第 23 卷 第 4 期
2002 年 4 月

小 型 微 型 计 算 机 系 统
MINI- MICRO SYSTEM

Vol. 23 No. 4
Apr. 2002

文 章 编 号: 1000-1220(2002)04-0435-03

贝叶斯网络拓扑结构确定方法的研究

王 玮 蔡莲红

(清华大学 计算机科学技术系, 北京 100084)

摘 要: 贝叶斯方法是概率统计学中一种很重要的方法, 贝叶斯网络是一种将贝叶斯概率方法和有向无环图的网络拓扑结构有机结合的表示模型, 描述了数据项及其依赖关系, 并根据各个变量之间概率关系建立的图论模型, 但是如何获取它的网络拓扑结构是一个需要解决的问题, 本文提出一种如何寻找最有可能的贝叶斯网络模型方法, 并用启发式算法进行模型评估.

关 键 词: 贝叶斯概率; 贝叶斯网络; 拓扑结构
中图分类号: TP311 文献标识码: A

1 引 言

贝叶斯学派是数理统计学中的一大流派, 贝叶斯概率与传统概率之间最大的区别在于他们对某个事件概率的定义是不一样的, 经典概率定义一个事件的概率是确定的, 并且是客观的, 而贝叶斯概率认为, 一个事件的概率是确定这个概率的人的主观判断, 即传统概率是客观认识, 而贝叶斯概率是主观判断.

贝叶斯网络是一种将贝叶斯概率方法和有向无环图的网络拓扑结构有机结合的表示模型, 描述了数据库中数据项及其相互之间的依赖关系. 贝叶斯网络就是根据各个变量之间的概率关系建立起来的图论模型, 随着近年来数据库规模的不断扩大, 研究发现应用贝叶斯方法进行数据库的知识发现要优于目前应用在数据库的知识发现的方法, 这主要表现在以下几个方面:

1. 贝叶斯网络能够处理不完备数据集. 这是传统的指导性学习方法所无法解决的问题, 对于一般的指导性学习方法而言, 必须知道所有可能的数据输入, 如果缺少其中的某一输入就会对建立的模型产生偏差. 贝叶斯方法则可以解决这个问题, 因为贝叶斯网络反映的是整个数据域中数据间的概率关系, 即使缺少某一数据变量仍然可以建立精确的模型.

2. 贝叶斯网络是根据因果关系进行学习的. 在数据分析处理中获得变量域的理解是十分重要的, 而且贝叶斯网络可以在缺少插入值的情况下进行决策.

3. 贝叶斯网络和贝叶斯统计是紧密相关的, 这促进了知识和数据域之间的关联关系. 通常必须在知道处理数据域的先验知识的基础上才能建立正确的预测模型, 由于贝叶斯网络具有语义的因果关系因而可以直接地进行因果先验知识的分析, 所以在贝叶斯网络中可以获得较全面的先验知识.

4. 贝叶斯网络可以有效地避免数据溢出的情况.

综上所述, 可以看出贝叶斯网络具有一般数据挖掘方法所不具备的特点, 因此如何获取它的网络拓扑结构是一个值得研究的问题. 本文通过对贝叶斯网络知识的研究寻找出最有可能的贝叶斯网络模型及其延伸方法, 并提出一种启发式算法进行模型评估.

2 贝叶斯方法

贝叶斯网络就是依据马尔可夫假设寻找的满足条件独立限制的模型结构. 举例来说, 对于给定两个数据变量域 X 和 Y 判断它们是否绝对独立, 可以用它们之间的概率来反映其独立性. 而且, 由于贝叶斯方法用的是概率网络, 无需对变量域中的个别变量的独立性进行判定, 只需根据变量间的概率分布来计算整个网络模型的独立性.

贝叶斯方法的模型描述如下:

变量域 $X=\{X, \cdots, X_n\}$, 数据域 $D=\{x_1, \cdots, x_N\}$, 数据库中的数据是从变量域 X 中随机抽取的概率分布. 若 D 中的每个 x 包含变量 X 中所有可能的情况, 未知的概率分布能够用某些因果模型结构加以描述, 这些假定的因果模型是根据马尔可夫条件建立的有向无环图, 但是无法确定其模型参数和结构, 应用贝叶斯方法可以变量之间的概率关系确定模型的参数和结构.

