parameters of BN can be updated to get $\theta = \{\theta_{ijk}\}$.

Let's continue with the equation (9). Substitute θ_{ijk} into the constraint equation (**), then $\sum_k \left(\frac{N_{ijk}}{N \cdot \bar{\theta}_{ijk}} \cdot \eta + \gamma_{ij} + \bar{\theta}_{ijk}\right) = 1$, that is

$$\gamma_{ij} = -\frac{\eta}{K \cdot N} \sum_k \frac{N_{ijk}}{\bar{\theta}_{ijk}} \tag{10}$$

Substituting equation (10) into θ_{ijk} 's computing equation (*), θ_{ijk} can be solved. Therefore, the solution of θ_{ijk} is only related to the set of parent nodes of node V_i and V_i, and is independent of other nodes. The calculation of the update probability parameter can be localized, so the above analysis is still valid for the incomplete data set (the data set contains only a subset of the node set in the BN).

$$\theta_{ijk} = \eta\left(\frac{N_{ijk}/N}{\bar{\theta}_{ijk}} - \frac{1}{K} \sum_k \frac{N_{ijk}/N}{\bar{\theta}_{ijk}}\right) + \bar{\theta}_{ijk} \tag{11}$$

The selection of η : when $\eta < 1$, the updated θ_{ijk} can be regarded as the weighted sum of the probability parameter and the original probability parameter in the new database. So, θ_{ijk} is in the interval of these two values, and the convergence is slower. When $\eta < 1$, it means that the updated θ_{ijk} is more remote from the original probability parameter than the probability parameter contained in the new database, and the convergence speed is faster.

III. Parameter Updating Algorithm for the Complete Data Set

Suppose there are no absence of data in the new data set D, that is, if the node V_i in BN is contained in the data set D, then each instance in D records the value of the node. From the above discussion, we establish the BN probability parameter updating algorithm. According to (11), the update of the probability parameter $\bar{\theta}_{ijk}$ of V_i only needs to obtain N_{ijk}, $k = 1, \ldots, val(V_i)$, where $val(V_i)$ is the number of V_i. Therefore, you can record the N_{ijk} of each node by traversing the database, and then update $\bar{\theta}_{ijk}$ by (11).

Observe equation (11). When the initial probability parameter is zero, the division operation cannot be performed. Therefore, when the initial probability parameter is zero, we do not update the parameter, that is, when $\bar{\theta}_{ijk} = 0$, $\theta_{ijk} = 0$.

When the data set D is complete, that is, all nodes in the BN are included, the probability parameter updating algorithm of BN is shown as follows.

Step 1: For each node V_i, the structure $StructNode(i)$ is set up. The structure contains nodes, parent nodes, counting matrix, node values, and parent node values.

Step2: Specify the η. Initialize the counting matrix. $Num(i, j, k) = 0$ where $i = 1, \ldots, n$, $j = 1, \ldots, value(V_i)$ and $k = 1, \ldots, value(Pa(V_i))$, where $value(V_i)$ is the number of node V_i.

Step 3: Go through each record in D. For each record y_l, the corresponding $Num(i,j,k) = Num(i,j,k) + P(v_{ik}, pa_{ij} \mid y_l)$ is calculated for each node V in BN and its parent node set.

Step 4: Solve for $\theta = \{\theta_{ijk}\}$ by equation (11).

IV. Parameter Updating Algorithm for the Incomplete Data Set

Since there is a large number of incomplete data in engineering practice, it is necessary to discuss whether the parameter updating algorithm based on maximum likelihood estimation is applicable to incomplete data sets.

According to the foregoing discussion, the update of the probability parameters of each node is only related to the node and its set of parent nodes. Therefore, if the data set D does not contain all nodes in BN, the algorithm in figure 1 is applied to the node subset $V' = \{V_i, V_i \in V \wedge V_i \in Node(D) \wedge Par(V_i) \subset Node(D)\}$, where $Node(D)$ is all nodes contained in the data set D, and the probability parameters of the node subset V' of BN can be updated.

Node 1

value	present	absent
probability	0.2	0.8

Node 2

Parent node (1)	increased	Not increased
present	0.8	0.2
absent	0.2	0.8

Node 3

Parent node (1)	increased	absent
present	0.2	0.8
absent	0.05	0.95

Node 4

Parent node (1)	Parent node (3)	increased	absent
present	increased	0.5	0.5
present	absent	0.5	0.5
absent	increased	0.5	0.5
absent	absent	0.5	0.5

Node 5

Parent node (3)	present	absent
present	0.5	0.5
absent	0.5	0.5

Figure 1. Initial probability parameters.

V. A Case Study

We choose one typical Bayesian network case-Cancer network from GeNIe [4], a graph processing software developed by Decision System Laboratory in University of Pittsburgh to show how the algorithm discussed in this paper works. And BNT [5] is used to help program, which is an opening code Matlab Bayes Net toolbox compiled by Kevin Murphy.

