ISSN1000-9825 Journal of Software 软件学报 2000,11(5):660~666

用于数据挖掘的贝叶斯网络*

慕春棣[1] 戴剑彬[1] 叶俊[2]

[1](清华大学自动化系 北京 100084)

[2](清华大学应用数学系 北京 100084)

E-mail: muchd@tsinghua.edu.cn

摘要 贝叶斯网络是用来表示变量集合的连续概率分布的图形模式,它提供了一种自然地表示因果信息的方法,用来发现数据间的潜在关系.贝叶斯网络的学习也就是要找出一个能够最真实反映现有数据库中各数据变量相互之间的依赖关系的贝叶斯网络模型,即根据数据样本 D 和先验知识 ξ,找出后验概率 $p(s^h|D,\xi)$ 最大的贝叶斯网络 S.该文在数学上对贝叶斯网络的学习方法进行了严格的推导,用一个实例来说明贝叶斯网络的计算过程,并介绍了贝叶斯网络在数据挖掘领域内的应用.

关键词 数据挖掘,贝叶斯网络,贝叶斯概率,先验概率,后验概率.

中图法分类号 TP18

贝叶斯网络是用来表示变量集合的连接概率分布的图形模型,它提供了一种自然地表示因果信息的方法.贝叶斯网络本身并没有输入和输出的概念,各结点的计算是独立的,因此,贝叶斯网络的学习既可以由上级结点向下级结点推理,也可以是由下级结点向上级结点的推理.用于数据挖掘的贝叶斯网络方法主要有以下几个特点[1]:

(1) 贝叶斯网络可以处理不完整和带有噪声的数据集.它用概率测度的权重来描述数据间的相关性,从而解决了数据间的不一致,甚至是相互对立的问题.

(2) 贝叶斯网络用图形的方法描述数据间的相互关系,语义清晰,可理解性强,这将有助于利用数据间的因果关系来进行预测分析.

(3) 由于贝叶斯网络具有因果和概率性语义,它有助于先验知识和概率的结合,容易与优化决策方法相结合.

贝叶斯网络最初是由 R. Howard 和 J. Matheson 于 1981 年提出来的.早期的贝叶斯网络主要在专家系统中用来表述不确定的专家知识.90 年代以来,对贝叶斯网络学习的方法研究有了很大的进展,主要的文献有 D. Heckerman 等人[2-6],W. Buntine[7],G. Cooper, E. Herskovits[8]等人的论著进行了严格的推导,并用一个实例来说明贝叶斯网络的计算过程,介绍了贝叶斯网络在数据挖掘领域内的应用.

1 贝叶斯概率

简单地说,贝叶斯概率是观测者对某一事件的发生的相信程度.观测者根据先验知识和现有的统计数据,用概率的方法来预测未知事件发生的可能性.贝叶斯概率不同于事件的客观概率.客观概率是在多次重复实验中事件发生的频率的近似值,而贝叶斯概率则是利用现有的知识对未知事件的预测.

记 $D=\{X_1=x_1, X_2=x_2, \ldots, X_m=x_m\}$ 为重复 m 次实验所得到的观测样本,其中 X 为事件变量,x 为变量值

* 本文研究得到国家 CIMS 工程研究中心基金(No. CIMS JJ. 96-001)资助.作者慕春棣,女,1946 年生,教授,主要研究领域为自动控制理论、优化、调度与决策支持.戴剑彬,1972 年生,博士生,主要研究领域为数据仓库,数据挖掘.叶俊,1965 年生,博士,副教授,主要研究领域为随机过程,统计分析.

本文通讯联系人:慕春棣,北京 100084,清华大学自动化系

本文 1999-03-15 收到原稿,1999-06-07 收到修改稿

或状态.记参数 θ 为事件 $X=x$ 发生的客观概率或先验概率,$p(\theta|\zeta)$为它的概率密度函数,其中 ζ 为观测者的先验知识.这样,贝叶斯概率的计算问题可以陈述如下:已知先验概率密度 $p(\theta|\zeta)$和样本 D,求第 $m+1$ 次实验中的事件 $X_{m+1}=x_{m+1}$发生的概率 $P(X_{m+1}=x_{m+1}|D,\zeta)$.

