Evolution of parameters in Bayesian Overlay Model

Loc Nguyen[1], **Phung Do**[2]
[1]Faculty of Information Technology, The University of Natural Science, Ho Chi Minh city, Vietnam
[2]Faculty of Information System, The University of Information Technology, Ho Chi Minh city, Vietnam

Abstract - *Adaptive learning systems require well-organized user model along with solid inference mechanism. Overlay modeling is the method in which the domain is decomposed into a set of elements and the user model is simply a set of masteries over those elements. The combination between overlay model and Bayesian network (BN) will make use of the flexibility and simplification of overlay modeling and the power inference of BN. However, it is compulsory to pre-define parameters, namely, Conditional Probability Tables (CPTs) in Bayesian network but no one ensured absolutely the correctness of these CPTs. This paper discusses about how to enhance parameters' quality in Bayesian overlay model, in other words, this is the evolution of CPTs.*

1. Introduction

User model is the core of almost adaptive learning system. There are some effective modeling methods, such as: stereotype, overlay, plan recognition but overlay model is proved soundness due to two its properties: flexible graphic structure and reflecting comprehensibly the domain knowledge in education course. The basic ideology of overlay model is to represent user knowledge as subset of domain model. Bayesian network (BN) is the directed acyclic graph (DAG) in which the nodes are linked together by arcs, each arc expresses the dependence relationships between nodes. The strengths of dependences are quantified by Conditional Probability Table (CPT). BN have excellent inference mechanism based on Bayesian law. The combination between overlay model and BN [Nguyen 2008] will make use of each method's strong points and restraints drawbacks.

- The structure of overlay model is translated into BN, each user knowledge element becomes an node in BN.
- Each prerequisite relationship between domain element in overlay model becomes an conditional dependence assertion signified by CPT of each node in BN.
- Domain elements are defined as hidden nodes and other learning objects which are used to assess user's performance are consider as evidence nodes in BN.

(See figure 7 and table 3)

In process of parameter specification by weighting arcs, the gained CPTs are confident but it is necessary to improve them after inference tasks from collected evidences. This trend relates to learning parameters, that's to say, the evolution of CPTs.

Section 2: related work in this BN and overlay model

Section 3: learning parameters in Bayesian model

Section 4: specifying parameters and their evolution

2. Related work

Some systems applied successfully BN to build up user models but most of them don't have mechanism for the evolution of BN:

- HYDRIVE [2] models a student's competence at troubleshooting an aircraft hydraulics system. Student's knowledge is characterized in terms of general constructs (dimensional variables). BN is used to update these student model dimensional variables, using as evidence student's actions.
- ANDES [3] is Intelligent Tutoring System that teaches Newtonian Physics via coached problem solving and uses BN to do long-terms knowledge assessment, plan recognition.
- KBS Hyperbook [4], a very success student guidance in e-learning, models students by classifying them according to their knowledge level K into the categories: novice (N), beginner (B), intermediate (I) and advanced (A). Of course, each student's knowledge item is represented by node Ki taking values {N, B, I, A}. All nodes Ki and their conditional dependences constitute the BN.

3. Learning parameters in Bayesian model

3.1. Learning parameters

Dummy variables and augmented BN

In continuous case, the CPT of each node is replaced by the probability density function (PDF). There is a family of PDF which quantifies and updates the strength of conditional dependencies between nodes by natural way is called beta density function, denoted as *β(f; a, b)* or *Beta(f; a, b)* with parameters *a, b, N=a+b* where *a, b* should be integer number > 0.

$$\beta(f) = \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} f^{a-1}(1-f)^{b-1} \quad (1)$$

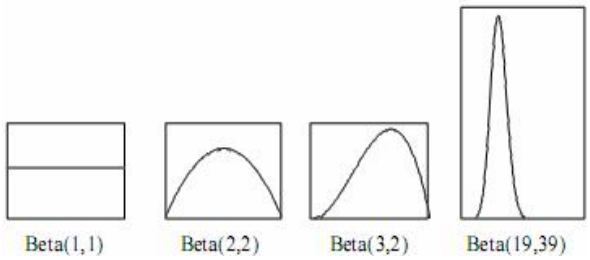

Fig. 1 Beta functions

It means that, there are *"a"* successful outcomes (for example, *f =1*) in *"a+b"* trials. Higher value of *"a"* is, higher ratio of success is, so, the graph leans forward right. Higher value of *"a+b"* is, the more the mass is concentrate around *a/(a+b)* and the more narrow the graph is. Definition of beta function is based on gamma function described below:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

The integral will converges if *x>0*, at that time, $\Gamma(x) = (x-1)!$. Of course, we have $\frac{\Gamma(x+1)}{\Gamma(x)} = x$ (2).

