2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Bayesian Approach to Learn Bayesian Networks Using Data and Constraints

Gao Xiao-guang
Northwestern Polytechnical University
Xi'an, China
Email: cxg2012@nwpu.edu.cn

Yang Yu
Northwestern Polytechnical University
Xi'an, China
Email: youngiv@126.com

Guo Zhi-gao
Northwestern Polytechnical University
Xi'an, China
Email: guozhigao2004@163.com

Chen Da-qing
London South Bank University
London, UK
Email: chend@lsbu.ac.uk

***Abstract*—One of the essential problems on Bayesian networks (BNs) is parameter learning. When purely data-driven methods fail to work, incorporating supplemental information, like expert judgments, can improve the learning of BN parameters. In practice, expert judgments are provided and transformed into qualitative parameter constraints. Moreover, prior distributions of BN parameters are also useful information. In this paper we propose a Bayesian approach to learn parameters from small datasets by integrating both parameter constraints and prior distributions. First, the feasible parameter region is derived from constraints. Then, using the prior distribution, a posterior distribution over the feasible region is developed based on the Bayes theorem. Finally, the parameter estimations are taken as the mean values of the posterior distribution. Learning experiments on standard BNs reveal that the proposed method outperforms most of the existing methods.**

I. Introduction

Bayesian networks [1] are efficient tools for performing inferences [2]. For BN learning, there are two directions: parameter learning and structure learning. In practice, domain experts tend to have problems providing the numerical conditional probabilities of a BN [3] when the structure is known previously. In this paper, we focus on parameter learning with fixed structures. In the procedure of data collecting, sufficient data may be unavailable, or gathered data are unrepresentative. In those scenarios, data can not reveal underlying true parameters, so purely data-driven methods often fail to work. Therefore, enhancing the precision of parameter learning by incorporating supplemental information is necessary.

Instead of numerical parameters, domain experts prefer to provide qualitative knowledge about parameters, which can be transformed into parameter constraints [4]. To detail the transformation from expert judgments to parameter constraints, preferred qualitative influence is considered. For two binary variables X and Y (Y is the cause and X is the effect), Y imposes a positive (or negative) influence on X if a greater value of Y tends to make X greater (or smaller). Then, a parameter constraint is obtained as $p(X = 1|Y = 1, s) \geq p(X = 1|Y = 0, s)$ for the positive qualitative influence, or $p(X = 1|Y = 1, s) \leq p(X = 1|Y = 0, s)$ for the negative qualitative influence, where s is the configuration for the causes of X other than Y. Analogously, we can transform other judgments into qualitative constraints. In addition, precise prior distributions of parameters are also helpful for improving parameter learning. However, they are frequently not readily available.

A new parameter learning method is proposed in this paper based on Bayesian estimation incorporating both parameter constraints and prior distributions. First, feasible parameter regions are defined by constraints. Then, posterior distributions over the feasible regions are built based on the prior distributions. Finally, parameter estimations are taken as mean values of the posterior distributions. Incorporating prior knowledge, the proposed method can remarkably improve parameter learning when data are insufficient or sparse.

The rest of this paper is organized as follow: In section 2 we discuss the related work. In section 3 we introduce basic concepts of BNs. In section 4 we propose a new parameter learning method. In section 5 some learning experiments on standard BNs are implemented. In section 6 we briefly summarize the work in this paper and indicate the future work.

II. Related Work

Several methods using prior information have been proposed to improve parameter learning accuracy.

Altendorf [5], Witting [6], Niculescu [7] and Liao [8] apply CO model to express parameter learning[1]. Then use the adaptive probabilistic network method (APN) [9], a gradient-based algorithm, to solve it. Chang [10] proposed a QMAP method to learn parameters from data and expert judgments, whose result is an average of all MAP estimations with different prior Dirichlet distributions.

Feelders [11] proposes modelling parameter learning with qualitative influences as a case of IR, and uses the minimum

[1]Niculescu and Liao explore the CO method to address parameter learning from incomplete data

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

lower sets (MLS) algorithm [12] to solve it. However, de Campos [13], [14] suggests modelling parameter learning as a case of convex optimization. Zhou [15]–[17] first reconstructs an auxiliary BN about querying parameters. Then use a dynamic discretization junction tree (DDJT) [18] algorithm to infer the posterior distributions of parameters. Finally, mean values of the posterior distributions are taken as parameter estimations.