定义 1: 离散变量 M 的 m 个状态对应于 m 种可能的模型结构, 用 $p(m)$ 表示其概率分布.

定义 2: 对于每一种模型结构 m, 存在一个连续向量值变量 θ_m, 其中 θ_m 值对应的是可能模型的真实参数.

定义 3: 对于不确定的 θ_m 使用概率密度函数 $p(\theta_m|m)$ 进行编码.

给定随机样本数据域 D, 对每个 m 和 θ_m 使用贝叶斯规则计算后验分布如下:

$$p(m|D)=\frac{p(m)p(D|m)}{\sum_{m'}p(m')p(D|m')} \quad (1)$$

$$p(\theta_m|D,m)=\frac{p(\theta_m|m)p(D|\theta_m,m)}{p(D|m)} \quad (2)$$

收稿日期: 2000-12-21 基金项目: 国家自然科学基金(69875008)资助. 作者简介: 王 玮, 博士后, 主要研究领域为数据挖掘, 语音合成. 神经网络. E-mail: cliff- wang@263.net; 蔡莲红, 教授, 博士生导师, 主要研究领域为多媒体信息处理.

其中 $p(D|m)=\int p(D|\theta_m,m)p(\theta_m|m)\mathrm{d}\theta_m$.

给定需求假设 h(如 X 会产生 Z 的可能性), 平均所有可能模型和参数如下:

$$p(h|D)=\sum_m p(m|D)p(h|D,m) \tag{3}$$

$$p(h|D,m)=\int p(h|\theta_m,m)p(\theta_m|D,m)d\theta_m \tag{4}$$

公式中的 $p(h|\theta_m,m)$ 是模型的可能性, 在特定的假设条件下, 这些计算是很有效的并且有闭形式. 假设 $p(x|\theta_m,m)$ 因子如下:

$$p(x|\theta_m,m)=\prod_{i=1}^{n}p(x_i|pa_i,\theta_i,m) \tag{5}$$

式中 $p(x_i|pa_i,\theta_i,m)$ 是幂级数, pa_i 表示对应变量 x_i 的父变量的构型, θ_i 表示对变量 x_i 的局部可能性参数, 当每个 $X_i\in X$ 是离散值时, 共有 r_i 种可能的值 $x_i^1,\cdots\cdots,x_i^{r_i}$, 每种局部可能性是多项式的集合, 对于每种构型 pa_i 的一种分布是:

$$p(x_i^k|pa_i^j,\theta_i,m)=\theta_{ijk>0} \tag{6}$$

其中 $pa_i^1,\cdots,pa_i^{q_i}$ $\left(q_i=\prod_{x_i=pa_i}r_i\right)$ 表示 pa_i 的构型, $\theta_i=((\theta_{ijk})_{k=2}^{r_i})_{j=1}^{q_i}$ 是参数. 参数 θ_{ij1} 可定义如下:

$$\theta_{i1}=1-\sum_{k=2}^{r_i}\theta_{ijk} \tag{7}$$

定义 4: 参数向量 $\theta_{ij}=(\theta_{ij2},\cdots,\theta_{ijr_i})\ \forall\ i,j$;

假设离散的二项式分布的参数向量 $\theta_{i,j}$ 是相互独立的, 则对于给定的随机抽样完备数据域 D, 其中参数相互独立, 有:

$$p(\theta_m|D,m)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}p(\theta_{ij}|D,m) \tag{8}$$

因此根据式(8)独立更新参数向量 θ_{ij}, 若每种向量 θ_{ij} 具有 Dirichlet 分布 $Dir(\theta_{ij}|\alpha_{ij1},\cdots,\alpha_{ijr_i})$, 可以得到参数的后验分布公式:

$$p(\theta_{ij}|D,m)=Dir(\theta_{ij}|\alpha_{ij1}+N_{ij1},\cdots,\alpha_{ijr_i}+N_{ijr_i}) \tag{9}$$

这里的 N_{ijk} 是数据域 D 的 case 数, $X_i=x_i^k$, $Pa_i=pa_i^k$, N_{ijk} 是对模型 m 的统计数据, 于是得到:

$$p(D|m)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{10}$$

式中的 $\alpha_{ij}=\sum_{k=1}^{r_i}\alpha_{ijk}$, $N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$, 根据公式(1)和公式(10)就可以计算出后验概率 $p(m|D)$.