From formula 1, $\int_0^1 f^a (1-f)^b df = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$ (3)

Proof,

$$\int_0^1 f^a (1-f)^b df = \int_0^1 \frac{\Gamma(a+1+b+1)}{\Gamma(a+1)\Gamma(b+1)} f^a (1-f)^b \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+1+b+1)} df$$

$$= \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} \int_0^1 \beta(f; a+1, b+1) df \quad \text{(due to formula 1)}$$

$$= \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} \Pr(0 \le f \le 1) = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$$

(due to $\Pr(0 \le f \le 1) = 1$)

Suppose there is one binary variable *X* in network and the probability distribution of *X* is considered as relative frequency having values in [0, 1] which is the range of variable *F*. We add a dummy variable *F* (whose space consists of numbers in [0, 1], of course) which acts as the parent of *X* and has a beta density function *β(f; a, b)*, so as to:

Pr(X=1|f) = f, where *f* denotes values of *F*

X and *F* constitute a simple network which is referred as augmented BN. So *X* is referred as real variable (hypothesis) opposite to dummy variable.

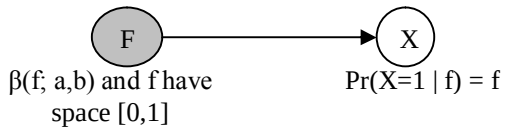

Fig. 2 The simple augmented BN with only one hypothesis node *X*

Obviously, *P(X=1) = E(F)* where *E(F)* is the expectation of *F*

Proof, owing to the law of total probability

$$\Pr(X=1) = \int_0^1 \Pr(X=1|f)\beta(f)df = \int_0^1 f\beta(f)df = E(F)$$

Due to F is beta function, $E(F) = \frac{a}{N}$, so, $\Pr(X=1) = \frac{a}{N}$ (4)

Proof.

$$E(F) = \int_0^1 f\beta(f)df = \int_0^1 f \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} f^{a-1}(1-f)^{b-1} df$$

$$= \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} \int_0^1 f^a (1-f)^{b-1} = \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(N+1)} \quad \text{(due to formula 3)}$$

$$= \frac{a}{N} \quad \text{(applying formula 2)}$$

The ultimate purpose of Bayesian inference is to consolidate a hypothesis (namely, variable) by collecting evidences. Suppose we perform *M* trials of a random process, the outcome of u^{th} trial is denoted $X^{(u)}$ considered as evidence variable whose probability $Pr(X^{(u)} = 1 \mid f) = f$. So, all $X^{(u)}$ are conditionally dependent on *F*. The probability of variable *X*, *Pr(X=1)* is learned by these evidences.

We denote the vector of all evidences as $E = (X^{(1)}, X^{(2)}, \ldots, X^{(M)})$ which is also called the sample of size *M*. Given this sample, *β(f)* is called the prior density function, and $Pr(X^{(u)} = 1) = a/N$ (due to formula 1) is called prior probability of $X^{(u)}$. *It is necessary to determine the posterior density function β(f|E) and the posterior probability of X, namely Pr(X|E). The nature of this process is the parameters learning.* Note that *Pr(X|E)* is referred as *Pr(X(M+1) | E)*.

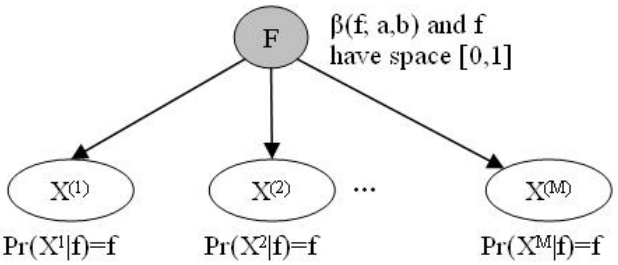

Fig. 3 The sample $E=(X^{(1)}, X^{(2)}, \ldots, X^{(M)})$ size of M

We only surveyed in the case of binomial sample, in other words, *E* having binomial distribution is called binomial

sample and the network in figure 3 becomes a binomial augmented BN. Then, suppose *s* is the number of all evidences $X^{(i)}$ which have value *1* (success), otherwise, *t* is the number of all evidences $X^{(i)}$which have value *0* (failed). Of course, *s + t = M*.