由全概率公式得

$$p(X_{m+1}=x_{m+1}|D,\zeta)=\int p(X_{m+1}=x_{m+1}|\theta,D,\zeta)p(\theta|D,\zeta)\mathrm{d}\theta=\int\theta p(\theta|D,\zeta)\mathrm{d}\theta=E_{p(\theta|D,\zeta)}(\theta). \tag{1}$$

这说明,事件 $X_{m+1}=x_{m+1}$发生的贝叶斯概率即为先验概率 θ 相对于后验概率的期望值.根据贝叶斯规则,由先验概率密度 $p(\theta|\zeta)$计算后验概率密度 $p(\theta|D,\zeta)$的公式为

$$p(\theta|D,\zeta)=\frac{p(\theta|\zeta)p(D|\theta,\zeta)}{p(D|\zeta)}=\frac{p(\theta|\zeta)p(D|\theta,\zeta)}{\int p(D|\theta,\zeta)p(\theta\ \zeta)\mathrm{d}\theta}. \tag{2}$$

在先验概率 θ 已知的条件下,样本 D 中各事件 $X=x$ 的条件独立.如果事件变量 X 为二点分布,即事件只有发生或不发生两种情况,则

$$p(D|\theta,\zeta)=\theta^h(1-\theta)^t, \tag{3}$$

其中 h 为样本 D 中事件发生的次数,$h+t=m$.现假设先验概率为 β 分布,即

$$p(\theta|\zeta)=Beta(\theta|\alpha_h,\alpha_t)=\frac{\Gamma(\alpha)}{\Gamma(\alpha_h)\Gamma(\alpha_t)}\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}. \tag{4}$$

其中 $\alpha_h>0,\alpha_t>0$ 为 β 分布的参数,$\alpha=\alpha_h+\alpha_t$.显然,β 分布 $Beta(\theta|\alpha_h,\alpha_t)$的期望值为$\frac{\alpha_h}{\alpha}$.由式(2)~(4)可得,后验概率也为 β 分布,即

$$p(\theta|D,\zeta)=\frac{\Gamma(\alpha+m)}{\Gamma(\alpha_h+h)\Gamma(\alpha_t+t)}\theta^{\alpha_h+h-1}(1-\theta)^{\alpha_t+t-1}=Beta(\theta|\alpha_h+h,\alpha_t+t). \tag{5}$$

于是,预测事件的贝叶斯概率为

$$p(X_{m+1}=x_{m+1}|D,\zeta)=\int\theta Beta(\theta|\alpha_h+h,\alpha_t+t)\mathrm{d}\theta=\frac{\alpha_h+h}{\alpha+m}. \tag{6}$$

现在,我们讨论事件变量 X 取值为有限的情况,即 X 有 $x^1,x^2,\ldots,x^r$ 共 r 个可能的状态,参数矢量为 $\theta=(\theta_1,\theta_2,\ldots,\theta_r)$,其中

$$\theta_k=p(X=x^k|\theta,\zeta),\quad k=1,2,\ldots,r, \tag{7}$$

记统计数 N_i 为样本 D 中事件 $X=x^i$ 发生的次数,$i=1,2,\ldots,r$.现假设先验概率为 Dirichlet 分布,即

$$p(\theta\ \zeta)=\mathrm{Dir}(\theta|\alpha_1,\alpha_2,\ldots,\alpha_r)=\frac{\Gamma(\alpha)}{\prod_{k=1}^{r}\Gamma(\alpha_k)}\prod_{k=1}^{r}\theta_k^{\alpha_k-1}, \tag{8}$$

其中 $\alpha=\sum_{k=1}^{r}\alpha_k$,且 $\alpha_k>0,k=1,2,\ldots,r$.当 $r=2$ 时,Dirichlet 分布即为 β 分布,则后验概率也为 Dirichlet 分布

$$p(\theta|D,\zeta)=\mathrm{Dir}(\theta|\alpha_1+N_1,\alpha_2+N_2,\ldots,\alpha_r+N_r). \tag{9}$$

于是,预测事件的贝叶斯概率为

$$p(X_{m+1}=x^k|D,\zeta)=\int\theta_k\mathrm{Dir}(\theta|\alpha_1+N_1,\alpha_2+N_2,\ldots,\alpha_r+N_r)\mathrm{d}\theta=\frac{\alpha_k+N_k}{\alpha+N}. \tag{10}$$

2 贝叶斯网络

贝叶斯网络是描述数据变量之间依赖关系的图形模型,其描述由以下两部分组成:

(1) 网络结构 S.S 是一个有向图,其中每一个结点代表一个数据变量 X_i,P_{a_i}为 S 中结点 X_i 的父结点的集合(图中的结点和其对应的数据变量都用同一符号表示).