3 算法设计

这里我们讨论在给定了一个数据库以后, 如何找到"最有可能"的贝叶斯网络结构 B_s, 即找到的 B_s 能正确反映数据库的域之间的关系时, 它就是最有可能的.

计算出任意两个 $P(B_{S_i}|D)$ 和 $P(B_{S_j}|D)$ 之间的比率, 那么就可以对它们进行排序, 公式(11)只要计算出 $P(B_{S_i}|D)$ 和 $P(B_{S_j}|D)$ 就可以得到上述的比率, 从而可以得到 $P(B_S|D)$ 的排序, 进而得到最大可能的 B_S.

$$\frac{P(B_{S_i}|D)}{P(B_{S_j}|D)}=\frac{\frac{P(B_{S_i},D)}{P(D)}}{\frac{P(B_{S_J},D)}{P(D)}}=\frac{P(B_{S_i},D)}{P(B_{S_J},D)} \tag{11}$$

计算 $P(B_S|D)$ 的公式(12)如下:

$$P(B_S,D)=P(B_S)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}! \tag{12}$$

根据公式(12)求最大可能 B_S 的公式, 并应用一个优化算法以求出相应于一个给定数据库的最大可能网络结构.

本文采用的算法是一种优化算法, 其算法基础是公式(13):

$$\max_{B_S}\{P(B_S,D)\}=c\prod_{i=1}^{n}\max_{\pi_i}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r-1)!}\prod_{k=1}^{r_i}N_{ijk}! \tag{13}$$

我们假设各个节点是有顺序的, 并且所有的网络结构都等同, 对公式(13)的修改是求最大值的方法, 首先假设节点的父节点集合是空集, 然后向该项集合中加入一个能使结果的概率最大的父节点, 此过程一直加到父节点已经不再使结果增加为止, 于是我们就得到一个可能的网络结构.

定义函数 $g(r_i,\pi_i)$ 表达式为

$$g(r_i,\pi_i)=\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)}\prod_{k=1}^{r_i}N_{ijk}! \tag{14}$$

算法设计如下:

```
1.  for(i=1; i<n; i++);
2.      πi=φ;
3.      Pold=g(i,π), 其中 g(ri,πi)=∏_{j=1}^{qi} (ri-1)!/(Nij+ri-1) ∏_{k=1}^{ri} Nijk! ;
4.      set Proceed=true;
5.      While (Proceed &&|∏i|<s)do;
6.          z=Pred(xi)-πi;
7.          Pnew=g(i,πi∪{z});
8.          if(Pnew>Pold)then;
9.              Pold=Pnew;
10.             πi=πi∪{z};
11.         else Proceed=false;
12.     end(while);
13.     Print f(xi,πi);
14.  end(for)
```

下面对这个算法的实现的复杂进行一些分析:

首先假设公式(13)中的所有阶乘都已经预算法并存入一个数组中, 因为 N_{ij} 不可能大于 m, 因此公式(14)中的阶乘没有大于 $(m+r-1)!$ 的, 即我们可以在 $O(m+r-1)$ 的时间里把 1 到 $m+r-1$ 的整数阶乘计算并存储到数组中, 其次算法中的函数 g 最多被调用 $n-1$ 次, 因为 x_i 最多有 $n-1$ 个父节点, 因此 P_{new} 最多需要 $O(m,u,n,r)$ 的时间就可以完成所有的操作. While 循环中的语句是以 $O(1)$ 的时间执行的, 每次进入 while 循环, 需要循环的次数是 $O(u)$ 次, for 语句共需要循环 n 次. 综合以上因素, 启发式算法的时间复杂度应该是 $O(m+r-1)+O(m,u,n,r)O(u)n=O(m,u^2,n^2,r)$. 即使在最

坏的情况下($u=n$)时的时间复杂度为 $O(m, n^4, r)$.