Owing the law of total probability, we have

$$E(f^s(1-f)^t) = \int_0^1 f^s(1-f)^t \beta(f)df$$

$$= \int_0^1 f^s(1-f)^t \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} f^{a-1}(1-f)^{b-1}df \quad \text{(applying formula 1)}$$

$$= \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} \int_0^1 f^{a+s-1}(1-f)^{b+t-1}df$$

$$= \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+s)\Gamma(b+t)}{\Gamma(a+b+s+t)} \quad \text{(due to formula 3)}$$

$$= \frac{\Gamma(\mathrm{N})}{\Gamma(\mathrm{N}+\mathrm{M})} \frac{\Gamma(a+s)\Gamma(b+t)}{\Gamma(a)\Gamma(b)} \quad (\text{due to } \mathrm{s}+\mathrm{t} = M)$$

(5)

And,

$$\Pr(E) = \int_0^1 \Pr(E|f)\beta(f)df = \int_0^1 \prod_{i=1}^{M} \Pr(X^i \mid f)\beta(f)df$$

$$= \int_0^1 f^s(1-f)^s \beta(f)df = E(f^s(1-f)^t), \text{ due to } \prod_{i=1}^{M} \Pr(X^i \mid f) = f^s(1-f)^t$$

(6)

Computing posterior density function

Now, we need to compute the posterior density function *β(f|E)* and the posterior probability *Pr(X=1|E)*. It is essential to determine the probability distribution of *X*.

$$\beta(f \mid E) = \frac{\Pr(E \mid f)\beta(f)}{\Pr(E)} \quad \text{(Bayes' law)}$$

$$= \frac{f^s(1-f)^t\beta(f)}{E(f^s(1-f)^t)} \quad (\text{due to } \Pr(\mathrm{E} \mid \mathrm{f}) = \prod_{i=1}^{M} \Pr(X^i \mid f) = f^s(1-f)^t \text{ and apply formua 6})$$

$$= \frac{f^s(1-f)^t \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} f^{a-1}(1-f)^{b-1}}{\frac{\Gamma(N)}{\Gamma(N+M)} \frac{\Gamma(a+s)\Gamma(b+t)}{\Gamma(a)\Gamma(b)}} \quad \text{(apply formula 1, 5)}$$

$$= \frac{\Gamma(\mathrm{N}+\mathrm{M})}{\Gamma(\mathrm{a}+\mathrm{s})\Gamma(\mathrm{b}+\mathrm{t})} f^{a+s-1}(1-f)^{b+t-1} = \beta(f; a+s, b+t)$$

(7)

Then the posterior density function is *β(f; a+s, b+t)* where the prior density function is *β(f; a, b)*. According to formula 4, the posterior probability

$$Pr(X=1|E) = E(\beta(f|E)) = \frac{a+s}{a+s+b+t} = \frac{a+s}{N+M} \quad (8)$$

In general, you should merely engrave the formula 1, 4, 7, 8 and the way to recognize prior density function, prior probability of *X* and posterior density function, posterior probability of *X*, respectively on your memory.

3.2. Expanding augmented BN with more than one hypothesis node

Suppose we have a BN with two binary random variables and there is conditional dependence assertion between these nodes. See the network and CPTs in the figure below

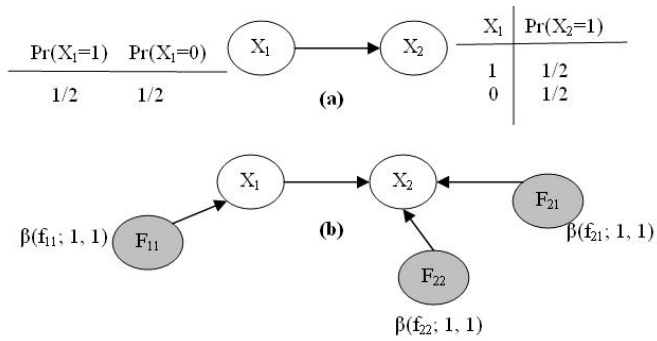

Fig. 4 BN (a) and expended augmented BN (b)

For every node (variable) X_i, we add dummy parent nodes to X_i, obeying two ways below:

- If X_i has no parent (not conditionally dependent on any others), we add only one dummy variable denoted F_{i1} having the probability density function $\beta(f_{i1}; a_{i1}, b_{i1})$ so as to: $Pr(X_i=1|f_{i1})= f_{i1}$.
- If Xi has a set of k_i parents and each parent pa_{il} ($l=\overline{1,k_i}$) is binary, we add a set of $c_i=2k_i$ dummy variables $F_i = \{f_{i1}, f_{i2},\ldots, f_{ic_i}\}$, in turn, instantiations of parents $PA_i= \{pa_{i1}, pa_{i2}, pa_{i3},\ldots, pa_{ic_i}\}$. In other words, c_i denotes the number of instantiations of the parents PA_i. We have $Pr(X_i=1|pa_{ij}, f_{i1},\ldots, f_{ij},\ldots, f_{ic_i})=f_{ij}$. where $\beta(f_{ij}) = \frac{\Gamma(N_{ij})}{\Gamma(a_{ij})\Gamma(b_{ij})} f^{a_{ij}-1}(1-f_{ij})^{b_{ij}-1}$