III. Preliminaries

A. Bayesian Networks

A BN can be defined as a dual (G, θ). The first term G, which is a directed acyclic graph, can be further represented by another dual (U, E), where $U = \{X_1, \cdots, X_n\}$ is the set of nodes (or stochastic variables) and $E = \{X_i \to X_j | X_i, X_j \in U, i \neq j\}$ is the set of directed edges that express relevance or dependency relationships among variables. The second term θ is the set of parameters associated with G. The set consisting of all parents of X_i is denoted as pa_i. According to the Markov condition, the joint distribution of U can be decomposed as shown by Equation (1):

$$p(\{X_1, \cdots, X_n\}) = \prod_{i=1}^{n} p(X_i | pa_i) \tag{1}$$

We denote the cardinality of states of X_i and the cardinality of configurations of pa_i respectively as r_i and q_i. Specifically, a single parameter is marked as θ_{ijk} in this paper, where $i = 1, \cdots, n$ ranges over all the variables in a BN, $j = 1, \cdots, q_i$ ranges over all the possible configurations of pa_i, and $k = 1, \cdots, r_i$ ranges over all the possible states of X_i.

Let N_{ijk} be the number of data records corresponding to θ_{ijk}; i.e., the count of observations where $X_i = k$ and $pa_i = j$. Similarly N_{ij} notates the number of data records for $pa_i = j$. Then the maximum likelihood estimation (MLE) of a single parameter θ_{ijk} is given by Equation (2):

$$\theta_{ijk} = \frac{N_{ijk}}{N_{ij}} \tag{2}$$

B. Constraints on Parameters

Domain experts can provide knowledge about BN parameters from their experience; however, experts have some problems giving quantitative prior knowledge, such as the prior distribution of parameters. In practice experts more easily give qualitative prior knowledge, which can be transformed into parameter constraints. Most of these constraints can be expressed by a linear inequality function [10] given by Equation (3):

$$f(\theta) \leq c \tag{3}$$

where f is a linear function, and c is a scalar. Three types of typical constraints can be derived from this equation.

(1) An ***approximate equality constraint*** is an assertion that one parameter is very close to another one, and can be given by Equation (4):

$$\theta_{ijk} \approx \theta_{i'j'k'}, \forall ijk \neq i'j'k' \tag{4}$$

Alternatively, we introduce a small positive rational number ε, then rewrite equation (4) as Equation (5):

$$|\theta_{ijk} - \theta_{i'j'k'}| < \varepsilon, \forall ijk \neq i'j'k' \tag{5}$$

(2) A ***range constraint*** defines the upper and lower bounds of a single parameter, and can be expressed by Equation (6):

$$\alpha < \theta_{ijk} < \beta, \forall ijk \tag{6}$$

where $\alpha < \beta \in [0, 1]$.

(3) An ***inequality constraint*** claims an inequality relationship between two parameters, and can be represented by Equation (7):

$$\theta_{ijk} < \theta_{i'j'k'}, \forall ijk \neq i'j'k' \tag{7}$$

***Definition* 1.** A convex constraint is a constraint such that the region restricted by it is a convex set.

Since the parameter learning method we propose in this paper will compute linear combinations of random samples from the regions restricted by parameter constraints, emphasizing the convex nature of parameter constraints is essential. A non-convex constraint may make the linear combinations violate the constraints, which is unacceptable.

For further information about BN parameter constraints, please refer to Druzdzel's paper [4].

IV. The New Method

It's hard to build a precise object function from sparse data combined with qualitative parameter constraints. Hence, instead of pursuing a perfect object function, we fuzz the parameter learning in view of the lack of information. Assuming the true value of θ satisfies given constraints, we consider that an arbitrary value in the feasible region restricted by constraints can be the true value with a certain probability. We are working on finding the posterior distribution of all possible values of θ, rather than deciding which one is the real value.

