计算机科学2000Vol.27№.10

贝叶斯学习、贝叶斯网络与数据采掘*)

Bayesian Networks Construction and Their Applications in Data Mining

林士敏 田凤占 陆玉昌

(清华大学计算机科学与技术系 智能技术与系统国家重点实验室 北京100084)

Abstract Recently Bayesian networks(BN) become a noticeable research direction in Data Mining. In this paper we introduce the structure of Bayesian networks, and the process of constructing a BN, with the emphasis on the basic methods of learning from prior knowledge and sample data, using Bayesian learning approach, to identify the structures and probabilities of BN. The merits of Bayesian networks are that prior knowledge can be combined with observed data, which is important especially when data is scarce or expensive, that causal relationships among data can be learned, and incomplete data set can be readily handled, which other models are disable to do so. It can foresee that Bayesian networks will become a powerful tools in Data Mining.

Keywords Bayesian networks, Data mining, Knowledge discovery, Machine learning

自从50～60年代贝叶斯学派形成后，关于贝叶斯分析的研究久盛不衰。早在80年代，贝叶斯网络就成功地应用于专家系统，成为表示不确定性专家知识和推理的一种流行方法。90年代以来，贝叶斯学习一直是机器学习研究的重要方向。由于概率统计与数据采掘的天然联系，数据采掘兴起后，贝叶斯网络日益受到重视，再次成为引人注目的热点。近两年研究者们进一步研究了直接从数据中学习并生成贝叶斯网络的方法，包括贝叶斯方法、类贝叶斯方法和非贝叶斯方法，为贝叶斯网络用于数据采掘和知识发现开辟了道路。这些新的方法和技术还在发展之中，但是已经在一些数据建模问题中显示出令人瞩目的效果。

1 贝叶斯网络的结构及建立方法

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：(1)一个表示 X 中的变量的条件独立断言的网络结构 S；(2)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：

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

P 表示(1)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)。

为了建立贝叶斯网络，第一步，必须确定与建立模型有关的变量及其解释。为此，需要：(1)确定模型的目标，即确定问题相关的解释；(2)确定与问题有关的许多可能的观测值，并确定其中值得建立模型的子集；(3)将这些观测值组织成互不相容的而且穷尽所有状态的变量。这样做的结果不是唯一的。

第二步，建立一个表示条件独立断言的有向无环图。根据概率乘法公式有：

$$p(x)=\prod_{i=1}^{n} p(x_i|x_1,x_2,\cdots,x_{i-1})$$

*)国家重点基础研究发展计划项目、国家自然科学基金项目、"九五"国家攀登计划预选项目。林士敏 副教授，访问学者，研究方向：机器学习、数据采掘与知识发现。田凤占 博士研究生，研究方向：数据采掘与知识发现。陆玉昌 教授，研究方向：数据采掘与知识发现，机器学习，知识工程。

$$=p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)\cdots p(x_n|x_1,x_2,\cdots,x_{n-1}) \quad (2)$$

对于每个变量 X_i，如果有某个子集 $\Pi_i\subseteq\{X_1,X_2,\cdots,X_{i-1}\}$ 使得 $X_i|X_1,X_2,\cdots,X_{i-1}\backslash\Pi_i$ 是条件独立的，即对任何 X，有

$$p(x_i|x_1,\cdots,x_{i-1})=p(x_i|\pi_i) \quad (i=1,2,\cdots,n) \quad (3)$$

则由(2)(3)两式可得：$p(x)=\prod_{i=1}^{n}p(x_i|\pi_i)$。变量集合 $(\Pi_1,\cdots,\Pi_n)$ 对应于父节点 $(Pa_1,\cdots,Pa_n)$，故又可以写成：$p(x)=\prod_{i=1}^{n}p(x_i|Pa_i)$，于是，为了决定贝叶斯网络的结构，需要：(1)将变量 $X_1,X_2,\cdots,X_n$ 按某种次序排序；(2)决定满足(3)式的变量集 $\Pi_i(i=1,2,\cdots,n)$。