All f_{ij} have no parent and are mutually independent, so, $\beta(f_{i1}, f_{i2},\ldots, f_{ic_i}) = \beta(f_{i1})\ \beta(f_{i2})\ldots\ \beta(f_{ic_i})$. Besides this local parameter independence, we have the global paramter independece if reviewing all variables *Xi*, such below:

$$\beta(F_1, F_2,\ldots, F_n)= \beta(f_{11}, f_{12},\ldots, f_{ic_{in}})= \beta(f_{i1})\ \beta(f_{i2})\ldots\ \beta(f_{ic_{in}})$$

All variables *Xi* and their dummy variables form the expended augmented BN representing the trust BN in figure 4. In the trust BN, the conditional probability of variable *Xi* with the instantiation of its parent pa_{ij}, in other words, the ij^{th} conditional distribution is given by $Pr(X_i=1|\ pa_{ij}=1) = E(F_{ij})= \frac{a_{ij}}{N_{ij}}$, that's to say the expected value of F_{ij} (8).

Proof,

$$\Pr(X_i = 1 \mid pa_{ij} = 1) = \int_0^1 \dots \int_0^1 \Pr(X_i = 1 \mid pa_{ij} = 1, f_{i1}, \dots, f_{ic_i}) \beta(f_{i1}) \dots \beta(f_{ic_i}) df_{i1} \dots df_{ic_i}$$

$$= \int_0^1 \dots \int_0^1 f_{ij} \beta(f_{i1}) \dots \beta(f_{ic_i}) df_{i1} \dots df_{ic_i} = E(F_{ij})$$

(due to F_{ij}s are mutually independent, $\Pr(X_i = 1 \mid pa_{ih_i} = 1, f_{i1}, \dots, f_{ic_i}) = \Pr(X_i = 1 \mid pa_{ij} = 1, f_{ij}) = f_{ij}$)

Suppose we perform M trials of random process, the outcome of i^{th} trial which is BN like figure 4 is represented as a random vector $X^{(u)} = \begin{pmatrix} X_1^{(u)} \\ \dots \\ X_n^{(u)} \end{pmatrix}$ containing all hypothesis variables in network. $X^{(u)}$ is also called evidence vector (or evidence, briefly). M trials constitute the sample of size M which is the set of random vectors denoted as $E=\{X^{(1)}, X^{(2)}, \dots, X^{(M)}\}$. E is also called evidence matrix. We review only in case of binomial sample, it means that E is the binomial BN sample of size M. For example, this sample corresponding to the network in figure 4 is showed below:

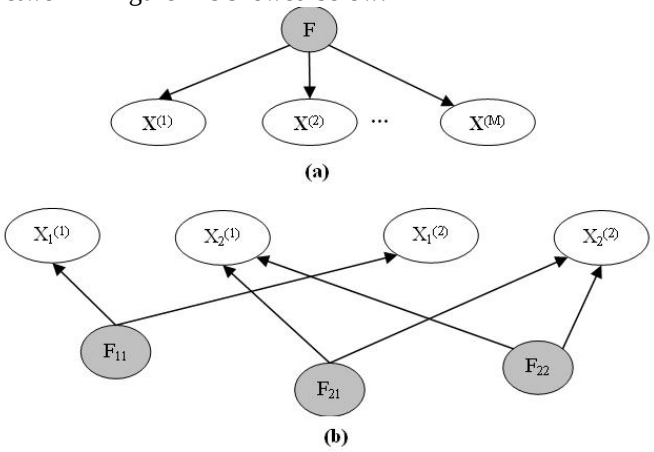

Fig. 5 Expanded binomial BN sample of size M

After occurring M trial, the augmented BN was updated and dummy variables' density functions and hypothesis variables' conditional probabilities changed. We need to compute the posterior density function $\beta(f_{ij}|E)$ of each dummy variable F_{ij} and the posterior condition probability $Pr(X_i=1 \mid pa_{ij} = 1, E)$ of each variable X_i. Note that the samples $X^{(u)}$ s are mutually independent with all given F_{ij}. We have,

$$\prod_{u=1}^{M} \Pr(X_i^{(u)} \mid pa_i, F_i) = \prod_{j=1}^{c_i} (f_{ij})^{s_{ij}} (1 - f_{ij})^{t_{ij}}$$

where

- c_i is the number of instances of $X_i^{(u)}$'s parents. In binary case, each $X_i^{(u)}$'s parent has two instances/values, namely, 0 and 1.
- s_{ij}, respective to f_{ij}, is the number of all evidences that variable $X_i = 1$ and $pa_{ij} = 1$
- t_{ij}, respective to f_{ij}, is the number of all evidences that variable $X_i = 1$ and $pa_{ij} = 0$.