Based on the above perspectives, a novel approach, called the *constrained Bayesian estimation* (CBE), is proposed for BN parameter learning from small datasets incorporating convex parameter constraints and prior distributions.

A. Assumptions

***Assumption* 1.** The prior distribution $\pi(\theta|G)$ of θ is a Dirichlet distribution with hyperparameters $\tau = \{\tau_{ijk} | i = 1, \cdots, n, j = 1, \cdots, q_i, k = 1, \cdots, r_i\}$. That is

$$\pi(\theta|G) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{\tau_{ijk}}$$

***Assumption* 2.** The true value of θ satisfies all the given parameter constraints, which means we assume the constraints are absolutely correct.

***Assumption* 3.** θ is globally independent, which is

$$p(\theta) = \prod_{i=1}^{n} p(\theta_i)$$

***Assumption* 4.** θ_i is locally independent, which is

$$p(\theta_i) = \prod_{j=1}^{q_i} p(\theta_{ij})$$

For assumption 1, if an accurate prior Dirichlet distribution is unavailable, a uniform distribution is an excellent alternative. Assumption 2 is significant for the asymptotic correctness of the CBE method. Otherwise, parameter estimations of the proposed method will not converge to true values. Assumption 3 and assumption 4 are applied to decompose a higher dimensional integral into lower ones with independent relationships expressed by the hypotheses.

B. The CBE Approach

Considering the definition of likelihood functions, the conditional distribution of data D given $\forall\theta \in \Theta_G$ (G is a fixed structure of a BN, and Θ_G is the corresponding parameter space) is equal to the likelihood function, which is defined by Equation (8):

$$q(D|\theta) = l(\theta, D) \propto \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \tag{8}$$

For convenience, we are, from now on, going to omit the common condition G, as it is always satisfied.

According to the *Bayes theorem*, the joint probability $f(\theta, D)$ of θ and D is the product of the prior distribution $\pi(\theta)$ and $q(D|\theta)$. Due to assumption 1, $\pi(\theta)$ is a Dirichlet distribution with hyperparameters. Then we can express $f(\theta, D)$ by Equation (9):

$$f(\theta, D) = \pi(\theta)q(D|\theta) \propto \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}+\tau_{ijk}} \tag{9}$$

After introducing parameter constraints Ω, there are actually two transformations on $f(\theta, D)$. The first one is that for $\forall\theta \notin \Theta_G^\Omega$ (Θ_G^Ω denotes the feasible region restricted by constraints), $f(\theta, D)$ tends to zero from a positive value, because these values of θ are certainly not the true values considering assumption 2. The second one is that for $\forall\theta \in \Theta_G^\Omega$, $f(\theta, D)$ changes because of the decreasing of the feasible region. Nevertheless, since likelihoods of values of θ remain unchanged, we legitimately suppose $f(\theta, D)$ for $\forall\theta \in \Theta_G^\Omega$ are increase with the same proportion. Thus we derive Equation (10):

$$f(\theta, D|\Omega) \propto f(\theta, D) \propto \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}+\tau_{ijk}}, \forall\theta \in \Theta_G^\Omega \tag{10}$$

where $f(\theta, D|\Omega)$ is the joint distribution of θ and D restricted by Ω.

To obtain the posterior distribution of θ, first the marginal probability $m(D|\Omega)$ of D given Ω should be determined. $m(D|\Omega)$ can be calculated by integrating $f(\theta, D|\Omega)$ on θ over the feasible region Θ_G^Ω as shown by Equation (11):

$$m(D|\Omega) = \int_{\Theta_G^\Omega} f(\theta, D|\Omega)d\theta \tag{11}$$

According to the *Bayes theorem*, the posterior distribution $h(\theta|D, \Omega)$ of θ is the quotient of $f(\theta, D|\Omega)$ and $m(D|\Omega)$, which can be written as

$$h(\theta|D, \Omega) = \frac{f(\theta, D|\Omega)}{m(D|\Omega)} \tag{12}$$

Seen from equations (10) and (11), $f(\theta, D|\Omega)$ and $m(D|\Omega)$ possess the same nonzero constant coefficient, and it can be eliminated by the division operation shown in Equation (12). Then, the posterior distribution of θ can be implemented as Equation (13):

$$h(\theta|D, \Omega) = \frac{\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}+\tau_{ijk}}}{\int_{\Theta_G^\Omega}\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}+\tau_{ijk}}d\theta} \tag{13}$$