从原理上说，如何从 n 个变量中找出适合条件独立的顺序，是一个组合爆炸问题，因为要比较 n! 种变量顺序。不过，通常可以在现实问题中决定因果关系，而且因果关系一般都对应于条件独立断言。因此，可以从原因变量到结果变量划一个带箭头的弧来直观表示变量之间的因果关系。

第三步，指派局部概率分布 $p(x_i|Pa_i)$，在离散的情形，需要为每一个变量 X_i 的各个父节点的状态指派一个分布。

显然，以上各步可能交叉进行，而不是简单的顺序进行可以完成的，因为网络的结构和参数都是根据背景知识和经验确定的，这样建立的网络又称为先验贝叶斯网络。

2 贝叶斯网络的语义

(1)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：

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

(2)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。

(3)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：

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

其中 θ_i 是分布 $p(x_i|pa_i,\theta_i,S^h)$ 的参数向量，θ_S 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 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 贝叶斯网络的推断

关于变量组 X 的贝叶斯网络表示了 X 的联合概率分布，所以，无论是从先验知识、数据或两者的综合建立的贝叶斯网络，原则上都可以用它来推断任何感兴趣的概率。

下面看一个简化的例子。考虑如何发现信用卡使用中的骗局问题。首先决定模型的变量，假定取5个变量：

变量名	意义
F(fraud)	是否当前的一笔买卖是骗局
G(gas)	是否在24小时中有一笔汽油买卖
J(jewelry)	是否在24小时中有一笔珠宝买卖
A(age)	信用卡持有者的年龄
S(sex)	信用卡持有者的性别

利用关于变量因果关系的先验知识分析有关数据和变量之间的关系后，决定变量的顺序为：(F,A,S,G,J)，并决定变量之间的条件独立关系：

$$p(a|f)=p(a)$$
$$p(s|f,a)=p(s)$$
$$p(g|f,a,s)=p(g|f)$$
$$p(j|f,a,s,g)=p(j|f,a,s)$$

据此得到网络结构。最后，为每一个变量指派局部概率分布，就得到一个如图1的完整的贝叶斯网络。

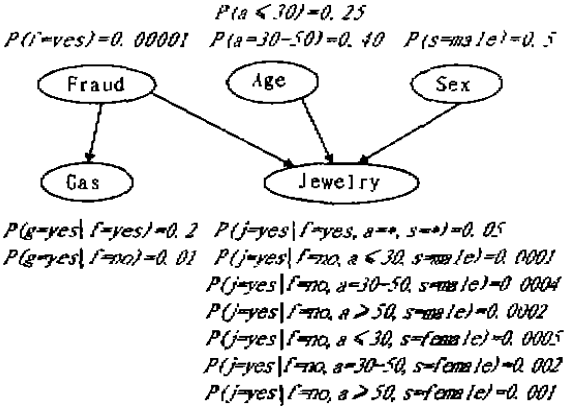

图1 侦测信用卡骗局的贝叶斯网络

已知其他变量的观测值，假如要想知道骗局发生的概率，就可以通过已经建立的贝叶斯网络计算推导

出来。从图1可知，所求的概率可以表示为：

$$p(f|a,s,g,j)=\frac{p(f,a,s,g,j)}{p(a,s,g,j)}=\frac{p(f,a,s,g,j)}{\sum_{f'}p(f',a,s,g,j)} \quad (4)$$

其中 f' 表示 f 所有可能的状态。在一般的多变量问题，以上直接计算的方法往往是困难的。不过，利用已经确定的条件独立关系，上式变为：

$$p(f|a,s,g,j)=\frac{p(f)p(a)p(s)p(g|f)p(j|f,a,s)}{\sum_{f'}p(f')p(a)p(s)p(g|f')p(j|f',a,s)}$$

$$=\frac{p(f)p(g|f)p(j|f,a,s)}{\sum_{f'}p(f')p(g|f')p(j|f',a,s)} \quad (5)$$

此时计算已得到简化。

1990年与1994年 Cooper、Dagum 和 Luby 已经分别证明，离散变量的任意贝叶斯网络的精确或近似推断（比如使用 Monte-Carlo 方法）都是 NP 难题。目前的解决办法是使用条件独立和一些技巧以简化计算，或面向特定的推断要求建立简单的网络拓扑，或在不牺牲太多精确性的前提下，简化网络的结构等。显然如此，一般仍然需要可观的计算时间。对某些问题如朴素贝叶斯分类器使用条件独立则有明显的效果。

