第 16 卷第 4 期,增刊　　广西科学院学报　　Vol. 16, No. 4 & Supplement
2000 年 11 月　　Journal of Guangxi Academy of Sciences　　November 2000

贝叶斯网络中的贝叶斯学习*

Bayesian Learing of Bayesian Network

胡振宇　　林士敏
Hu Zhenyu　　Lin Shimin

(广西师范大学计算机科学系　桂林　541004)
(Dept. of Computer Science, Guangxi Normal University, Guilin, 541004)

摘要　从机器学习的角度研究贝叶斯方法及其学习机制,着重讨论了具有完整数据、不完整数据集,及在结构不确定时贝叶斯网络进行学习的方法,表明贝叶斯网络在数据采掘中是一个有力的工具。文后给出一个基于贝叶斯网络的学习的实例。

关键词　贝叶斯网络　贝叶斯学习　机器学习　数据采掘
中图法分类号　TP 301.6

Abstract　The Bayesian method and its learning mechanism from the view of machine learning is explored. The Bayesian leaning approaches for different conditions such as complete data, incomplete data as well as uncertain network structure are discussed. It shows that Bayesian network is a powerful tool in data mining. An example from reality is given.

Key words　Bayesian approach, Bayesian network, machine learning, data mining

1　贝叶斯方法的学习机制

在经典的概率统计中,连续随机变量的贝叶斯定理有如下的形式:

$$\pi(\theta|x)=\frac{p(x|\theta)\pi(\theta)}{\int_{\theta}p(x|\theta)\pi(\theta)\mathrm{d}\theta},\tag{1}$$

其中 $\pi(\theta)$ 为先验分布密度,$p(x|\theta)\pi$ 为样本信息,而 $\pi(\theta|x)$ 是在给定样本 $X=(X_1,X_2,\cdots,X_n)$ 的条件下,θ 的条件分布密度函数,称为 θ 的后验分布密度函数,或简称为后验分布。

任何系统经过运行改善其行为,都是学习。从贝叶斯公式可以看出,后验信息是由先验信息与样本数据综合得到的。也就是说,可以通过样本信息使先验信息得以改善。因此贝叶斯公式中蕴含有一种学习机制。使用贝叶斯公式的学习机制来改善系统的功能,就称为贝叶斯学习。现以正态分布为例,分析其学习机制。

设 $X_1,X_2,\cdots,X_n$ 是来自正态分布 $N(\theta,\sigma^2)$ 的一个样本,其中 σ^2 已知,θ 未知。为了求 θ 的估

2000-09-01 收稿。
* 清华大学智能技术与系统国家重点实验室开放课题资助(编号 99002)。

计量 $\tilde{\theta}$，取另一个正态分布 $N(\mu_0,\sigma_0^2)$ 作为该正态均值 θ 的先验分布，即取先验为：

$$\pi(\theta) = N(\mu_0,\sigma_0^2).$$

用贝叶斯公式可以计算出后验仍为正态分布，后验密度为：

$$h(\theta|\bar{x}) = N(\alpha,\ d^2),$$

其中 $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$，$\alpha = (\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x})/(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2})$，$d^2 = (\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2})^{-1}$。用后验 $h(\theta|\bar{x})$ 的数学期望作为 θ 的估计值，有：

$$\tilde{\theta} = E(\theta|\bar{x}) = (\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x})/(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}).$$

由此可见，这样得到的估计 $\tilde{\theta}$ 是先验分布中的期望 μ_0 与样本均值 $\bar{x}$ 按各自的精度的加权平均。因为 σ_0^2 是 $N(\mu_0,\sigma_0^2)$ 的方差，它的倒数 $1/\sigma_0^2$ 就是 μ_0 的精度。样本均值 $\bar{x}$ 的方差是 σ_0^2/n，它的倒数 n/σ_0^2 就是样本均值 $\bar{x}$ 的精度。方差越小者在后验均值中所占的比重越大，方差越大者在后验均值中所占的比重越小。此外，样本数目 n 越大则 σ_0^2/n 越小，则样本均值 $\bar{x}$ 在后验均值中所占的比重越大。如果 n 无限增大，则先验均值在后验中的影响将变得很小。这说明贝叶斯公式求出的后验确实对先验信息和样本数据进行了合理的综合，其得到的结果比先验信息或样本数据都更完善。在先验分布密度函数得到合理确定的条件下，后验分布密度函数比单纯使用样本信息更符合实际，比单纯先验密度分布函数也更接近于实际。在采用其他共轭先验分布的情况下，也有类似的结果。