The Bayesian estimation of θ is the mean value of the posterior distribution as shown by Equation (14):

$$\hat{\theta} = \int_{\Theta_G^\Omega} \theta h(\theta|D, \Omega)d\theta \tag{14}$$

The calculation result of Equation (14) is a vector, which is composed of components shown by Equation (15):

$$\hat{\theta}_{ijk} = \frac{\int_{\Theta_G^\Omega} \theta_{ijk}\prod_{l=1}^{n}\prod_{m=1}^{q_i}\prod_{w=1}^{r_i} \theta_{lmw}^{N_{lmw}+\tau_{lmw}}d\theta}{\int_{\Theta_G^\Omega}\prod_{l=1}^{n}\prod_{m=1}^{q_i}\prod_{w=1}^{r_i} \theta_{lmw}^{N_{lmw}+\tau_{lmw}}d\theta} \tag{15}$$

Integrating parameters except θ_{ij} (assume the results is $\varphi(\theta_{ij})$), Equation (15) can be rewritten as

$$\hat{\theta}_{ijk} = \frac{\int_{\Theta_G^\Omega} \varphi(\theta_{ij})\theta_{ijk}\prod_{w=1}^{r_i} \theta_{ijw}^{N_{ijw}+\tau_{ijw}}d\theta_{ij}}{\int_{\Theta_G^\Omega} \varphi(\theta_{ij})\prod_{w=1}^{r_i} \theta_{ijw}^{N_{ijw}+\tau_{ijw}}d\theta_{ij}} \tag{16}$$

$\varphi(\theta_{ij})$ is often not a constant, which means parameter estimations of θ_{ij} are impacted by data and prior distributions corresponding to other parameters. It seems to contradict the global and local independences. To decouple θ_{ij} from other parameters, let

$$N_{lmk} = 0, \tau_{lmk} = 0$$

for $l \neq i$ or $m \neq j$ when estimating θ_{ij}. Then $\varphi(\theta_{ij})$ degenerates into a constant, which is

$$\varphi(\theta_{ij}) = c$$

Thus $\varphi(\theta_{ij})$ can be eliminated from Equation (16). Finally, we obtain the estimation for every single parameter θ_{ijk} by the CBE approach as given by Equation (17):

$$\hat{\theta}_{ijk} = \frac{\int_{\Theta_G^\Omega} \theta_{ijk}\prod_{w=1}^{r_i} \theta_{ijw}^{N_{ijw}+\tau_{ijw}}d\theta_{ij}}{\int_{\Theta_G^\Omega}\prod_{w=1}^{r_i} \theta_{ijw}^{N_{ijw}+\tau_{ijw}}d\theta_{ij}} \tag{17}$$

Notice that the integral domain in Equation (17) is still Θ_G^Ω rather than $(\Theta_{ij})_G^\Omega$, because part of the parameter constraints Ω don't just involve parameters θ_{ij}. Hence, each single parameter θ_{ijk} is probably not uniform in its feasible region when no data given. Take a simple BN $X_1 \rightarrow X_2$ with two binary nodes as an example (refer to Figure 1). Let there be a constraint $\theta_{211} < \theta_{221}$, then we find that $0 < \theta_{211} < 1$.

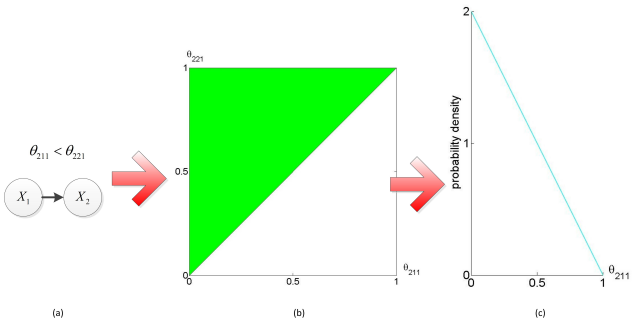

Fig. 1: The distribution of a parameter from a two binary nodes BN. (a) A small BN with a constraint. (b) The feasible region of $(\theta_{211}, \theta_{221})$. (c) The distribution of θ_{211}.