4 学习贝叶斯网络

学习贝叶斯网络指的是利用样本数据更新网络原有参数或结构的先验分布。比较简单的问题是：给定贝叶斯网络的结构，利用给定样本数据学习网络的参数（概率分布）。更为复杂的问题是：网络的结构也没有确定，利用给定样本数据学习网络的结构和参数。由于数据采掘面对的是大量数据，一时往往难以断定变量之间的关系，因此后一个问题更具有现实意义。

学习贝叶斯网络使用的是贝叶斯学习的方法，即利用贝叶斯定理综合先验信息和样本数据去改善已有知识的技术，这些技术也用于数据采掘的其他问题。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：

$$p(x|\theta_S,S^h)=\prod_{i=1}^{n}p(x_i|pa_i,\theta_i,S^h) \quad (6)$$

其中 θ_i 是分布 $p(x_i|pa_i,\theta_i,S^h)$ 的参数向量，θ_S 是参数组 $(\theta_1,\theta_2,\cdots,\theta_n)$ 的向量，而 S^h 表示物理联合分布可以依照 S 被分解的假设。此外，假设从 X 的物理联合概率分布得到一个随机样本 $D=\{X_1,\cdots,X_m\}$。D 的一个元素 X_i 表示样本的一个观测值，称为一个案例。定义一个取向量值的变量 Θ_S 对应于参数向量 θ_S，并指派一个先验概率密度函数 $p(\theta_S|S^h)$ 表示对 Θ_S 的不确定性。于是贝叶斯网络的参数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 $p(\theta_S|D,S^h)$。

假定每个变量 $X\in X_i$ 是离散的，有 r_i 个可能的值 $x_i^1,x_i^2,\cdots,x_i^{r_i}$，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

$$p(x_i^k|pa_i^j,\theta_i,S^h)=\theta_{ijk}>0$$
$$(i=1,2,\cdots,n;j=1,2,\cdots,q_i;k=1,2,\cdots,r_i) \quad (7)$$

其中 $pa_i^1,pa_i^2,\cdots,pa_i^{q_i}$ 表示 pa_i 的构成，$q_i=\prod_{X_l\in Pa_i}r_l$，$\theta_i=((\theta_{ijk})_{k=2}^{r_i})_{j=1}^{q_i}$ 是参数，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，在以下两个假设下，可以封闭地计算后验分布 $p(\theta_S|D,S^h)$：

（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} 是相互独立的，即 $p(\theta_S|S^h)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}p(\theta_{ij}|S^h)$，这就是参数独立假设。

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：

$$p(\theta_S|D,S^h)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}p(\theta_{ij}|D,S^h) \quad (8)$$

于是可以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[3] $Dir(\theta_{ij}|\alpha_{ij1},\alpha_{ij2},\cdots,\alpha_{ijr_i})$，经过计算得到后验分布，从而求得问题的解为：

$$p(\theta_{ij}|D,S^h)=Dir(\theta_{ij}|\alpha_{ij1}+N_{ij1},\alpha_{ij2}+N_{ij2},\cdots,\alpha_{ijr_i}+N_{ijr_i}) \quad (9)$$

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。

当样本数据不完全时，除了少数特例外，一般要借助于近似方法，如 Monte-Carlo 方法，Gaussian 逼近，以及 EM（期望-极大化）算法求 ML（极大似然）或 MAP（极大后验）等。尽管有成熟的算法，其计算开销也是比较大的。

当不能确定贝叶斯网络的结构时，用贝叶斯学习的方法从给定数据学习网络的结构和概率分布也是可能的。

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。然后计算后验概率分布 $p(S^h|D)$ 和 $p(\theta_S|D,S^h)$，其中 θ_S 是参数向量。