(2) X 的局部概率分布 P.P 中的每一元素为数据变量 X_i 的条件概率密度 $p(X_i|P_{a_i},\zeta)$.由概率的链规则得

$$p(X|\zeta)=p(X_1,X_2,\ldots,X_n|\zeta)=\prod_{i=1}^{n}p(X_i|X_1,X_2,\ldots,X_{i-1},\zeta). \tag{11}$$

对于任一数据变量 X_i，必可以找到一个与 X_i 条件都不独立的最小子集 $\pi_i \subseteq \{X_1, X_2, \ldots, X_{i-1}\}$，使得

$$p(X_i|X_1, X_2, \ldots, X_{i-1}, \zeta) = p\{X_i|\pi_i, \zeta\}. \tag{12}$$

此时，π_i 中的变量就为贝叶斯网络中的 X_i 的父结点 P_{a_i}，故

$$p(X|\zeta) = \prod_{i=1}^{n} p\{X_i|P_{a_i}, \zeta\}, \tag{13}$$

由 (S, P) 确定了一个贝叶斯网络. 下面，我们用一个简单的实例来描述贝叶斯网络模型.

例：在某生产企业采用了一项新技术后，经过一段实验性的生产，需对该技术的有效性进行评估. 现有以下几个数据变量：

(1) 新技术的有效值 T；

(2) 技术原因引起的产品不合格数 G；

(3) 非技术原因引起的产品不合格数 J；

(4) 操作人员的年龄 A；

(5) 操作人员的性别 S.

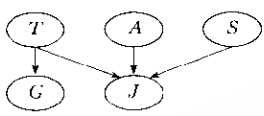


Fig. 1　Example's Bayesian network structure
图1　实例的贝叶斯网络结构

根据已有的经验知识，我们找出数据变量之间的因果关系，得出贝叶斯网络结构 S，如图 1 所示.

现在可以根据图 1 中的数据变量之间的因果关系得到以下的条件独立关系：

$$p(a|t) = p(a),$$
$$p(s|t) = p(s),$$
$$p(g|t, a, s) = p(g|t),$$
$$p(j|t, a, s, g) = p(j|t, a, s).$$

利用贝叶斯网络进行所需的概率计算. 例如，在数据库中已有 A, S, G, J 的统计数据，需要计算新技术的有效值 T 的概率密度.

$$\begin{aligned} p(t|a, s, g, j) &= \frac{p(t, a, s, g, j)}{\int p(t', a, s, g, j)\mathrm{d}t'} = \frac{p(t)p(a)p(s)p(g|t)p(j|t, a, s)}{\int p(t')p(a)p(s)p(g|t')p(j|t', a, s)\mathrm{d}t'} \\ &= \frac{p(t)p(g|t)p(j|t, a, s)}{\int p(t')p(g|t')p(j|t', a, s)\mathrm{d}t'}. \end{aligned} \tag{14}$$

3　贝叶斯网络的学习

贝叶斯网络的学习也就是找出一个能够最真实地反映现有数据库中各数据变量之间的依赖关系的贝叶斯网络模型. 记 $X = \{X_1, X_2, \ldots, X_n\}$ 为数据变量集，对于每一个 X_i，它的值域为 $\{x_i^1, x_i^2, \ldots, x_i^{r_i}\}$. $D = \{C_1, C_2, \ldots, C_m\}$ 为数据样本，其中的元素 C_l 为一事例. D_l 为前 $l-1$ 个事例集. 记 S^h 为数据样本 D 由贝叶斯网络结构 S 所产生的事件. 贝叶斯网络的学习过程也就是根据数据样本 D 和先验知识 ζ，找出后验概率 $p(S^h|D, \zeta)$ 最大的贝叶斯网络结构 S 的过程. 由贝叶斯概率公式得