在贝叶斯方法之下，可以将先验信息和样本数据统一起来。在共轭先验的前提下，可以将得到的后验信息作为新一轮计算的先验，与进一步获得的样本信息综合，求得下一个后验信息。随着计算次数的增多，先验信息的影响逐渐减弱，样本信息的影响越来越显著。如果样本的噪音很小，得到的后验信息将越来越接近于实际。

在样本很多的情况下，先验分布密度对结果的影响变得很小。换句话说，即使任意估计先验分布密度，多次使用贝叶斯定理后也可以得到接近实际的结果，只不过需要大量的计算而已。但是，在样本不多的情况下，先验分布密度的估计的好坏对计算量和计算结果的影响就比较大。计算实践表明，如果能恰当地估计出先验分布密度，就可以使用少量样本数据，进行几次计算就得到比较满意的结果。这在样本不容易获得的情况下特别有用。因此，先验分布的确定是贝叶斯方法的一个重要问题。

2 贝叶斯网络及其语义

贝叶斯网络是一个带有概率注释的有向无环图。这种概率图模型能表示变量之间的联合概率分布（物理的或贝叶斯的），分析变量之间的相互关系，利用贝叶斯定理揭示的学习和统计推断功能，实现预测、分类、聚类、因果分析等数据采掘任务。

关于一组变量 $X = \{X_1, X_2, \cdots, X_n\}$ 的贝叶斯网络由以下两部分组成：(i) 一个表示 X 中的变量的条件独立断言的网络结构 S；(ii) 与每一个变量相联系的局部概率分布集合 P。两者定义了 X 的联合概率分布。S 是一个有向无环图，S 中的节点一对一地对应于 X 中的变量。以 X_i 表示变量以及该变量对应的节点，Pa_i 表示 S 中的 X_i 的父节点。S 的节点之间缺省弧线则表示条件独立。X 的联合概率分布表示为：

$$p(x) = \prod_{i=1}^{n} p(x_i|pa_i), \tag{2}$$

如果 P 表示(3)式中的局部概率分布，即乘积中的项 $p(x_i|Pa_i)(i=1,2,\cdots,n)$，则二元组$(S,P)$表示了联合概率分布 $p(X)$。当仅仅从先验信息出发建立贝叶斯网络时，该概率分布是贝叶斯的(主观的)。当从数据出发进行学习，进而建立贝叶斯网络时，该概率是物理的(客观的)。贝叶斯网络的语义如下：

(i) 贝叶斯网络对给定网络结构 S 编码了一组变量 $X=\{X_1,X_2,\cdots,X_n\}$ 的联合概率分布：

$$p(x)=\prod_{i=1}^{n}p(x_i|pa_i).$$

(ii) 贝叶斯网络表示条件独立及因果关系。

所谓 X_i 对于 $\{X_1,X_2,\cdots,X_{i-1}\}\backslash \Pi i$ 条件独立意味着变量 X_i 只依赖于变量集 $\{X_1,X_2,\cdots,X_{i-1}\}$ 中的某些变量 $\Pi i(i=1,2,\cdots,n)$，而与 $\{X_1,X_2,\cdots,X_{i-1}\}\backslash \Pi i$ 中的变量无关。前一种情况在贝叶斯网络中表现为变量之间有弧线连接，而后一种情况表现为变量之间无弧线连接。

(iii) 贝叶斯网络是概率的分类 / 回归模型。

假设一组变量 $X=(X_1,X_2,\cdots,X_n)$ 的物理联合概率分布可以编码在某个网络结构 S 中：

$$p(x|\theta s,S^h)=\prod_{i=1}^{n}p(x_i|pa_i,\theta_i,S^h), \tag{3}$$

其中 θ_i 是分布 $p(x_i|pa_i,\theta_i,S^h)$ 的参数向量，θs 是参数组$(\theta_1,\theta_2,\cdots,\theta_n)$构成的向量，而 S^h 表示物理联合分布可以依照 S 分解的假设。将分布 $p(x_i|pa_i,\theta_i,S^h)$ 看成的 θ_i 函数，并称为局部分布函数。局部分布函数其实只是一个概率分类或回归函数，在离散变量情形是分类，在连续变量情形是回归。于是，贝叶斯网络可以看成由条件独立关系组成的概率分类 / 回归模型的集合。如线性回归、扩展的线性回归、概率神经网络、概率决策树等，都是该集合的例子。在大多数情形，都可以用贝叶斯方法进行学习。