$p(\theta_S|D,S^h)$ 的计算方法与上一节类似。$p(S^h|D)$ 的计算至少在原理上是简单的。根据贝叶斯定理有：

$$p(S^h|D)=p(S^h,D)/p(D)$$
$$=p(S^h)p(D|S^h)/p(D) \quad (10)$$

其中 $p(D)$是一个与结构无关的正规化常数，$p(D|S^h)$是边界似然。于是确定网络结构的后验分布只需要为每一个可能的结构计算数据的边界似然。

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：

$$p(D|S^h)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\cdot\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})} \qquad (11)$$

该公式首次为 Cooper 和 Herskovits 于1992年给出。

在一般情况下，n 个变量的可能的网络结构数目大于以 n 为指数的函数。逐一排除这些假设是困难的，可以使用两个方法来处理这个问题："模型选择"和"选择模型平均"方法。前一个方法是从所有可能的模型（结构假设）中选择一个"好的"模型，并把它当作正确的模型。后一个方法是从所有可能的模型中选择合理数目的"好"模型，并认为这些模型代表了所有情况。关于"好模型"已有一些不同的定义和相应的计算方法（例如使用评分函数）。若干研究者的工作表明，使用贪心搜索法选择单个好的假设通常会得到准确的预测。使用 Monte-Carlo 方法进行模型平均有时也很有效，甚至可以得到更好的预测。这些结果多少可以算是对目前用贝叶斯网络进行学习的莫大兴趣的回答。

Heckerman 于1995年提出，在参数独立、参数模块性、似然等价以及机制独立、部件独立等假设成立的前提下，可以将学习贝叶斯非因果网络的方法用于因果网络的学习。1997年又提出在因果马尔科夫条件下，可以由网络的条件独立和条件相关关系推断因果关系。这使得在干涉（扰动）出现时可以预测其影响。这是用贝叶斯网络进行因果分析的尝试。

小结 与其它用于数据采掘的表示法如规则库、决策树、人工神经网络相比，贝叶斯网络有如下特点：(1)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，也可避免只使用后验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，这在样本难得或者代价高昂时特别有用。(2)适合处理不完整数据集问题。(3)可以发现数据间的因果关系。后两点在实际问题中经常遇到，而且是用其他模型难以处理的。(4)有成熟有效的算法。虽然任意贝叶斯网络的概率推断是 NP-难题，但是很多问题加上一些限制后计算就可以简化，有些问题有近似解法，不过，贝叶斯网络的计算量较大，在某些其他方法也可以解决的问题求解中显得效率较低，先验密度的确定虽然已经有一些方法，但对具体问题要合理确定许多变量的先验概率仍然是一个较困难的问题，而这在样本难得时却特别重要。此外，贝叶斯网络需要多种假设为前提，如何判定某个实际问题是否满足这些假设，没有现成的规则，这给实际应用带来困难。这些都是需要进一步研究的问题。尽管如此，可以预见，在数据采掘和知识发现中，尤其在具有概率统计特征的数据采掘中，贝叶斯网络将成为一个有力的工具。

参 考 文 献

1 Heckerman D. Bayesian networks for data mining[J]. 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[J]. Machine Learning, 1995, 20: 196～243

3 Geiger D, Heckerman D. A characterization of the Dirichler distribution with application to learning Bayesian networks[A]. In: Proc. of Eleventh Conf. on Uncertainty in Artificial Intelligence[C], Montreal, QU, 1995. 196～207

4 Dagum P, Luby M. Approximating probabilistic inference in Bayesian belief networks is NP-hard[J]. Artificial Intelligence, 1993, 60: 141～153

5 Cooper G, Herskovits E. A Bayesian method for the induction of probabilistic networks from data [J]. Machine Learning, 1992, 9: 309～347

6 Chickering D. Learning equivalence classes of Bayesian-network structures[A]. In: Proc. of Twelfth Conf. on Uncertainty in Artificial Intelligence[C]. 1996

7 Heckerman D, Mamdani A, Wellman M. Real-world applications of Bayesian networks [J]. Communications of ACM, 1995, 38

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

9 Spirtes P, Glymour C, Scheines R. Causation, Predication, and Search[M]. New York: Springer-Verlag, 1993

10 Cheeseman P, Stutz J. Bayesian classification (AutoClass): Theory and results[A]. In: Fayyad U, et al, eds. Advances in Knowledge Discovery and Data Mining [C], Menlo Park, CA: AAAI Press, 1995