Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/080147

Study of Four Types of Learning Bayesian Networks Cases

Yonghui Cao[1,2,*]

[1] School of Economics and Management, Henan Institute of Science and Technology, Xinxiang 453003, P. R. China
[2] School of Management, Zhejiang University, Hangzhou 310058, P.R. China

Received: 5 Jul. 2013, Revised: 16 Nov. 2013, Accepted: 18 Nov. 2013
Published online: 1 Jan. 2014

Abstract: As the combination of parameter learning and structure learning, learning Bayesian networks can also be examined, Parameter learning is estimation of the dependencies in the network. Structural learning is the estimation of the links of the network. In terms of whether the structure of the network is known and whether the variables are all observable, there are four types of learning Bayesian networks cases. In this paper, first introduce two cases of learning Bayesian networks from complete data: known structure and unobservable variables and unknown structure and unobservable variables. Next, we study two cases of learning Bayesian networks from incomplete data: known network structure and unobservable variables, unknown network structure and unobservable variables.

Keywords: Bayesian networks, Network Structure, Observable Variables

1 Introduction

The main driving force to choose Bayesian networks is that Bayesian networks have a bidirectional message passing architecture. Learning from the evidence can be interpreted as unsupervised learning. Similarly, expectation of an action can be interpreted as supervised learning. Since Bayesian networks pass evidence (data) between nodes and use the expectations from the world model, they can be considered as bi-directional learning systems. In addition to bi-directional message passing, Bayesian networks have several important features such as allowing subjective a priori judgements, direct representation of causal dependence, nonmonotonic reasoning, distillation of sensory experience and the ability to imitate human thinking process.

A Bayesian network is a graphical model that finds probabilistic relationships among variables of the system. There are a number of models available for data analysis, including rule bases, decision trees and artificial neural networks. There are also several techniques for data analysis such as classification, density estimation, regression and clustering. One may wonder what Bayesian networks and Bayesian methods have to offer to solve such problems.

This paper is devoted to answering the question: how can Bayesian networks be learned from data? The process of learning Bayesian networks takes different forms in terms of whether the structure of the network is known and whether the variables are all observable. The structure of the network can be known or unknown, and the variables can be observable or hidden in all or some of the data points. The latter distinction can also be expressed as complete and incomplete data. Consequently, there are four cases of learning Bayesian networks from data; known structure and observable variables, unknown structure and observable variables, known structure and unobservable variables, and unknown structure and unobservable variables. Learning Bayesian networks can also be examined as the combination of parameter learning and structure learning. Parameter learning is estimation of the conditional probabilities (dependencies) in the network. Structural learning is the estimation of the links of the network. The four types of learning Bayesian networks cases are discussed in the following paragraphs.

* Corresponding author e-mail: caoyonghui2000@126.com

2 Four Types of Learning Bayesian Networks

2.1 Known Network Structure and Observable Variables

This is the easiest and the most studied case of learning Bayesian networks in the literature [1, 2]. The network structure is specified, and the inducer only needs to estimate the parameters. The problem is well understood and the algorithms are computationally efficient. Despite its simplicity, this problem is still extremely useful, because numbers are very hard to elicit from people. Additionally, it forms the basis for everything else in Bayesian learning.

Because every variable is observable, each data case can be pigeonholed into the CPT entries corresponding to the values of the parent variables at each node. The pigeonhole principle essentially states that if a set consisting of more than $k \cdot n$ objects is partitioned into n classes, then some classes receive more than k objects [3]. Therefore, estimations will be highly accurate since every variable is observable.

Learning is achieved simply by calculating conditional probability table (CPT) entries using estimation techniques such as Maximum Likelihood Estimation (MLE) and Bayesian Estimation. For simplicity, MLE and Bayesian estimators will be explained by employing parameter learning for a single parameter.

Assume that an experiment was conducted by flipping a thumbtack in the air. The thumbtack comes to land as either heads or tails. As usual, the different tosses are assumed to be independent, and the probability of the thumbtack landing heads is some real number .Therefore, the goal is to estimate . Assume that we have a set of instances $d[1],\cdots,d[M]$ such that each instance is sampled from the same distribution and independently from the rest. The goal is to find a good value for the parameter θ. A parameter is good if it predicts the data well. In other words, if data are very likely given the parameter, the parameter is a good predictor. The likelihood function is defined as

$$L(D|\theta) = P(D|\theta) = \prod_{m=1}^{M} P(d[m]\theta) \tag{1}$$

Thus, the likelihood for a sequence H,T,T,H,H is

$$L(D|\theta) = \theta(1-\theta)(1-\theta)\theta\theta \tag{2}$$

To calculate the likelihood we need to know number of heads N_h and the number of tails N_t.These are the sufficient statistics for this learning problem. A sufficient statistic is a function of the data that summarize the relevant information for computing the likelihood.