But the distribution of θ_{211} in interval $[0, 1]$ is not uniform. It's unbiased that the dual $(\theta_{211}, \theta_{221})$ is uniform in the top left corner of the unit square shown in Figure 1 (b), so the distribution of θ_{211} is a marginal distribution of $p(\theta_{211}, \theta_{221})$, which is

$$p(\theta_{211}) = \int_{\theta_{211}}^{1} 2d\theta_{221} = 2 - 2\theta_{211}$$

In Equation (17), if $\tau_{ijk} = 0$, the prior distribution degenerates into the uniform distribution. It's not a bad thing when an accurate prior distribution is unavailable, because a poor prior distribution may lead to terrible results. Actually, to find precise prior distributions of BN parameters is our future work to improve the basic CBE method.

There are two multidimensional definite integrals with the same form in Equation (17). If the integral domain is normative, it's easy to obtain results using the property of beta functions. However, it is hard to compute directly as the integral domain is not normative after introducing parameter constraints. Therefore, approximating the integral with a numerical integration [19] is an effective alternative. In this paper, we use a Monte Carlo (MC)method to compute the approximate result of a multidimensional definite integral.

Let $\{\vartheta^{(1)}, \cdots, \vartheta^{(M)}\}$ be M (a sufficient number, $M = 1,000$ in our experiments) uniform samples randomly generated from the feasible region Θ_G^{Ω} of BN parameters. Then, the numerical approximation of Equation (17) is shown in Equation (18):

$$\hat{\theta}_{ijk} = \frac{\sum_{m=1}^{M} \vartheta_{ijk}^{(m)} \prod_{w=1}^{r_i} (\vartheta_{ijw}^{(m)})^{N_{ijw}+\tau_{ijw}}}{\sum_{m=1}^{M} \prod_{w=1}^{r_i} (\vartheta_{ijw}^{(m)})^{N_{ijw}+\tau_{ijw}}} \qquad (18)$$

V. Experiments

We are going to compare the proposed method with the ML [20], [21], MAP [21], and CML [13] algorithms from learning experiments on standard BNs. Databases are synthesized by standard BNs with a common size of 50,000, and observations will be randomly sampled from databases in each learning case. Parameter constraints and prior distributions are representatively synthesized by certain rules described in the following. Although we don't consider physical processes

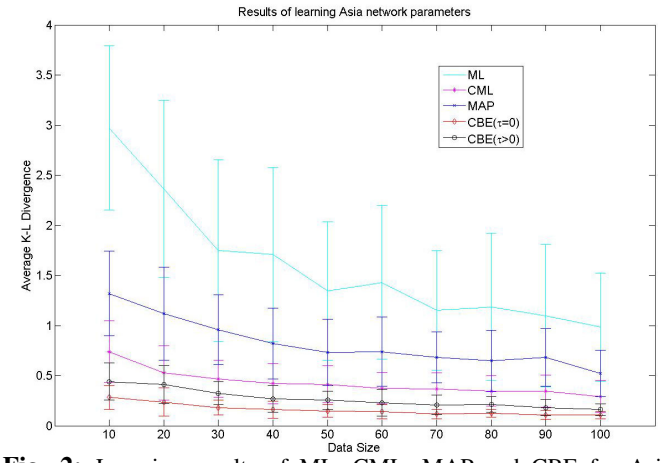

Fig. 2: Learning results of ML, CML, MAP and CBE for Asia network. The blue-green, carmine, and blue lines respectively represents the learning result of average baseline of ML, CML, and MAP approach, and the red and black line respectively shows learning results of the CBE algorithm with different prior distributions ($\tau = 0$ represents CBE method that doesn't use a prior distribution, $\tau > 0$ means CBE method that use a prior distribution with $K = 10$).

when they are synthesized, they possess identical mathematical properties with prior information provide by domain experts. To adequately demonstrate the differences in learning results, we repeatedly learn different standard BNs at various data sizes and record their average performances. The criterion for evaluating learning results is the Kullback-Leibler (K-L) divergence between parameter estimations and true parameters, which is widely used to describe the difference between two distributions.