$$p(S^h|D, \zeta) = \frac{p(S^h, D|\zeta)}{p(D|\zeta)}. \tag{15}$$

样本 D 的先验概率 $p(D|\zeta)$ 不依赖于网络结构 S，所以只需找出联合概率 $p(S^h, D|\zeta)$ 最大的网络结构 S. 记先验概率的参数变量

$$\theta_{ijk} = p(x_i^k|P_{a_i^j}, \theta_i, S^h, \zeta) > 0, \tag{16}$$

$$\sum_{k=1}^{r_i} \theta_{ijk} = 1, \tag{17}$$

其中 P_{a_i} 的值域为 $\{P_{a_i}^1, P_{a_i}^2, \ldots, P_{a_i}^{q_i}\}$，$q_i = \prod_{X_t \in P_{a_i}} r_t$ 为 P_{a_i} 所有可能状态的个数，则

$$p(X|\theta,S^h,\zeta)=\prod_{i=1}^{n}p(x_i|P_{a_i},\theta_i,S^h,\zeta). \tag{18}$$

对于贝叶斯网络的学习过程,我们提出以下3个假设条件:

(1) 随机样本 D 是完整的,即 D 中没有丢失的数据;

(2) 参数变量相互独立,即

$$p(\theta|S^h,\zeta)=\prod_{i=1}^{n}p(\theta_i|S^h,\zeta), \tag{19}$$

$p(\theta_i|S^h,\zeta)$为第 i 个变量的先验概率密度

$$p(\theta_i|S^h,\zeta)=\prod_{j=1}^{q_i}p(\theta_{ij}|S^h,\zeta), \tag{20}$$

$p(\theta_{ij}|S^h,\zeta)$为第 i 个变量 x_i,其父结点的取值为 $P_{a_i^j}$时的先验概率密度.

(3) 参数变量为 Dirichlet 分布,即

$$p(\theta_{ij}|S^h,\zeta)=\frac{\Gamma\left(\sum_{k=1}^{r_i}N'_{ijk}\right)}{\prod_{k=1}^{r_i}\Gamma(N'_{ijk})}\prod_{k=1}^{r_i}\theta_{ijk}^{N'_{ijk}-1}, \tag{21}$$

其中 $N'_{ijk}>0$ 为 Dirichlet 分布的指系数,它的大小与 S^h 和 ζ 有关. 当 $r_i=2$ 时,Dirichlet 分布即为 β 分布. 由参数变量的独立性假设得

$$p(C_l|D_l,\theta,S^h,\zeta)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{l_{ijk}}. \tag{22}$$

在事例 C_l 中,当 $X_i=x_i^K$,$P_{a_i}=P_{a_i}^k$时,$l_{\text{link}}=1$,在其他状态时,$l_{\text{link}}=0$. 记 N_{ijk}为数据样本 D 中 $X_i=x_i^k$,$P_{a_i}=P_{a_i}^k$,$i=1,2,\ldots,m$发生的次数,则

$$p(D|\theta,S^h,\zeta)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{N_{\text{link}}}. \tag{23}$$

下面,我们用前 $l-1$ 个事例来预测第 l 个事例发生的概率,类似于式(1)的处理,可知

$$p(C_l|D_l,S^h,\zeta)=\int p(C_l|\theta,D_l,S^h,\zeta)p(\theta|D_l,S^h,\zeta)\mathrm{d}\theta=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}E(\theta_{ijk}|D_l,S^h,\zeta)^{l_{\text{link}}}. \tag{24}$$

由假设的条件(3),即参数变量为 Dirichlet 分布,则由数据样本 D 预测未知事例 C_{m+1} 发生的概率为

$$p(C_{m+1}|D,S^h,\zeta)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\left(\frac{N'_{ijk}+N'_{ijk}}{N'_{ij}+N_{ij}}\right)^{l_{m+1,ijk}}, \tag{25}$$

其中 $N'_{ij}=\sum_{k=1}^{r_i}N'_{ijk}$,$N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$,则

$$\begin{aligned}p(D|S^h,\zeta)&=\prod_{l=1}^{m}p(C_l|D_l,S^h,\zeta)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\prod_{l=1}^{m}E(\theta_{ijk}|D_l,S^h,\zeta)^{l_{\text{link}}}\\&=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})},\end{aligned} \tag{26}$$