The Maximum Likelihood Estimation (MLE) principle tells us to choose θ that maximizes the likelihood function. The MLE is one of the most commonly used estimators in statistics. For the above problem, the estimation of the parameter is as expected.

$$\hat{\theta} = \frac{N_h}{N_h + N_t} \tag{3}$$

The MLE estimate seems plausible, but is overly simplistic in many cases. Assume that the experiment with the thumbtack is done and 3 heads out of 10 are recorded. It may be quite reasonable to conclude that the parameter θ is 0.3. On the other hand, what if the same experiment is done with a dime and also 3 heads are recorded. We would be much less likely to jump the conclusion that the parameter of the dime is 0.3 because we have a lot more experience with tossing dimes. Thus, we have a lot more prior knowledge about their behavior.

Using MLE, we cannot make the following distinctions: between a thumbtack and a dime, and between 10 tosses and 1,000,000 tosses of a dime. On the other hand, there is another method recommended by Bayesian statistics. The MLE is a frequentist approach since it relies on the frequency in the data. Another approach is the Bayesian approach that assumes that there is unknown but fixed parameter θ.It estimates the parameter with some confidence, i.e., it calculates a range such that, if the parameter is out of this range, the probability of the data is very low.

The Bayesian approach deals with uncertainty over anything that is unknown by putting a distribution over it. In other words, the parameter θ is treated as a random variable and a distribution $P(\theta)$ is defined over it. Therefore, we can tell how likely the parameter is to take on one value versus another. In other words, we now have a joint probability space that contains both the tosses and the parameter. This joint probability is easy to find given our prior distribution over θ. Let $X[1],\cdots,X[M]$ be our coin tosses. The conditional probabilities $P(X[M]\theta)$are according to θ, i.e., $P(X[M] = H|\theta) = \theta$. Now, the value of the next toss $X[M+1]$can be predicted by

$$P(X[M+1]|X[1],\cdots,X[M]) = \int P(X[M+1]|\theta)P(\theta|D)d\theta \tag{4}$$

where

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{5}$$

The first term in the numerator is the likelihood, the second is the prior over parameters, and the third is a normalizing factor, which is the marginal probability of the data. If we reconsider the thumbtack problem again with a uniform prior over θ in the interval $[0,1]$then $P(D|\theta) = \theta^{N_k}(1-\theta)^{N_t}$ is proportional to the likelihood . After plugging this into the integral and doing all the math and normalizing, it can be shown that the following equation holds.

$$P(X[M+1]|D) = \frac{N_k + 1}{N_k + N_t + 2} \tag{6}$$

Clearly, as the number of samples grows, the Bayesian estimator and the MLE estimator converge to each other. This result depends on the use of uniform prior. In the Bayesian networks literature, the most commonly used class of priors are the Dirichlet priors because it turns out that most of the interesting calculations can be done in closed form. The conjugacy of the Dirichlet priors allows us to have the posterior probabilities in the same form as prior probabilities. Therefore, we can do sequential updating within the same representations and the closed form solution can found both for the update and the prediction problem in many cases.

Recall that a multinomial is parameterized via a set of parameters $\theta_1 \cdots, \theta_k$ such that $\sum_i \theta_i = 1$; θ_i corresponds to the probability of ith outcome. A Dirichlet distribution over this set of parameters $\alpha_1, \cdots, \alpha_k$ is defined via a set of hyper parameters .Then, the generalization can be written as

$$Dir(\theta|\alpha_1, \cdots, \alpha_k) = \frac{\Gamma(\alpha)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1} \tag{7}$$

All of the results regarding prediction and computing the posterior extend in the obvious way. That is, if θ is distributed as in (7), then

$$P(x_i) = \frac{\alpha_i}{\sum_j \alpha_j} \tag{8}$$

and if there is a data set D whose sufficient statistics are $N_1, \cdots, N_k$, then

$$P(\theta|D) = Dir(\theta|\alpha_1 + N_1, \cdots, \alpha_k + N_k) \tag{9}$$

To generalize these results for a Bayesian network, we need to define the sufficient statistic as $N(x,u)$ for the event $X = x$ and the parents $U = u$. In the MLE case, the estimation of the parameters can be calculated as

$$\hat{\theta}_{x|u} = \frac{N(x,u)}{N(u)} \tag{10}$$

Similarly, in the Bayesian case, the parameter estimation is calculated as

$$\hat{\theta}_{x|u} = Dir(\alpha_1 + N(x_1,u), \cdots, \alpha_k + N(x_k,u)) \tag{11}$$

If the data were actually generated from the given network structure, then both methods converge asymptotically to the correct parameter setting. If not, then they converge to the distribution with the given structure that is closest to the distribution from which the data were generated. Both estimations can be implemented online by accumulating sufficient statistics.