We have,

$$\Pr(E \mid F_1, \dots, F_n) = \prod_{i=1}^{n} \prod_{u=1}^{M} \Pr(X_i^{(u)} \mid pa_i, F_i) = \prod_{i=1}^{n} \prod_{j=1}^{c_i} (f_{ij})^{s_{ij}} (1 - f_{ij})^{t_{ij}} \quad (9)$$

$$\Pr(E) = \prod_{i=1}^{n} \prod_{j=1}^{c_i} E(f_{ij}^{s_{ij}} (1 - f_{ij})^{t_{ij}}) \quad (10)$$

Proof,

$$\Pr(E) = \prod_{i=1}^{n} \left(\int_{F_i} \prod_{u=1}^{M} \Pr(X_i^{(u)} \mid pa_i, F_i) \beta(F_i) dF_i \right)$$

(due to the law of total probability and the joint probability distribution)

$$= \prod_{i=1}^{n} \left(\int_{F_i} \prod_{j=1}^{c_i} (f_{ij})^{s_{ij}} (1 - f_{ij})^{t_{ij}} \beta(F_i) dF_i \right)$$

(applying formula $\prod_{u=1}^{M} \Pr(X_i^{(u)} \mid pa_i, F_i) = \prod_{j=1}^{c_i} (f_{ij})^{s_{ij}} (1 - f_{ij})^{t_{ij}}$)

$$= \prod_{i=1}^{n} \prod_{j=1}^{c_i} \int_0^1 (f_{ij})^{s_{ij}} (1 - f)^{t_{ij}} \beta(f_{ij}) df_{ij}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{c_i} E(f_{ij}^{s_{ij}} (1 - f_{ij})^{t_{ij}})$$

There is the question "how to determine $E(f_{ij}^{s_{ij}} (1 - f_{ij})^{t_{ij}})$". Applying formula 5, we have:

$$E(f_{ij}^{s_{ij}} (1 - f_{ij}^{t_{ij}})) = \frac{\Gamma(N_{ij})}{\Gamma(N_{ij} + M_{ij})} \frac{\Gamma(a_{ij} + s_{ij}) \Gamma(b_{ij} + t_{ij})}{\Gamma(a_{ij}) \Gamma(b_{ij})} \quad \text{where } N_{ij} = a_{ij} + b_{ij} \text{ and } M_{ij} = s_{ij} + t_{ij} \quad (11)$$

Updating posterior density function $\beta(f_{ij}|E)$

$$\beta(f_{ij} \mid E) = \frac{(f_{ij})^{s_{ij}} (1 - f_{ij})^{t_{ij}} \beta(f_{ij})}{E(f_{ij}^{s_{ij}} (1 - f_{ij})^{t_{ij}})} = beta(f_{ij}; a_{ij} + s_{ij}, b_{ij} + t_{ij}) \quad (12)$$

Proof,

$$\beta(f_{mn} \mid E) = \frac{\Pr(E \mid f_{mn}) \beta(f_{mn})}{\Pr(E)} \quad \text{(Bayes' law)}$$

$$= \frac{\left(\int_0^1 \dots \int_0^1 \Pr(E \mid F_1, F_2, \dots, F_n) \prod_{ij \neq mn} \beta(f_{ij}) df_{ij} \right) \beta(f_{mn})}{\Pr(E)} \quad \text{(law of total probability)}$$

$$= \frac{(f_{mn})^{s_{mn}} (1 - f_{mn})^{t_{mn}} \left(\prod_{ij \neq mn} \int_0^1 (f_{ij})^{s_{ij}} (1 - f_{ij})^{t_{ij}} \beta(f_{ij}) df_{ij} \right) \beta(f_{mn})}{\prod_{i=1}^{n} \prod_{j=1}^{c_i} E(f_{ij}^{s_{ij}} (1 - f_{ij})^{t_{ij}})}$$

(apply formula 9, 10)

$$= \frac{(f_{mn})^{s_{mn}} (1 - f_{mn})^{t_{mn}} \beta(f_{mn})}{E(f_{mn}^{s_{mn}} (1 - f_{mn})^{t_{mn}})}$$

$$= \frac{(f_{mn})^{s_{mn}} (1 - f_{mn})^{t_{mn}} \frac{\Gamma(N_{mn})}{\Gamma(a_{mn}) \Gamma(b_{mn})} (f_{mn})^{a_{mn}-1} (1 - f_{mn})^{b_{mn}-1}}{\frac{\Gamma(N_{mn})}{\Gamma(N_{mn} + M_{mn})} \frac{\Gamma(a_{mn} + s_{mn}) \Gamma(b_{mn} + t_{mn})}{\Gamma(a_{mn}) \Gamma(b_{mn})}}$$