由此得出

$$p(S^h,D|\zeta)=p(S^h|\zeta)p(D|S^h,\zeta)=p(S^h|\zeta)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})}. \tag{27}$$

可见,联合概率 $p(S^h,D|\zeta)$只是由 Dirichlet 分布的指系数 N'_{ijk}来决定的,这表明,贝叶斯网络的学习过程也就是寻找合适的指系数 N'_{ijk},使联合概率 $p(S^h,D|\zeta)$最大. 记

$$g(X_i)=\prod_{j=1}^{q_i}\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})}, \tag{28}$$

则

$$p(S^h,D|\xi)=p(S^h|\xi)\prod_{i=1}^{n}g(X_i). \tag{29}$$

$g(X_i)$为数据变量X_i对联合概率$p(S^h,D|\xi)$的贡献值.各$g(X_i)$的计算是独立的,我们对每一数据变量X_i,逐个找出能使$g(X_i)$值增大的其他数据变量X_j,直到$g(X_i)$的值不再增加为止.这些X_j即为X_i的父结点,由此也就得出了贝叶斯网络的结构S.

4 计算实例

我们用一个社会调查研究的例子来说明贝叶斯网络的计算方法.通过对某地区的中学生进行调查,找出以下几个变量因素对学生的就学情况产生影响:

- 性别(X_1):male,female
- 智商(X_2):low,lower middle,upper middle,high
- 家庭经济(X_3):low,lower middle,upper middle,high
- 家庭鼓励(X_4):low,high
- 是否打算上大学(X_5):yes,no

表1是对10 318名学生的统计结果.表1中的第1格数据表示X_1=male,X_2=low,X_3=low,X_4=low,X_5=yes的学生个数为4,第2格数据表示X_1=male,X_2=low,X_3=low,X_4=low,X_5=no的学生个数为349,...,依此类推,在表1的下半部分(即5~8行),X_1的取值都为female.

Table 1

表1

4	349	13	64	9	207	33	72	12	126	38	54	10	67	49	43
2	232	27	84	7	201	64	95	12	115	93	92	17	79	119	59
8	156	47	91	6	120	74	110	17	92	148	100	6	42	198	73
1	48	49	57	5	47	123	90	9	41	224	65	8	17	414	54
5	454	39	44	5	312	14	47	8	216	20	35	13	96	28	24
11	235	29	61	19	236	47	88	12	164	62	85	15	113	72	50
7	133	36	72	13	193	75	90	12	174	91	100	20	81	142	77
6	50	36	58	5	70	110	76	12	48	123	81	13	49	360	98

用贝叶斯网络作数据挖掘就是要找出这些变量之间的因果关系,具体计算过程如下:

(1) 根据先验知识,选择合适的网络结构.

对于有n个变量的数据样本,可能组成的网络结构有$n!$种,要对每一个网络结构进行计算是不可能的.利用现有的专家知识,就可以排除大量的不合理组合.例如,在本例中,学生的性别和家庭的经济状况是没有关系的,所以在贝叶斯网络中,不会有X_1和X_3之间的联系.在下面的计算中,我们只选择了S_1和S_2两种网络结构,如图2所示.它们唯一的区别是学生的智商与家庭经济的因果关系不同.

(2) Dirichlet分布指数N'_{ijk}的计算.

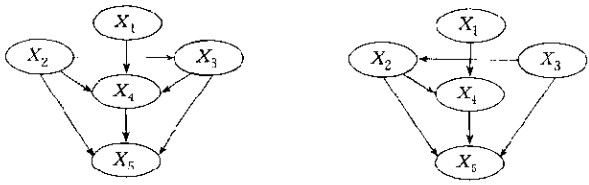
由式(21)可知,$g(X_i)$的计算只与N'_{ijk}有关,现在需要对其进行估计.在假定的网络结构已知的情况下,对第l个事例$C_l(l=2,3,\ldots,m)$的预测公式为:

$$p(C_l|D,S^h,\xi)=\prod_{i=1}^{n}\frac{N'_{ijk}+N_{ijk}}{N'_{ij}+N_{ij}}, \tag{30}$$

其中$N'_{ij}=\sum_{k=1}^{r_i}N'_{ijk}$,$N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$.统计数据$N_{ijk}$为已知,预测概率$p(C_l|D,S^h,\xi)$也很容易计算得出.例如,某一事例$C_l$($X_1$=male,$X_2$=low,$X_3$=low,$X_4$=low,$X_5$=no)的概率为349/10318=0.03382.因此,由式(25)可列出$m-1$个方程式,可以用最小方差估计出N'_{ijk}值.