A. Experiments on Asia Network

Asia network models the interaction between risk factors, diseases and symptoms for the purpose of diagnosing the most likely reason for a patient entering a chest clinic. Each of the 8 nodes in Asia network has two states. There is a logical node in Asia network, we don't learn its parameters since they can be determined previously without any information.

For the Asia network, the data sizes are fixed at 10, 20, $\cdots$, 100, and we learn all parameters at each data size with 50 repetitions to get an average learning performance. Data are randomly sampled from the 50,000 size database. The parameter constraints and prior distributions are synthesized by the following rules.

***Rules for synthesizing parameter constraints*:** (a) Generate a constraint $0 \leq p < 0.05$ if p is a small value close to 0; (b) Generate a constraint $0.95 < p \leq 1$ if p is a great value close to 1; (c) Generate a constraint $p1 \approx p2$ if $p1 = p2$; (d) Generate a constraint $p1 < p2$ if $p2 - p1 > 0.2$, where p, $p1$, and $p2$ are single parameters.

***Rules for synthesizing prior distributions*:** First choose a constant K ($K = 10$ in our experiments) to multiply the true parameters to generate initial hyperparameters. Then random fluctuations are add onto the initial hyperparameters with an amplitude of a half of K.

TABLE I: Learning Results* for Standard BNs

BNs	Nodes	Edges	Paras	Data	ML	CML	MAP	CBE
Andes	223	338	1157	10	0.680 ± 0.045	0.229 ± 0.006	0.450 ± 0.002	$\mathbf{0.195} \pm 0.003$
				50	0.312 ± 0.029	$\mathbf{0.159} \pm 0.005$	0.332 ± 0.002	0.173 ± 0.007
Win95pts	76	112	574	10	5.370 ± 0.112	$\mathbf{1.441} \pm 0.025$	1.364 ± 0.096	1.517 ± 0.102
				50	4.448 ± 0.143	$\mathbf{1.328} \pm 0.048$	1.226 ± 0.129	1.447 ± 0.138
Hepar2	70	123	1453	10	0.259 ± 0.005	$\mathbf{0.111} \pm 0.004$	0.158 ± 0.005	0.292 ± 0.012
				50	0.228 ± 0.006	$\mathbf{0.116} \pm 0.004$	0.146 ± 0.005	0.285 ± 0.009
Hailfinder	56	66	2656	10	0.946 ± 0.062	$\mathbf{0.322} \pm 0.009$	0.606 ± 0.025	0.442 ± 0.022
				50	0.429 ± 0.043	$\mathbf{0.241} \pm 0.014$	0.518 ± 0.023	0.406 ± 0.019
Alarm	37	46	509	10	0.596 ± 0.040	0.053 ± 0.004	0.217 ± 0.022	$\mathbf{0.042} \pm 0.002$
				50	0.431 ± 0.024	0.044 ± 0.003	0.186 ± 0.019	$\mathbf{0.037} \pm 0.002$
Insurance	27	52	984	10	2.327 ± 0.100	0.760 ± 0.038	1.030 ± 0.071	$\mathbf{0.712} \pm 0.059$
				50	1.376 ± 0.077	$\mathbf{0.508} \pm 0.034$	0.943 ± 0.071	0.650 ± 0.043
Boerlage92	23	36	86	10	0.388 ± 0.096	0.148 ± 0.017	0.201 ± 0.008	$\mathbf{0.096} \pm 0.005$
				50	0.134 ± 0.020	0.084 ± 0.012	0.125 ± 0.006	$\mathbf{0.076} \pm 0.004$
Sachs	11	17	178	10	3.188 ± 0.291	0.757 ± 0.053	1.444 ± 0.023	$\mathbf{0.582} \pm 0.009$
				50	2.901 ± 0.179	0.724 ± 0.069	1.537 ± 0.029	$\mathbf{0.602} \pm 0.035$
Survey	6	6	21	10	0.679 ± 0.198	0.443 ± 0.095	0.321 ± 0.009	$\mathbf{0.227} \pm 0.027$
				50	0.287 ± 0.108	0.220 ± 0.074	0.222 ± 0.012	$\mathbf{0.166} \pm 0.042$
Cancer	5	4	10	10	0.430 ± 0.198	0.088 ± 0.034	0.227 ± 0.013	$\mathbf{0.045} \pm 0.013$
				50	0.239 ± 0.140	0.080 ± 0.043	0.132 ± 0.013	$\mathbf{0.031} \pm 0.009$
Earthquake	5	4	10	10	0.512 ± 0.128	0.069 ± 0.041	0.251 ± 0.020	$\mathbf{0.026} \pm 0.005$
				50	0.455 ± 0.143	0.103 ± 0.046	0.114 ± 0.008	$\mathbf{0.023} \pm 0.006$
Weather	4	4	9	10	2.558 ± 4.027	0.470 ± 0.202	2.344 ± 0.222	$\mathbf{0.048} \pm 0.022$
				50	0.216 ± 0.082	0.106 ± 0.060	1.338 ± 0.166	$\mathbf{0.075} \pm 0.033$