3　贝叶斯网络的参数学习

现在考虑这样的问题：给定贝叶斯网络的结构，如何利用给定样本数据去学习网络的参数的概率分布，即更新网络变量原有的先验分布。我们假设变量组 $X=(X_1,X_2,\cdots,X_n)$ 的物理联合概率分布可以编码在某个网络结构 S 中：

$$p(x|\theta s,S^h)=\prod_{i=1}^{n}p(x_i|pa_i,\theta_i,S^h), \tag{4}$$

其中各项的解释与(3)式相同。S^h 表示物理联合分布可以依照 S 被分解的假设。此外，假设我们从 X 的联合概率分布得到一个随机样本 $D=\{X_1,\cdots,X_n\}$。D 的一个元素 X_l 表示样本的一个观测值，称为一个例子或案例。我们定义一个取向量值的变量 Θs 表示对于参数向量 θs 的不确定性，并指派一个先验概率密度函数 $p(\theta s|S^h)$。于是在贝叶斯网络中对参数的学习问题可以简单地表示成：给定随机样本 D，计算后验分布 $p(\theta s|D,S^h)$。

3.1　具有完整数据的贝叶斯网络的学习

我们采用无约束多项分布来讨论学习参数的基本思想。假定每个变量 $X\in X_n$ 是离散的，有 r_i 个可能的值 $x_i^1,x_i^2,\cdots,x_i^{ir}$，每个局部分布函数是一组多项分布的集合，一个分布对应于 pa_i 的一个构成(即一个分量)。也就是说，假定

$$p(x_i^k|pa_i^j,\theta_i,S^h)=\theta_{ijk}>0\ (i=1,2,\cdots,n;j=1,2,\cdots,q_i;k=1,2,\cdots,r_i), \tag{5}$$

其中 $pa_i^1,pa_i^2,\cdots,pa_i^{qi}$ 表示 pa_i 的构成，$q_i=\prod_{Xi\in Pai}r_i$。$\theta_i=((\theta_{ijk})_{k=2}^{rl})_{j=1}^{qi}$ 是参数，θ_{ij1} 没有列入，因为

$\theta_{ij1} = 1 - \sum_{k=1}^{r_j} \theta_{ijk}$，可以通过计算得到。为方便起见，定义参数向量：

$$\theta_{ij} = (\theta_{ij2}, \theta_{ij3}, \cdots, \theta_{ijr_i}) \ (i = 1, 2, \cdots, n; j = 1, 2, \cdots, q_i).$$

给定以上的局部分布函数后，在以下两个假设之下，我们可以以封闭的形式有效地计算后验分布 $p(\theta s | D, S^h)$：

1）在随机样本 D 中没有缺损数据，这时又称 D 是完全的；

2）参数向量 θ_{ij} 量是相互独立的，即所谓参数独立假设，也就是说：

$$p(\theta s | S^h) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\theta_{ij} | S^h).$$

在以上两个假设下，对于给定的随机样本 D，参数仍然保持独立：

$$p(\theta s | D, S^h) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\theta_{ij} | D, S^h),$$

于是我们可以像一个变量的情形那样相互独立地更新每一个参数向量 θ_{ij}。假设每一个参数向量 θ_{ij} 有先验 Dirichlet 分布 $\mathrm{Dir}(\theta_{ij} | \alpha_{ij1}, \alpha_{ij2}, \cdots, \alpha_{ijr_i})$，我们得到后验分布：

$$p(\theta_{ij} | D, S^h) = \mathrm{Dir}(\theta_{ij} | \alpha_{ij1} + N_{ij1}, \alpha_{ij2} + N_{ij2}, \cdots, \alpha_{ijr_i} + N_{ijr_i}), \tag{6}$$

其中 N_{ijk} 是当 $X_i = x_i^k$ 且 $Pa_i = pa_i^j$ 时 D 中的案例数目。由于无约束多项分布属于指数家族，上面的计算将变得简单。