The process above is the method by which Bayesian network parameters are learned when the network topology is known and all variables are fully observable. The next section provides an overview of some proposed methods in the literature if the structure of the network is not known in advance.

2.2 Unknown Network Structure and Observable Variables

In this case, the inducer is given the set of variables in the model, and needs to select the arcs between them and estimate the parameters. This problem is very useful for a variety of applications; in general, when we are given a new domain with no available domain expert, and want to get all of the benefits of a BN model. It is also useful for data-mining style applications, where there are masses of data available and we would like to interpret them. In addition to providing a model that will allow us to predict behavior of cases that we have not seen, the structure also gives the expert some indication of what attributes are correlated. The algorithms for this problem are combinatorially expensive. They basically reduce to a heuristic search over the space of BN structures.

There has been some attention given to the problem of unknown network structure in the literature. The key aspect of the problem is to reconstruct the topology of the network from fully observable variables. In the literature, this is considered as a discrete optimization problem solved by a greedy search algorithm in the space of structures. Some examples of the greedy search algorithm can be found in [5, 6].

A MAP (Maximum a Posterior) analysis of the most likely network structure has been studied in [5] and [6] when the data are fully observable. The resulting algorithms are capable of recovering fairly large networks from large data sets with a high degree of accuracy [7]. On the other hand, they usually adopt a greedy approach to choosing the set of parents for a given node because the problem of finding the best topology is intractable.

There are two main approaches to structure learning in BNs:

Constraint based: Perform tests of conditional independence on the data, and search for a network that is consistent with the observed dependencies and independencies.

Score based: Define a score that evaluates how well the (in) dependencies in a structure match the data, and search for a structure that maximizes the score.

Constraint-based methods are more intuitive. They follow the definition of a BN more closely. They also separate the notion of the independence from the structure construction. The advantage of score-based methods is that they less sensitive to errors in individual tests. Compromises can be made between the extent to which variables are dependent in the data and the cost of adding the edge.

The score-based methods operate on the same principle: a scoring function is defined for each network structure, representing how well it fits the data. The goal is to find the highest-scoring network structure. The space of Bayesian networks is a combinatorial space, consisting of a super exponential number of structures. Thus, it is not clear how one can find the highest-scoring network even with a scoring function. In general, the problem of finding

the highest-scoring network structure is NP-hard. On the other hand, the problem of searching a combinatorial space with the goal of optimizing a function is very well studied in AI literature. Consequently, the answer is to define a search space, and then do heuristic search.

In light of the above statements, a BN structure learning algorithm requires the following components be determined:

i) Scoring function for different candidate network structures.

ii) The definition of the search space: operators that take one structure and modify it to produce another.

iii) A search algorithm that does the optimization search.

Each component will be discussed separately. The three main scoring functions commonly used to learn Bayesian networks are the log-likelihood , the one based on the principle of minimal description length (MDL) [8] which is equivalent to Schwarz' Bayesian information criterion (BIC), and Bayesian score.

The log-likelihood function is simply the log of the likelihood function. That is,

$$l(D|B,\theta_B) = \log L(D|B,\theta_B) \tag{12}$$

The log-likelihood is easier to analyze than the likelihood, because the logarithm turns all the products into sums. Therefore,

$$L(D|B,\theta_B) = \prod_m P(d[m]|B,\theta_B) \tag{13}$$

and, the following equation can be written:

$$L(D|B,\theta_B) = \sum_m logP(d[m]|B,\theta_B) \tag{14}$$

There are a couple of important things to note about the log-likelihood. The log-likelihood increases linearly with the length of data, M. The higher scoring networks are those where the node and the parents are highly correlated. Adding a node to the networks always increases the log-likelihood. As a result, the network structure that maximizes the likelihood is often the fully connected network. This is the deficiency of the log-likelihood score and is not desired. Thus, a score that makes it harder to add edges is necessary. In other words, we would like to penalize structures with too many edges.