* The first value is K-L divergence, and the second value is mean square deviation of it.

Learning results of ML, CML, MAP, and CBE algorithms are shown in Figure 3. We can see that the average error of ML is always maximal; on the contrary, the CBE algorithm performs best at all data sizes (tested in this set of experiments). The performances of CML and MAP are between ML and CBE. What's more, the CBE method without prior distributions remarkably performs better than the CBE method using prior distributions, which means that imprecise prior distributions will hold back learning process.

B. Experiments on Standard BNs

In this part, we learn twelve standard BNs, whose basic information is listed in Table 1, to compare different parameter learning methods. The above BNs are available at the BN repository[2] in addition to Boerlage92 [22], and they are widely used to evaluate learning algorithms. For each BN, data sizes are fixed at 10 and 50, and we repeat each learning case 20 times to compute the average learning performance. To make the K-L divergences of learning all the BNs have similar orders of magnitude, the learning results of the first eight BNs in Table 1 are further divided by the numbers of parameters. In addition, the parameter constraints and prior distributions are synthesized by rules mentioned in previous experiments.

Table 1 shows the learning results for different BNs of ML, CML, MAP, and CBE (without prior distributions). We can see that the CBE method achieves the best performances in experiments for learning Weather, Earthquake, Cancer, Sachs, Boerlage92, and Alarm BNs at data sizes of 10 and 50. CBE also overtakes other methods in experiments of learning such as Insurance and Andes BNs at 10 data, but the CML method performs better than CBE in the learning case at 50 data. It's no denying that the CBE method does not achieve the best performance in experiments of learning Hailfinder, hepar2, and Win95pts BNs.

CML method performs better than proposed method on K-L divergence in some cases. However, it doesn't mean estimations of CML are more accurate. K-L divergences are computed in this paper by following formula:

$$KL = \sum \hat{\theta}_{ijk} \log \frac{\hat{\theta}_{ijk}}{\theta_{ijk}}$$

When training data are sparse, numerous estimations of CML are zeros (however, true parameters are often not zeros), which don't accumulate the total K-L divergence.

VI. Conclusion

We propose a CBE method for parameter learning incorporating expert judgments, which is a derivation approach of the Bayesian estimation. We first establish the posterior probability distributions of parameters over feasible regions

[2]http://www.bnlearn.com/bnrepository/

with limited training data. Then, mean values of the posterior distributions are taken as parameter estimations.

In the experiments of learning standard BNs, it is proved that the CBE method performs best in small BNs learning but not so good in big BNs learning. In fact, a big BN means more parameters and constraints, which increase the difficulty in generating sample by MC method. Since rejection/acceptance approach fails to work, we generate samples using a linear transformation approach. Because non-uniform samples are generated by this sampling approach when BNs are big, it lowers the performance of CBE. What's more, the computing way of K-L divergences is in favour of CML method.

For possible future work, improving quality of samples generating from parameter feasible region may improve the CBE method.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (61573285) and Fundamental Research Funds for the Central Universities (3102015BJ(II)GH01). Yang Yu is sponsored by Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University (CX201619). The authors would like to thank anonymous reviewers for their valuable feedback.