3.2 具有不完整数据的贝叶斯网络学习

仍以无约束多项分布为例来讨论具有不完整数据的统计学习问题。设 $Y \subset X$、$Z \subset X$ 分别表示观察到的和未观察到的变量，假定缺损的数据与变量的状态无关，并假定参数相互独立，可用如下的式子来计算参数 θs 对于不完整数据 Y 的后验分布：

$$p(\theta s | Y, S^h) = \sum_{z} p(Z | Y, S^h) p(\theta s | Y, Z, S^h).$$

一般来说，对于任意的局部分布和先验分布，要精确计算参数 θs 后验分布是比较困难的，此时常借助于一些近似方法来处理。

4 具有不确定结构的贝叶斯网络学习

如果对网络的结构也不能确定，那么也可能通过学习来获得网络结构，建立网络。

首先假定网络结构是可以改进的。按照贝叶斯方法，定义一个离散变量表示我们对于网络结构的不确定性，其状态对应于可能的网络结构假设 S^h，并赋予先验概率分布 $p(S^h)$。给定随机样本 D（D 来自 X 的物理概率分布），参数后验概率分布 $p(\theta s | D, S^h)$ 的计算方法与上一节类似。结构后验分布 $p(S^h | D)$ 的计算至少在原理上是简单的。根据贝叶斯定理有：

$$p(S^h | D) = p(S^h, D)/p(D) = p(S^h) p(D | S^h)/p(D), \tag{7}$$

其中 $p(D)$ 是一个与结构无关的正规化常数，$p(S^h)$ 是结构先验，$p(D | S^h)$ 是边界似然。

在指数分布族、参数独立、参数共轭先验和完整数据的条件下，边界似然可以封闭的形式有效地计算。在无约束多项分布、参数独立、采用 Dirichlet 先验和数据完整的前提下，参数向量 θ_{ij} 可以独立地更新。数据的边界似然正好等于每一个 $i - j$ 对的边界似然的乘积：

$$p(D | S^h) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}, \tag{8}$$

对网络结构不确定的贝叶斯网络学习的主要困难也在于计算量的繁重，也常采用近似计算。

5　学习实例

下面是一个使用贝叶斯网络进行数据采掘和知识发现的应用实例[3]。数据来自华盛顿高级中学的 10 318 名高年级学生。每个学生用表 1 中所列变量及其应的状态来描述。

表 1　变量及性状描述

变　量	性状描述	变　量	性状描述
性别(SEX)	男、女	升学计划(CP)	是、否
社会经济状态(SES)	低、中下、中上、高	家长的鼓励(PE)	低、高
智商(IQ)	低、中下、中上、高		

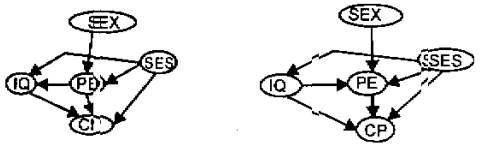
目的是从数据中发现影响高中学生上大学意向的因素，即存在于这些变量之间的可能的因果关系。数据已经整理成如表 2 的充分统计量。

表 2　充分统计量

男生	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	166	47	91	6	120	74	110	17	92	148	100	6	42	198	73
	4	48	39	57	5	47	132	90	9	41	224	65	8	17	414	54
女生	5	454	9	44	5	312	14	47	8	216	20	35	13	96	28	24
	11	285	29	61	19	236	47	88	12	164	62	85	15	113	72	50
	7	163	36	72	13	193	75	90	12	174	91	100	20	81	142	77
	6	50	36	58	5	70	110	76	12	48	230	81	13	49	360	98

表中每个数据表示：对于 5 个变量取值的某种组合（构成）统计所得到的人数。例如，第一个数据表示对（SEX＝男，SES＝低，IQ＝低，PE＝低，CP＝是）这种组合统计得到的人数为 4 人。第二个数据则表示对（SEX＝男，SES＝低，IQ＝低，PE＝低，CP＝否）这种组合统计得到的人数为 349 人。其后的数据依次表示轮换每个变量可能的状态统计得到的人数。变量依照从右到左的顺序轮换，状态则按照表 1 列出各变量状态的顺序轮换。如此等等，依次类推。