One possible formulation of this idea is called the MDL score. It is defined as:

$$Score_{MDL}(B:D) = l(D|B,\hat{\theta}_B) - \frac{logM}{2} Dim(B) - DL(B) \tag{15}$$

Where $Dim(B)$ is the number of independent parameters in B and $DL(B)$ is the number of bits (the description length) required to represent the structure of .The abbreviation MDL stands for minimum description length. The MDL score is a compromise between fit to data and model complexity. Adding a variable as a parent causes the log-likelihood term to increase, but so does the penalty term. [6] There will be an edge addition if its increase to the likelihood is worth it.

Another commonly used score is called Bayesian score. In this case, the network score is evaluated as the probability of the structure given the data. The Bayesian score has the following form:

$$Score_{BDE}(B:D) = P(B|D) = \frac{P(D|B)P(B)}{P(D)} \tag{16}$$

As usual $P(D)$ is constant, so it can be ignored when different structures are compared. Therefore, the model maximizes $P(D|S)P(S)$, where S represents a structure. The ability to ascribe a prior over structures gives us a way of preferring some structures to others. Here, the probability $P(D|B)$ can be calculated as

$$P(B|D) = \int P(D|\theta_B,B)P(\theta_B|B)d\theta_B \tag{17}$$

From Equation (17), one can see that the more parameters we have the more variables we are integrating over. As a result, each dimension causes the value of the integral go down because the "hill" of the likelihood function is a smaller fraction of the space. Therefore, this idea gives preference to networks with fewer parameters. It can be shown that the Bayesian score is a general form of MDL score. The MDL score can be viewed as an approximation of the Bayesian score. Therefore, the Bayesian score is also a compromise between the model complexity and fit to the data.

Several ways of scoring different Bayesian network structures have been explained. Different scores have been explored in terms of the network complexity and how the network fits to the correlation in the data. Now, the goal is to find the network that has the highest score. In other words, training data D, the scoring function, and a set of possible structures are the inputs of the search algorithm while the desired output is a network that maximizes the score. It can be shown that finding maximal scoring network structures where nodes are restricted to having at most k parents is NP-hard for any $k > 1$. Therefore, a heuristic search is resorted to for this optimization problem. A search space is defined, where the states in the space are possible structures and the operators denote the adjacency of structures. This space is traversed looking for high-scoring functions to complete the optimization. The obvious operators in the search spaces are add an edge, delete an edge, and reverse an edge. The search starts with some candidate network, which may be the empty one, or one that some expert has provided as a starting point. [7, 8]Then, applying the operators, the high-scoring network is searched in the space. The parameters of the network are calculated by using training data D.

The most commonly used algorithm for optimization search is simple greedy hill climbing. Even though the hill-climbing method is commonly used, it has several

key problems such as local maxima where all one-edge changes reduce the score and plateaus where a large set of neighboring networks that have the same score. There are some clever tricks that avoid some of these problems such as TABU-search, random restart, and simulated annealing. In general, greedy hill climbing with random start works quite well in practice. In a world, we examined methods for learning a Bayesian network from fully observable data.

2.3 Known Network Structure and Unobservable Variables

The learning of Bayesian networks with known structure and unobservable variables has been studied by Lauritzen [9, 10], Olesen et al. [11], and Spiegelhalter and Cowel [12]. The algorithm that these papers describe is the expectation maximization (EM) algorithm [13] The EM algorithm is an iterative method to calculate maximum likelihood estimates (MLEs) and MAP estimates of the network parameters. The EM algorithm alternates an expectation step a maximization step. In the expectation step, unknown quantities depending on missing entries are replaced by their expectations in the likelihood. In the maximization step, the likelihood completed in the expectation step is maximized with respect to the unknown parameters, and the resulting estimates are employed to replace unknown quantities in the next expectation step. The algorithm continues until the difference between successive estimates is smaller than a fixed threshold. Lauritzen states some difficulties with the use of EM algorithm such as slow convergence rate and local maxima. He then suggests that the gradient descent algorithm can be used as a possible alternative.