(expansion of $\beta(f_{mn})$ and applying formula 11 to $E(f_{mn}^{s_{mn}} (1 - f_{mn})^{t_{mn}})$)

$$= \frac{\Gamma(N_{mn} + M_{mn})}{\Gamma(a_{mn} + s_{mn}) \Gamma(b_{mn} + t_{mn})} (f_{mn})^{a_{mn}+s_{mn}-1} (1 - f_{mn})^{b_{mn}+t_{mn}-1}$$

$$= beta(f_{mn}; a_{mn} + s_{mn}, b_{mn} + t_{mn})$$

According to formula 8 and 12,

$Pr(X_i=1|\ pa_{ij}=1,\ E) = E(F_{ij}) = E(\beta(f_{ij}|E) = \frac{a_{ij}+s_{ij}}{a_{ij}+s_{ij}+b_{ij}+t_{ij}} = \frac{a_{ij}+s_{ij}}{N_{ij}+M_{ij}}$ (13)

In short, in case of binomial distribution, if we have the real/trust BN embedded in the expanded augmented network such as figure 4 and each dummy node F_{ij} has a prior beta distribution $\beta(f_{ij};\ a_{ij},\ b_{ij})$ and each hypothesis node X_i has the prior conditional probability $Pr(X_i=1|\ pa_{ij}=1)=E(\beta(f_{ij}))=\frac{a_{ij}}{N_{ij}}$, the parameter learning process based on a set of evidences is to update the posterior density function $\beta(f_{ij}|E)$ and the posterior conditional probability $Pr(X_i=1|\ pa_{ij}=1,\ E)$. Indeed, $\beta(f_{ij}\mid E) = beta(f_{ij}; a_{ij}+s_{ij}, b_{ij}+t_{ij})$ and $Pr(X_i=1|\ pa_{ih_i}=1,\ E) = E(\beta(f_{ij}|E)) = \frac{a_{ij}+s_{ij}}{N_{ij}+M_{ij}}$

Example

Suppose we have the set of 5 evidences $E=\{X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}, X^{(5)}\}$ owing to network in figure 4

Table 1: Set of evidences E corresponding to 5 trials (sample of size 5)

	$\mathbf{x_1}$	$\mathbf{x_2}$
$\mathbf{X^{(1)}}$	$X_1^{(1)}=1$	$X_2^{(1)}=1$
$\mathbf{X^{(2)}}$	$X_1^{(2)}=1$	$X_2^{(2)}=1$
$\mathbf{X^{(3)}}$	$X_1^{(3)}=1$	$X_2^{(3)}=1$
$\mathbf{X^{(4)}}$	$X_1^{(4)}=1$	$X_2^{(4)}=0$
$\mathbf{X^{(5)}}$	$X_1^{(5)}=0$	$X_2^{(5)}=0$

Note that the first evidence $X^{(1)}=\begin{pmatrix} X_1^{(1)}=1 \\ X_2^1=1 \end{pmatrix}$ implies that variable $X_2=1$ given $X_1=1$ occurs in the first trial. We need to compute all posterior density functions $\beta(f_{11}|E),\ \beta(f_{21}|E),\ \beta(f_{22}|E)$ and all conditional probabilities $Pr(X_1=1),\ Pr(X_2=1|X_1=1),\ Pr(X_2=1|X_1=0)$ from prior density functions $\beta(f_{11};\ 1,1),\ \beta(f_{21};\ 1,1),\ \beta(f_{22};\ 1,1)$. In fact,

$s_{11}=1+1+1+1+0=4 \quad t_{11}=0+0+0+0+1=1$
$s_{21}=1+1+1+0+0=3 \quad t_{21}=0+0+0+0+1=1$
$s_{22}=0+0+0+0+0=0 \quad t_{21}=0+0+0+0+1=1$

$\beta(f_{11}|E) = \beta(f_{11};a_{11}+s_{11},\ b_{11}+t_{11})=\beta(f_{11};\ 1+4,\ 1+1)=\beta(f_{11};\ 5,\ 2)$
$\beta(f_{21}|E) = \beta(f_{21};a_{21}+s_{21},\ b_{21}+t_{21})=\beta(f_{21};\ 1+3,\ 1+1)=\beta(f_{11};\ 4,\ 2)$
$\beta(f_{22}|E) = \beta(f_{22};a_{22}+s_{22},\ b_{22}+t_{22})=\beta(f_{22};\ 1+0,\ 1+1)=\beta(f_{11};\ 1,\ 2)$

and $Pr(X_1=1),\ Pr(X_2=1|X_1=1),\ Pr(X_2=1|X_1=0)$ are expectations of $\beta(f_{11}|E),\ \beta(f_{21}|E),\ \beta(f_{22}|E)$. Then,