首先假定没有隐藏变量。为了生成网络参数的先验，使用容量为 5 的等价样本和一个带有一致的 $p(X|S_c^h)$ 的先验网络。除了排除的 SEX 和 SES 有父节点、CP 有子节点的结构之外，假定所有的网络结构都同等可能。因为数据集是完整的，可以用（7）式、（8）式来计算网络结构的后验概率。进行所有网络结构的穷举搜索后，发现两个最相似的网络结构如图 1。请注意，最相似的结构的后验概率也是极端接近的。如果采纳因果马尔可夫假设，并假设没有隐藏变量，则在两个图中的弧都可以有因果的解释。其中一些结果，例如社会经济状态和智商对升学愿望的影响，并不使人意外。另一些结果更有趣：从两个图中都可以得到性别对升学愿望的影响仅是不直接地通过父母的影响体现出来。此外，两个图的不同仅在于 PE 和 IQ 之间的弧的方向。两个不同的因果关系似乎都是有道理的。右边的网络曾经由 Spirtes 等[4]用非贝叶斯方法于 1993 年选出。



$\text{Log}\ p(D|S_1^h) \approx -45\,653$　　$p(S_1^h|D) \approx 1.0$

$\text{Log}\ p(D|S_2^h) \approx -45\,699$　　$p(S_2^h|D) \approx 1.2 \times 10^{-10}$

图 1　没有隐藏变量的后验最可能的网络结构

最值得怀疑的结果是：社会经济状况对智商有直接的影响。为了考证这个结果，将图 1 中原来模型的直接影响用一个指向 SES 和 IQ 的隐藏变量代替。考虑这样的模型的各种最可能的情况，使用 Laplace 逼近的 Cheeseman-Stutz 变体[5]计算这些模型的后验概率。使用 EM 算法，并在带有不同的随机初始化的 θ_S 的100 次运行中取最大局部极大，来找最大后验 $\bar{\theta}_S$。结果证实确有一个隐藏变量。分析图 2 的概率可知，隐藏变量对应于“家长的素质”。

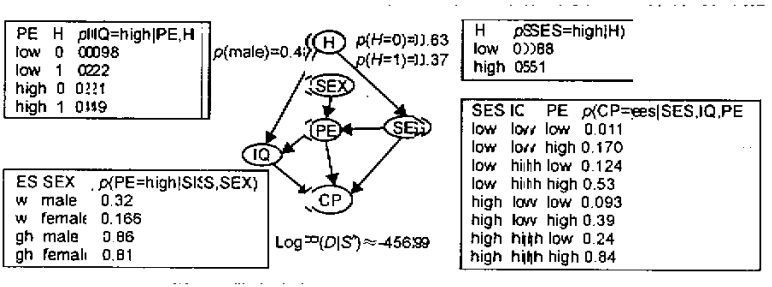


图 2 带有隐藏变量的后验最可能的网络结构

6 结语

与非贝叶斯方法相比，贝叶斯方法的特出特点是其学习机制可以综合先验信息和后验信息，既可避免只使用先验信息可能带来的主观偏见，和缺乏样本信息时的大量盲目搜索与计算，也可避免只使用样本信息带来的噪音的影响。只要合理地确定先验，就可以进行有效的学习。因此，适用于具有概率统计特征的数据采掘和机器发现问题，尤其是样本难得的问题。贝叶斯方法遇到的一个重大的问题是先验分布的确定依据的只是一些准则，没有可操作的完整的理论，在许多情况下先验分布的合理性和准确性难以评价。将贝叶斯方法用于学习，其学习机制的基础理论还有待进一步深入研究。

参考文献

1 Heckerman D. Bayesian networks for data mining. Data Mining and Knowledge Discovery, 1997, 1: 79～119.

2 Heckerman D, Geiger D, Chickering D. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning, 1995, 20: 196～243.

3 Sewell W, Shah V. Social class, parental encouragement and educational aspirations. American Journal of Sociology, 1968, 73: 559～572.

4 Spirtes P, Glymour C, Scheines R. Causation, Predication and Search. New York: Springer-Verlag, 1993.

5 Cheeseman P, Stutz J. Bayesian classification (AutoClass): theory and results. Advances in Knowledge Discovery and Data Mining, Menlo Park, CA: AAAI Press, 1995.

6 林士敏，王双成，陆玉昌. 贝叶斯方法的学习机制与问题求解. 清华大学学报, 2000, 40 (9): 61～64.

（责任编辑：黎贞崇）