The third possible approach, introduced by Heckerman [14], is to use Gibbs sampling (GS). Gibbs Sampling is one of the most popular Markov Chain Monte Carlo methods for Bayesian inference. The GS algorithm generates a value for the missing data from some conditional distributions and provides stochastic estimations of the posterior probabilities [15]. To illustrate Gibbs sampling, let us approximate the probability density $p(\theta_s|D,s^h)$ for the configuration of parameters of a particular network S^h given an incomplete data set $D=\{Y_1,\cdots,Y_N\}$ and a Bayesian network for discrete variables with independent Dirichlet priors. To approximate $p(\theta_s|D,S^h)$, we first initialize the states of the unobserved variables in each case somehow (e.g., at random). Therefore, we have a complete random sample D.Then, we choose some variable X_{il} (variable X_i in case l) that is not observed in the original random sample D, and reassign its states according to the probability distribution

$$p(x'_{il}|D_c\backslash x_{il},S^h)=\frac{p(x'_{il},D_c\backslash x_{il}|S^h)}{\sum_{x''_{il}}p(x''_{il},D_c\backslash x_{il}|S^h)} \qquad (18)$$

Where $D_c\backslash x_{il}$ denotes the data set D with observations x_{il} removed, and the sum in the denominator runs over all states of variable x_{il}.Then, this reassignment for all unobservable variables in D is repeated producing a new complete random sample D_c. Using this data set, the posterior density $p(\theta_s|D_c,S^h)$is computed. Finally, the three steps are iterated and the average of $p(\theta_s|D_c,S^h)$ is used as our approximation.

Both the GS and EM algorithms use a basic strategy called the missing information principle: fill in the missing observations on the basis of the available information. Unfortunately, these approximate methods are prone to errors when little and/or biased information is available about the pattern of the missing data [16].

In recent years, an exciting solution to this problem was proposed by Sabestiani and Ramoni [17]. The algorithm is called Bound and Collapse (BC), which is a deterministic method to estimate conditional probabilities from incomplete data. The method bounds the set of possible estimates consistent with the available information by computing the minimum and the maximum estimates that would be gathered from all possible completions of the database. These bounds then collapse into a unique value via a convex combination of the extreme points with weights depending on the assumed pattern of missing data [18].

The basic intuition behind BC is that an incomplete database is still able to constrain the possible estimates within a set and that, when exogenous information is available on the pattern of missing data, this can be used to select a point estimate within the set of possible ones. Let X be a variable in the set $X=\{X_1,\cdots,X_n\}$ with parent variable π_i. Sebastiani and Ramoni show that the maximum Bayesian estimate of $p(x_{ik}|\pi_{ij})$is

$$p^{\cdot}(x_{ik}|\pi_{ij},D)=\frac{\alpha_{ijk}+n(x_{ik}|\pi_{ij})+n^{\cdot}(x_{ik}|\pi_{ij})}{\alpha_{ij}+n(\pi_{ij})+n^{\cdot}(x_{ik}|\pi_{ij})} \qquad (19)$$

and the minimum Bayesian estimate is

$$p_{\cdot}(x_{ik}|\pi_{ij},D)=\frac{\alpha_{ijk}+n(x_{ik}|\pi_{ij})}{\alpha_{ij}+n(\pi_{ij})+n_{\cdot}(x_{ik}|\pi_{ij})} \qquad (20)$$

Where α_{ijk} are the Dirichlet hyperparameters, $n^{\cdot}(x_{ik}|\pi_{ij})$ and $n_{\cdot}(x_{ik}|\pi_{ij})$ are maximum and minimum achievable virtual frequencies of $(x_{ik}|\pi_{ij})$ in the incomplete data, respectively. The frequency $n(x_{ik}|\pi_{ij})$ is the number of occurrences of $(x_{ik}|\pi_{ij})$ in the data. The maximum and minimum values of the virtual frequency are calculated filling the missing entries in order to have maximum and minimum number of occurrences of and counting the number of occurrences of the entry $(x_{ik}|\pi_{ij})$, respectively. The probability interval defined by $[p_{\cdot}(x_{ik}|\pi_{ij},D),\ p^{\cdot}(x_{ik}|\pi_{ij},D)]$ contains all possible estimates consistent with D, therefore it is sound and it is the tightest estimable interval.

The main feature of the BC method is its independence of the distribution of missing data because it does not attempt to infer them: with no information on the missing data mechanism, an incomplete database can only provide bounds on the possible estimates that could be learned [19]. A complete database is just a special case; within available data are enough to constrain the set of possible estimates to a single point. Another advantage of this method is that the width of each interval accounts for the amount of information available in D about the parameter to be estimated. Each interval represents a measure of quality of probabilistic information conveyed by the database about a parameter: the wider the interval, the greater the uncertainty due to the incompleteness of the database. In this way, intervals provide an explicit representation of the reliability of the estimates, which can be taken into account when the extracted BN is employed to perform a particular task.

The second step of the BC method collapses the intervals estimated in the bound step into point estimates employing a convex combination of the extreme estimates. This convex combination can be determined either by using external information about the pattern of missing data or by a dynamic estimation of this pattern from the available data.