$Pr(X_1=1)=\frac{5}{5+2}=\frac{5}{7}$ $\quad Pr(X_2=1|X_1=1)=\frac{4}{4+2}=\frac{2}{3}$ $\quad Pr(X_2=1|X_1=0)=\frac{1}{1+2}=\frac{1}{3}$

Network in figure 4 changed as follows:

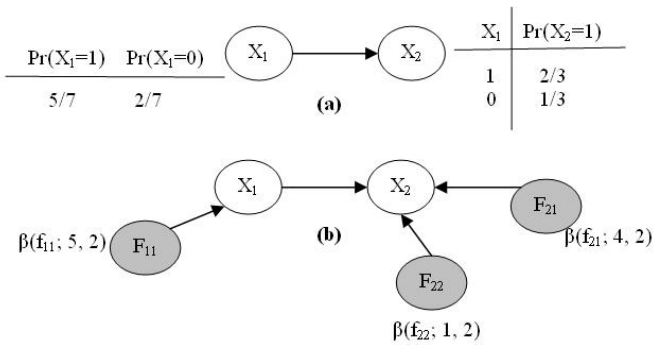

Fig. 6 Updated version of BN (a) and augmented BN (b) in fig. 4

3.3. Learning parameters in case of data missing

In practice there are some evidences in E such as $X^{(u)}$ s which lack information and thus, it stimulates the question "How to update network from data missing". We must address this problem by artificial intelligence techniques, namely, expectation maximization (EM) algorithm – a famous technique solving estimation of data missing. Like above example, we have the set of 5 evidences $E=\{X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}, X^{(5)}\}$ along with network in figure 4 but the evidences $X^{(2)}$ and $X^{(5)}$ have not data yet.

Table 2 Set of evidences E (for network in figure 4) with data missing

	$\mathbf{x_1}$	$\mathbf{x_2}$
$\mathbf{X^{(1)}}$	$X_1^{(1)}=1$	$X_2^{(1)}=1$
$\mathbf{X^{(2)}}$	$X_1^{(2)}=1$	$X_2^{(2)}=\boldsymbol{v_1}$**?**
$\mathbf{X^{(3)}}$	$X_1^{(3)}=1$	$X_2^{(3)}=1$
$\mathbf{X^{(4)}}$	$X_1^{(4)}=1$	$X_2^{(4)}=0$
$\mathbf{X^{(5)}}$	$X_1^{(5)}=0$	$X_2^{(5)}=\boldsymbol{v_2}$**?**

Table 3 New split evidences E' for network in figure 4

	$\mathbf{x_1}$	$\mathbf{x_2}$
$\mathbf{X^{(1)}}$	$X_1^{(1)}=1$	$X_2^{(1)}=1$
$\mathbf{X'^{(2)}}$	$X_1'^{(2)}=1$	$X_2'^{(2)}=1/2$
$\mathbf{X'^{(2)}}$	$X_1'^{(2)}=1$	$X_2'^{(2)}=1/2$
$\mathbf{X^{(3)}}$	$X_1^{(3)}=1$	$X_2^{(3)}=1$
$\mathbf{X^{(4)}}$	$X_1^{(4)}=1$	$X_2^{(4)}=0$
$\mathbf{X'^{(5)}}$	$X_1'^{(5)}=0$	$X_2'^{(5)}=1/2$
$\mathbf{X'^{(5)}}$	$X_1'^{(5)}=0$	$X_2'^{(5)}=1/2$

As known, s_{21} , t_{21} and s_{22} , t_{22} can't be computed directly, it means that it is not able to compute directly the posterior density functions $\beta(f_{21}|E)$ and $\beta(f_{22}|E)$. In evidence $X^{(2)}$, v_1 must be determined. Obviously, v_1 obtains one of two values which is respective to two situations:

- $X_1^{(2)} = 1$ and $X_2^{(2)}=1$, it is easy to infer that $v_1=Pr(X_2^{(2)}=1|X_1^{(2)}=1)=E(\beta_{21})=\frac{a_{21}}{a_{21}+b_{21}}=½$
- $X_1^{(2)} = 1$ and $X_2^{(2)}=0$, it is easy to infer that $v_1=Pr(X_2^{(2)}=1|X_1^{(2)}=0)=E(\beta_{22})=\frac{a_{22}}{a_{22}+b_{22}}=½$

We split $X^{(2)}$ into two $X'^{(2)}$ s corresponding to two above situations in which the probability of occurrence of $X_2=1$ given $X_1=1$ is estimated as ½ and the probability of occurrence of $X_2=0$ given $X_1=1$ is also considered as ½. We perform similarly this task for $X^{(5)}$.