(3) 网络S^h结构的选择.

各个$g(X_i)$的计算是独立的,可以根据式(28)来对各个网络结构进行计算,并选择出使$g(X_i)$最大的结构,



(a) S_1 network structure
(a) S_1 网络结构

(b) S_2 network structure
(b) S_2 网络结构

Fig. 2 Network structure
图2 网络结构图

这就是我们需要的贝叶斯网络.由此,各个数据变量的因果关系可以由网络图中得出.

在贝叶斯网络的实际计算中所面临的主要困难是根据式(25)对 N'_{ijk} 的估计.由于估计式是非线性的,在计算上有很大的困难.在文献[4,5]中,G. Cooper 和 E. Herskovits 提出用 $N'_{ijk}=1$ 来估计时,对网络的计算结果影响不大.在我们的例子中,用 $N'_{ijk}=1$ 代入式(30),分别对上述两个网络进行计算,并由式(15)和式(29)计算得出:

$$p(S_1^h|D,\xi)\cong 1.0,$$
$$p(S_2^h|D,\xi)\cong 1.2\times 10^{-12}.$$

由此得出,网络结构 S_1 更能反映变量之间的因果关系.同时,我们也注意到,虽然 S_1 和 S_2 的结构差别不大,但计算的结果截然相反,这说明贝叶斯网络有较好的敏感性.

5 结 论

用贝叶斯网络找出数据之间潜在的关系,正是数据挖掘所需要完成的功能[9].但是利用贝叶斯网络进行数据挖掘,主要问题是先验知识的重要性.由于我们不可能对所有的网络结构进行计算,特别是当变量增多时,可能的网络结构成倍增加,因此必须在现有的知识下进行网络选择,这在很大程度上依赖于专家知识.

参考文献

1 Chickering D. Learning equivalence classes of Bayesian networks structures. In: Horvitz E, Jensen F ed. Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers, Inc., 1996. 54~61

2 Geriger D, Hekerman D. A charactererization of the Dirichlet distribution with application to learning Bayesian networks. In: Besnard P, Hanks S eds. Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., 1995. 196~207

3 Heckman D. A Bayesian approach for learning causal networks. In: Besnard P, Hanks S eds. Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers, Inc., 1995. 285~295

4 Heckman D, Geiger D, Chickering D. Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning, 1995,20(3):197~243

5 Heckman D, Shachter R. Decision-Theoretic foundations for causal reasoning. Journal of Artificial Intelligence Research, 1995,3:405~430

6 Heckman D, Mandani A, Wellman M. Real-World applications of Bayesian networks. Communications of the ACM, 1995, 38(3):38~45

7 Buntine W. Theory refinement on Bayesian networks. In: Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence. Los Angeles, CA: Morgan Kaufmann Publishers, Inc., 1991. 52~61

8 Cooper G, Herskovits E. A Bayesian method for the introduction of probabilistic networks from data. Machine Learning, 1992,9(4):309~347

9 Russell S, Binder J, Koller D *et al*. Local learning in probabilistic networks with hidden variables. In: Cooper G F, Moral

S ed. Proceedings of the 14th International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers, Inc., 1998. 1146~1152

Bayesian Network for Data Mining

MU Chun-di[1] DAI Jian-bin[1] YE Jun[2]

[1](*Department of Automation Tsinghua University Beijing* 100084)

[2](*Department of Applied Mathematics Tsinghua University Beijing* 100084)

Abstract Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. It is a natural way to express the causal information, and to discover the hidden patterns among the data. Learning of Bayesian network is to find out a network model that best represents the dependent relationships of the variables in a database, that is, given sample D and prior knowledge ζ, to find a Bayesian network S that fits the maximum posterior probability $p(s^h|D,\zeta)$. In this paper, the learning process of the network is strictly derived, and a case study is presented to indicate the applications of Bayesian network in data mining.

Key words Data mining, Bayesian network, Bayesian probability, prior probability, posterior probability.