Assume that some external information is available on the pattern of missing data. One can encode this information as a probability distribution defining, for each datum in the database, the probability of the datum being missing as

$$p(x_{ik}|\pi_{ij}, X_i = ?) = \phi_{ijk}$$

Where $k = 1, \cdots, c_i$, the number of state in X_i is denoted by c_i and $\sum_k \phi_{ijk} = 1$.The notation $X_i = ?$ denotes that the state of X_i is missing. The probabilities ϕ_{ijk} can be employed to determine accurate estimates of θ_{ijk}, which is the probability of X_i being in the kth state given the parent states π_{ij}. A single probability for each state of the variable X_i given the parent states π_{ij} as

$$p_{k\cdot}(x_{il}|\pi_{ij}, D) = \frac{\alpha_{il} + n(x_{il}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij}) + n^{\cdot}(x_{ik}|\pi_{ij})} \tag{21}$$

for $l \neq k$. Therefore, the local minimum of $E(\theta_{ijk}|D)$ can be calculated as

$$p^{l}_{\cdot}(x_{ik}|\pi_{ij}, D) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij}) + max_{h\neq k} n^{\cdot}(x_{ik}|\pi_{ij})} \tag{22}$$

Which shows that the difference between $p_{k\cdot}(x_{il}|\pi_{ij}, D)$ and $p^{l}_{\cdot}(x_{ik}|\pi_{ij}, D)$ depends only on the cases in which the state of the child variable is known and the parent configuration is not.

The distribution of missing entries in terms of ϕ_{ijk} can be employed to identify a point estimate within the interval $[p^{l}_{\cdot}(x_{ik}|\pi_{ij}, D), p^{\cdot}(x_{ik}|\pi_{ij}, D)]$via convex combination of extreme probabilities:

$$\hat{p}(x_{ik}|\pi_{ij}, D, \phi_{ijk}) = \sum_{l\neq k} \phi_{ijk} p^{l}_{\cdot}(x_{ik}|\pi_{ij}, D) + \phi_{ijk} p^{\cdot}(x_{ik}|\pi_{ij}, D) \tag{23}$$

Finally, if data are missing only on the child variable ($n^{\cdot}(x_{ik}|\pi_{ij}) = n^{\cdot}_{ij}$), then we yet

$$\hat{p}(x_{ik}|\pi_{ij}, D, \phi_{ijk}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij}) + n^{\cdot}_{ij}\phi_{ijk}}{\alpha_{ij} + n(\pi_{ij}) + n^{\cdot}_{ij}} \tag{24}$$

so that the incomplete cases are distributed across the states of X according to the prior knowledge on the pattern of missing data. Note that Equation (24) is the expected Bayesian estimate given the assumed pattern of missing data.

If there is no external information about the pattern of missing data, the BC method works similar to EM and GS methods due to the use of the pattern of the available data. In this case, $\phi_{ijk} = p(x_{ik}|\pi_{ij})$ and it can be estimated from the available data as

$$\hat{\phi}_{ijk} = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij})} \tag{25}$$

This estimate can then be employed to compute the convex combination of the extreme probabilities. The estimate of $p(x_{ik}|\pi_{ij}, D)$ can be computed as

$$\hat{p}(x_{ik}|\pi_{ij}, D) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij}) + n^{\cdot}_{ij}\phi_{ijk}}{\alpha_{ij} + n(\pi_{ij}) + n^{\cdot}_{ij}} = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + n(\pi_{ij})} \tag{26}$$

which is a consistent estimate of θ_{ijk} since $\hat{p}(x_{ik}|\pi_{ij}, D)$ is a generalized version of the Maximum Likelihood Estimate of θ_{ijk}. If $\alpha_{ijk} = 0$, then the BC estimate becomes the classical MLE of θ_{ijk}. Clearly, the estimates of the conditional probabilities computed by Equation (26) are the expected estimates and, as the database increases, they will be the same estimates computed by GS.

Sebastiani and Romani compared the accuracy and the efficiency of EM, GS, and methods. They found that both EM and GS provide reliable estimates of the parameters and they are currently regarded as the most viable solutions to the missing data. On the other hand, both these iterative methods can be trapped into local minima and the convergence detection can be difficult. Furthermore, they assume that the missing data mechanism is ignorable; i.e., within each observed parent configuration, the available data is a representative sample of the complete database and the distribution of missing data can therefore be inferred from the available entries [20].When this assumption fails, and the missing data mechanism is hot ignorable (NI), the accuracy of these methods can drastically decrease. Additionally,

Sabestiani and Romani state that the computational cost of these methods depends mainly on the absolute number of missing data, and this dependency can prevent their scalability to large databases.