REFERENCES

[1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of plausible Inference*. San Mateo: Morgan Kaufmann, 1988.

[2] L. M. de Campos, "A scoring function for learning bayesian networks based on mutual information and conditional independence tests," *Journal of Machine Learning Research*, vol. 7, no. 2, pp. 2149–2187, 2006.

[3] M. Druzdel and L. V. D. Gaag, "Building probabilistic networks: Where do the numbers come from?" *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, no. 4, pp. 481–486, 2001.

[4] M. J. Druzdzel and L. C. V. D. Gaag, "Elicitation of probabilities for belief networks: Combining qualitative and quantitative information," *UU-CS*, vol. 23, pp. 141–148, 1995.

[5] E. E. Altendorf, A. C. Restificar, and T. G. Dietterich, "Learning from sparse data by exploiting monotonicity constraints," in *Conf. Uncertainty in Artificial Intelligence*, 2012, pp. 18–26.

[6] F. Writting and A. Jameson, "Exploiting qualitative knowledge in the learning of conditional probabilities of bayesian networks," in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*. San Francisco, USA: Morgan Kaufmann, 2000, pp. 644–652.

[7] R. S. Niculescu, T. M. Mitchell, and R. B. Rao, "A theoretical framework for learning bayesian networks with parameter inequality constraints." in *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, 2007, pp. 155–160.

[8] W. Liao and Q. Ji, "Learning bayesian network parameters under incomplete data with domain knowledge," *Pattern Recognition*, vol. 42, no. 11, pp. 3046–3056, 2009.

[9] J. Binder, D. Koller, S. Russell, and K. Kanazawa, "Adaptive probabilistic networks with hidden variables." *Machine Learning*, vol. 29, no. 2-3, pp. 213–244, 1997.

[10] R. Chang and W. Wang, "Novel algorithm for bayesian network parameter learning with informative prior constraints," in *The 2010 International Joint Conference on Neural Networks*, 2010, pp. 1–8.

[11] A. Feelders and L. C. V. D. Gaag, "Learning bayesian network parameters under order constraints," *International Journal of Approximate Reasoning*, vol. 42, no. 1C2, pp. 37–53, 2006.

[12] H. D. Brunk, "Maximum likelihood estimates of monotone parameters," *Annals of Mathematical Statistics*, vol. 26, no. 4, pp. 607–616, 1955.

[13] C. P. D. Campos, Y. Tong, and Q. Ji, *Constrained Maximum Likelihood Learning of Bayesian Networks for Facial Action Recognition*. Springer Berlin Heidelberg, 2008.

[14] C. P. De Campos and Q. Ji, "Improving bayesian network parameter learning using constraints," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008, pp. 1–4.

[15] Y. Zhou, N. Fenton, and M. Neil, *An Extended MPL-C Model for Bayesian Network Parameter Learning with Exterior Constraints*. Springer International Publishing, 2014.

[16] Y. Zhou, N. Fenton, and M. Neil, "Bayesian network approach to multinomial parameter learning using data and expert judgments," *International Journal of Approximate Reasoning*, vol. 55, no. 5, pp. 1252–1268, 2014.

[17] T. Hospedales, Y. Zhou, N. Fenton, and M. Neil, "Probabilistic graphical models parameter learning with transferred prior and constraints," in *Uncertainty in Artificial Intelligence*, 2015.

[18] M. Neil, M. Tailor, and D. Marquez, "Inference in hybrid bayesian networks using dynamic discretization." *Statistics and Computing*, vol. 17, no. 17, pp. 219–233, 2007.

[19] P. J. Davis and P. Rabinowitz, *Methods of numerical integration*. Academic Press,, 1975.

[20] J. Sijbers, A. J. D. Dekker, P. Scheunders, and D. V. Dyck, "Maximum-likelihood estimation of rician distribution parameters." *IEEE Transactions on Medical Imaging*, vol. 17, no. 3, pp. 357–361, 1998.

[21] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[22] B. Boerlage, *Link Strength in Bayesian Networks*. Unisversity of British Comlumbia Vancouver, Canada, 1994.