4　结果分析

这里我们基于上述算法进行最有可能贝叶斯网络结构的发现, 并对实验结果进行分析, 数据库是经过预处理的文本数据库, 为了简化我们只选择其中的一部分区域进行说明, 文本数据库中包含五个数据域: sex、class、intelligence、support、plan.

表 1　部分文本数据库

Sex	Class	Intelligence	Support	Plan
1	2	3	1	1
1	1	1	1	1
1	1	2	0	1
0	2	2	1	1
1	3	3	0	0
1	2	0	1	1

这里的数据表格只是整个数据库中的一部分, 主要目的是为了反映整个数据库的数据项目构成方式.

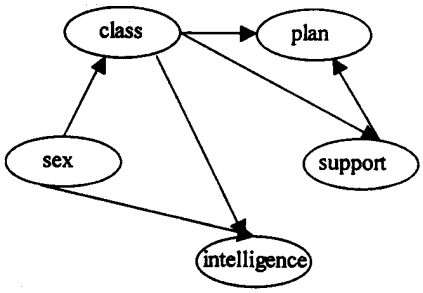

图 1　前 100 条记录的网络结构

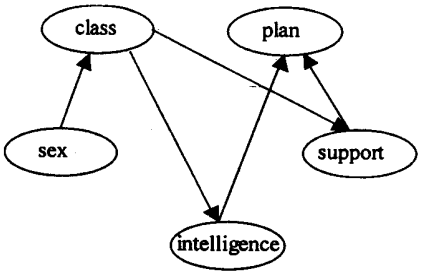

图 2　5000 条以后的 100 条记录的网络结构

从建立的图论模型可以看出, 首先网络的结构和整个数据的分布差异是有关系的, 即区域性的分布差异会对数据网络产生影响. 这对于不同时间段获取的数据进行分析, 会产生一些与时间有关的变化趋势. 其次, 数据的规模对于网络的结构也有较大的影响, 随着数据规模的增大, 节点之间的内在关系和长期关系也逐渐显露出来的, 产生一般性的规律变化模式, 通过结果分析我们可以发现, 贝叶斯网络能够很好地表示

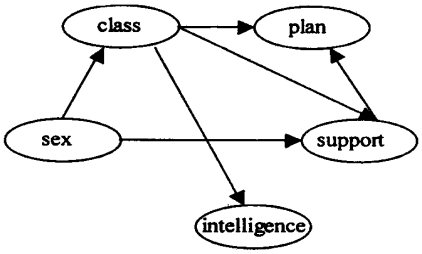

图 3　前 200 条记录的网络结构

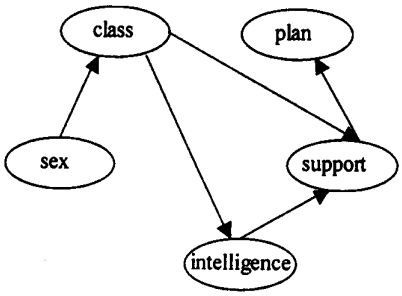

图 4　前 1000 条记录的网络结构

数据项之间的相互关系, 但是对数据库的规模比较敏感. 因此, 如何消除贝叶斯方法对数据库规模的敏感性是我们今后要研究的问题.

参　考　文　献

1　David Heckerman, Christopher Meek, Gregory Cooper. A bayesian approach to causal discovery [R]. Technical Report MSR-TR-97-05

2　David Heckerman. Bayesian networks for data mining [J]. Data Mining and Knowledge Discovery 1, 1997. 79~ 119

3　Pat Langley, Stephanie Sage. Induction of selective bayesian classifiers [C]. In Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, 1994

4　Gregory F. Cooper, Edward Herskovits. A bayesian method for the induction of probabilistic networks from data [J]. In Maching Learning 1992. 9, 309~ 347.

Study of Determining Bayesian Network Topology Structures

WANG　Wei, CAI　Lian-hong

(*Department of Computer Science, Tsinghua University, Beijing* 100084, *China*)

Abstract　Bayesian approach is an important method in statistics. A Bayesian network is a graphics model that encodes probabilistic relationships among variables of interest. But it is difficulty to determine its topology structure. In this paper, we use an approach for obtaining Bayesian network structure. Finally the method is used to test database and the result demonstrates the method is effective.

Key words　bayesian probability; bayesian network; topology structures