The most important characteristic of BC is its ability to represent the pattern of available data and the assumed pattern of missing data explicitly and separately. The BC algorithm provides probability intervals that can make the analyst aware of the range of possible estimates, and hence of the quality of information on which inference is based. The probability intervals used by BC provide a specific measure of the quality of information conveyed by the database and explicit representation of the impact of the assumption made on the pattern of missing data. Therefore, BC does not depend on the ignorability assumption. Furthermore, BC reduces the cost of estimating each conditional distribution of each variable X_i to the cost of one exact Bayesian updating and one convex combination for each state of X_i in each parent configuration. This deterministic process does not decrease the convergence rate and the convergence detection relative to stochastic processes. Additionally, BC the method's computational complexity is independent of the number of missing data.

Consequently, the BC algorithm gives almost the same results as EM and GS when the missing data is ignorable but it gives better results when the missing data mechanism is not ignorable. The convergence rate of BC is also better than EM and GS. Thus, BC learns the network faster than EM and GS methods. The experimental comparison with EM and GS proves that a substantial equivalence of the estimates provided by these three methods and a dramatic gain in efficiency using BC.

Ramoni and Sebastiani claimed the estimates provided by BC are more robust to departure of the data from the true pattern of missing data. The computational cost of BC is equal to the cost of two exact Bayesian updates-one for each extreme distribution-plus the cost of a convex combination for each parameter in the BN [21].

One may ask what happens if the network structure is unknown in addition to partially observable data. There is no easy answer to this question given in the literature. Some possibilities are explored in the next section.

2.4 Unknown Network Structure and Unobservable Variables

This is the most difficult case to resolve because the structure of the networks is unknown and the variables are not fully observable. There is no significant amount of research for this case. There are two recently developed methods that recover the Bayesian network structure with unobserved variables.

The first algorithm was proposed by Russell [22] and is called structural EM (SEM) algorithm. The algorithm combines the standard EM algorithm, which optimizes the network parameters, with structure search for model selection. The main idea of this method is that it attempts to maximize the expected score of models instead of their actual scores at iteration. Russell proves a theorem that the SEM algorithm makes progress in iteration on finding the better scoring network. Then, he states that if one chooses a model that maximizes the expected score at iteration, then a better choice is provably made in terms of the marginal score of the network. The SEM algorithm is exciting since it attempts to directly optimize the true Bayesian score within EM iteration rather than an asymptotic approximation.

The most problematic aspect of SEM is that it might converge to a sub-optimal model. This could happen if the model generates a distribution that causes other models to appear worse when the expected score is examined. This difficulty becomes more obvious when the ratio of missing information is higher. Russell suggests that, in practice, the algorithm needs to be run from several starting points to get a better estimate of the MAP model. Another restriction of the SEM is that it focuses on learning a single model. In practice, several high scoring models is necessary for better prediction. Additional to this deficiency, the algorithm requires large number of computations during learning. This is the main problem in applying this technique to large-scale domains. The following paragraphs provide a computationally cheaper method.

The second algorithm was proposed by Sebastiani and Marino. They were able to show that BC algorithm could also learn the structure of the network with small changes in the algorithm. This method is very similar to the search method which we had fully observed data. The only difference is that, in this case, we have partially observed data or incomplete data. Therefore, the estimation of the parameters of the network is also necessary. The BC method is employed to estimate the parameters of the network. The estimation process is performed in each step, i.e., after adding each edge to the network. Consequently, the method involves both parameter learning and structure learning. However, the main attention was given to the parameter estimation part since it is newly discovered method. The structure learning part can be modified as a greedy search algorithm. In that case, "delete an edge" operator and "reverse an edge" operator have to be incorporated to the algorithm.

There is a slight difference between SEM and BC methods and the problem of self- organizing agents in terms of required data structure. The SEM and BC algorithms require a certain minimum length database. Unfortunately, there will not be a prior database to work with at the beginning of the agents' exploration of the environment. Thus our learning method has to be online: estimation of the network structure and parameters will be performed simultaneously with the gathering of new entries in the database. So, our method has to learn the network while the agents are exploring environment and organizing themselves to manage a common task.