The topology and true probability parameters of the Cancer network are known. According to the real probability distribution $\theta = \{\theta_{ijk}\}$ of the Cancer network, the Monte Carlo method is used to simulate the sampling and obtain 1000 data, namely, the new data set D, $N = 1000$. The initial probability parameter set $\overline{\theta}$ for the Cancer network is given, as shown in Figure 1. For different learning speed η, we do multiple probability update learning, and define distance $\sum_{i,j,k}(\theta_{ijk} - \theta_{ijk})^2$ to measure the distance between the new probability parameter and the real parameter. Figure 2 shows the distance between the new probability parameter and the true probability parameter updated based on the initial parameters of Figure 1 for different η.

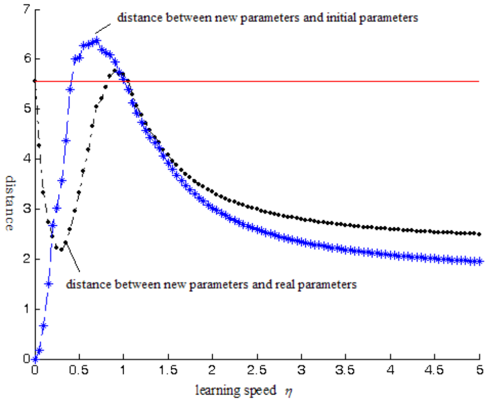

Figure 2. Distance between the probability parameters updated at different learning speed and the real value.

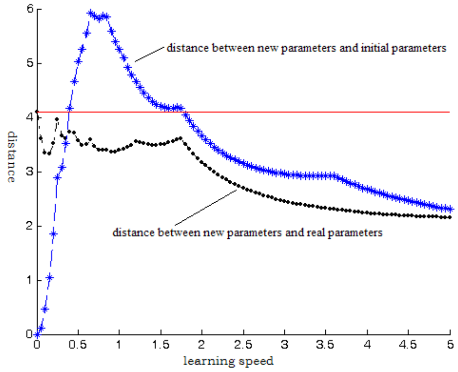

Figure 3. Learning effect 1.

In order to study the relationship between the update effect of probability parameters and different η values, the random probability distribution was used as the initial probability parameter, and several experiments were conducted. Figure 3 and Figure 4 are two update results for different initial probability parameters. The red line in the figure shows the distance between the initial probability parameter and the real probability parameter. The overall trend of the distance of the updated new parameter and the real parameter decreases with the increase of η value. However, the new parameters that can be learned in some η values are not as good as the initial parameters (as shown in Figure 4). Similar phenomenon is normal, as the parameter update principle is two different requirements: one is to balance the distance between the new parameters and the initial parameters, and the other is that the new parameters reflect the accuracy of the data set. No unilateral measurement can reflect the effect of parameter updates.

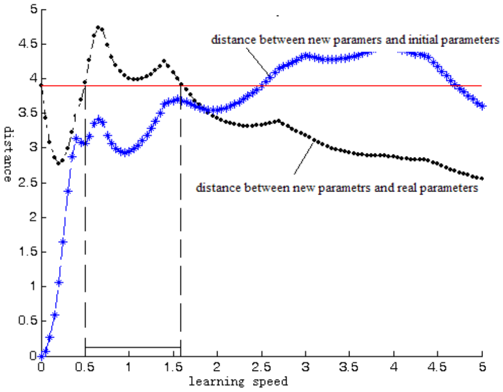

Figure 4. Learning effect 2.

The above test results show that when the learning speed $\eta \in (0,1)$, the new probabilistic parameters updated are between the original parameters and the true parameters (parameters of the data set implication). When $\eta > 1$ the updated parameters are further away from the original parameters than the probability parameters contained in the database. The larger the η, the greater the influence of the data set on the new parameters, the closer the new parameters are to the probability parameters contained in the database.

VI. Summary

The main goal of this paper is to propose a method to update BN's probability parameters with fixed structure and original parameters when new data set is available, thus to retain the probability information of the original model and also reflect the probability distribution contained in the new data set. After the principle of parameters updating based on MLE discussed, algorithms for complete data set and incomplete data set are put forward. A case study is carried out at last and learning effects by different learning speeds are discussed also.

References

[1] G.F. Cooper and E. Herskovitz. A Bayesian method for the induction of probabilistic networks from data. Machine learning, 9,1992: 309-347.

[2] Nevin Lianwen Zhang. Irrelevance and parameter learning in Bayesian networks. Artificial Intelligence, 1996: 359-373.

[3] Finn Verner Jensen. Gradient Descent Training of Bayesian Networks. ECSQARU 1999: Symbolic and Quantitative Approaches to Reasoning and Uncertainty: 190-200.

[4] Marek Jozef Druzdzel. GeNIe: A Development Enviroment for Graphical Decision-Analytic Models. AMIA Symposium, January 1999.

[5] Kevin P.Murphy. Machine Learning: A probabilistic Perspective. The MIT Press, Cambridge, Massachusetts, London, England, 2012.