So, we have $\begin{pmatrix} s'_{21} = 1+\frac{1}{2}+1=\frac{5}{2} \\ t'_{21} = \frac{1}{2}+1=\frac{3}{2} \end{pmatrix}$ and $\begin{pmatrix} s'_{22} = \frac{1}{2} \\ t'_{22} = \frac{1}{2} \end{pmatrix}$ where s'_{21}, t'_{21}, s'_{22}, t'_{22} are the counts in E'. Then

$\beta(f_{21}|E)= \beta(f_{21};a_{21}+s'_{21},b_{21}+t'_{21})=\beta(f_{21};1+5/2,1+3/2)= \beta(f_{21};7/2, 5/2)$

$\beta(f_{22}|E)= \beta(f_{22};a_{22}+s'_{22},b_{22}+t'_{22})=\beta(f_{22};1+1/2,1+1/2)= \beta(f_{22};3/2, 3/2)$

$$Pr(X_2=1|\, X_1=1)=E(\beta(f_{21}|E))= \frac{7/2}{7/2+5/2}=\frac{7}{12}$$

$$Pr(X_2=0|\, X_1=1)=E(\beta(f_{22}|E))= \frac{3/2}{3/2+3/2}=\frac{1}{2}$$

If there are more evidences, this task repeated more and more brings out the EM algorithm having two steps.

1. Step1. We compute s'_{ij} and t'_{ij} based on the expected value of given $\beta(f_{ij})$, $s_{ij}=E(\beta(f_{ij}))$ and $t'_{ij}=1- E(\beta(f_{ij}))$. Next, replacing missing data by s'_{ij} and t'_{ij} . This step is called the **E**xpectation step.
2. Step 2. We determine the posterior density function f_{ij} by computing its parameters $a_{ij}=a_{ij}+s_{ij}$ and $b_{ij}=b_{ij}+t_{ij}$. Note that s_{ij} and t_{ij} are recomputed absolutely together on occurrence of s'_{ij} and t'_{ij}. Terminating algorithm if the stop condition (for example, the number of iterations approaches *k times*) becomes true, otherwise, reiterating step 1. This step is called the **M**aximization step.

After k^{th} iteration, we have $\lim_{k\to\infty} Expectation_{ij} = \lim_{k\to\infty} \frac{a_{ij}+s_{ij}^{(k)}}{a_{ij}+s_{ij}^{(k)}+b_{ij}+t_{ij}^{(k)}}$ which will approache an certain limit. Don't worry about the case of infinite iterations, we will obtain approximate s'_{ij}, t'_{ij}, *posterior* f_{ij} if k is large enough due to certain value of $\lim_{k\to\infty} Expectation_{ij}$

4. Conclusion

BN is a powerful mathematical tool for reasoning but it is restricted by unimproved initial parameters. This paper suggests the approach to parameter evolution that uses the EM algorithm for beta functions. Note that the particular features of beta function make this suggestion feasible because it is possible to compute the expectation of beta function which is the conditional probability in BN. Whether the EM converges quickly or not depends on how to pre-define the parameters. So, we specify the initial parameters (a_{ij}, b_{ij}) by weights of arcs.

However, the qualitative model (graph structure) is now fixed. It is creative to apply learning machine algorithms to enhance entirely the structure of BN. That is learning structure process which will be represented in other papers.

5. Reference

[1] David Heckerman. A Tutorial on Learning With Bayesian Networks. Technical Report MSR-TR-95-06. Microsoft Research Advanced Technology Division, Microsoft Corporation.

[2] Mislevy, R., & Gitomer, D. H. (1996). The Role of Probability-Based Inference in an Intelligent Tutoring System. User Modeling and User-Adapted Interaction, 5, 253-282.

[3] Conati, C., Gertner, A., VanLehn, K., & Druzdzel, M. (1997). On-line student modelling for coached problem solving using Bayesian Networks. Proceedings of the 6th International Conference on User Modelling UM'97. Wien, New York: Springer Verlag

[4] Henze, N., Nejdl, W.: Student modeling for KBS Hyperbook system using Bayesian networks. Technical report, University of Hannover (1999).

[5] Richard E. Neapolitan. Learning Bayesian Networks. Northeastern Illinois University Chicago, Illinois 2003.

[6] Tomoyosi Akiba, Hozulni Tanaka. A Bayesian Approach for User Modeling in Dialogue Systems. COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics.

[7] Eric Horvitz, Jack Breese, David Heckerman, David Hovel, Koos Rommelse. The Lumière Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users. Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, July 1998, pages 256-265. Morgan Kaufmann: San Francisco.

[8] Roberto Tedesco, Peter Dolog, Wolfgang Nejdl, Heidrun Allert. Distributed Bayesian Networks for User Modeling. ELEARN 2006 : World Conference on E-Learning in Corporate, Government, Health Care, and Higher Education.