3 Conclusions

The process of learning Bayesian networks takes different forms in terms of whether the structure of the network is known and whether the variables are all observable. The structure of the network can be known or unknown, and the variables can be expressed as complete and incomplete data. In this paper, we introduce two cases of learning Bayesian networks from complete data: known structure and observable variables, unknown structure and observable variables. Next, we study two cases of learning Bayesian networks from incomplete data: known network structure and unobservable variables, unknown network structure and unobservable variables.

Acknowledgements

This work is financially supported by the National Natural Science Foundation of China (Project No. 90718038). Thanks for the help.

References

[1] Singh, Jagdip, "Performance Productivity and Quality of Frontline Employees in Service Organization," Journal of Marketing, **64**, 15-34 (2000).
[2] Jong, Ad de, Ko de Ruyter, and Jos Lemmink, "Antecedents and Consequences of the Service Climate in Boundary-Spanning Self-Managing Service Team" .Journal of Marketing, **68**, 18-35 (2004).
[3] Louis, Meryl R. and Robert I. Sutton, "Switching Cognitive Gears: From Habits of Mind to Active Thinking," Human Relations, **44**, 55-76 (1991).
[4] Argyris, C. and D. A. Schon, Organizational Learning: Theory, Method, and Practice. Reading, MA: Addison-Wesley, **2**, (1996).
[5] Friedman, Victor J., "The Individual as Agent of Organizational Learning," in Handbook of Organizational Learning and Knowledge, Meinolf Dierkes and Ariane Berthoin Antal and John Child and Ikujiro Nonaka, Eds. Oxford: University Press (2001).
[6] Tyre, M. J. and W.J. Orlikowski, "Windows of Opportunity: Temporal Patterns of Technological Adaptation in Organizations," Organization Science, **5**, 98-118 (1994).
[7] Edmondson, Amy, "Psychological Safety and Learning Behavior in Work Teams Administrative Science Quarterly, **44**, 350-383 (1999).
[8] Zeithaml, Valarie A. and Mary Jo Bitner, Services Marketing. New York: McGraw Hill (1996).
[9] S. L. Lauritzen, "The EM algorithm for graphical association models with missing data," Technical Report TR-91-05, Department of Statistics, Aalborg University, (1991).
[10] S. L. Lauritzen, "The EM algorithm for graphical association models with missing data," Computational Statistics and Data Analysis, **19**, 191-201 (1995).
[11] K. G. Olesen, S. L. Lauritzen and F. V. Jensen, "aHUGIN: A system for creating adaptive causal probabilistic networks," in Proceedings of the Eighth Conference on Uncertainty in AI (UAI '92), Stanford, CA: Morgan Kaufmann, (1992).
[12] D. J. Spiegelhalter and R.G. Cowell, "Learning in probabilistic expert systems," in J.M. Bernardo, J. O. Bergen A. P. Dawid, and A. F. M. Smith (Eds.), Bayesian Statistics, **4**, (1992).
[13] 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).
[14] D. Heckerman, "A tutorial on learning Bayesian networks," Technical Report MSR-TR-95-06, Microsoft Research, (1995).
[15] M. Ramoni and P. Sebastiani, "Learning Bayesian networks from incomplete data," Technical Report KMi-TR-43, Knowledge Median Institute, The Open University, February (1997).
[16] M. Ramoni and P. Sebastiani, "Discovering Bayesian networks in incomplete databases," Technical Report KMi-TR-46, Knowledge Median Institute, The Open University, March (1997).
[17] P. Sebastiani and M. Ramoni," Bayesian inference with missing data using bound and collapses," Technical Report KMi-TR-58, Knowledge Median Institute, The Open University, November (1997).
[18] M. Ramoni and P. Sebastiani, "Learning conditional probabilities from incomplete data: An experimental comparison," Technical Report KMi-TR-64 Knowledge Median Institute, The Open University, July (1998).
[19] M. Ramoni and P. Sebastiani, "Parameter estimation in Bayesian networks from incomplete databases," Technical Report KMi-TR-57, Knowledge Median Institute, The Open University, November (1997).
[20] R. J. A. Little and D. B. Rubin, Statistical Analysis with Data, Wiley, New York, (1987).
[21] M. Ramoni and P. Sebastiani, "Learning Bayesian networks from incomplete data," Technical Report KMi-TR-43, Knowledge Median Institute, The Open University, February (1997).
[22] N. Friedman, "The Bayesian structural EM algorithm," in G.F. Cooper and S. Moral (Eds.), Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI 98), San Francisco, CA: Morgan Kaufmann, (1998).

Yonghui Cao received the MS degree in business management from Zhejiang University in 2006. He is currently a doctorate candidate in Zhejiang University. His research interest is in the areas of management information systems.