第 36 卷 第 5 期 计 算 机 工 程 2010 年 3 月
Vol.36 No.5 Computer Engineering March 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)05—0165—03 文献标识码：A 中图分类号：TP393.02

非确定先验信息的贝叶斯网结构学习方法

刘明辉[1]，王 磊[2]，党林阁[1]，石景岚[1]
(1. 解放军 63891 部队，洛阳 471003；2. 国防科技大学信息系统与管理学院，长沙 410073)

摘 要：针对非确定先验结构信息下的贝叶斯网络学习问题，提出一种非确定先验结构信息贝叶斯网络的结构学习方法。为更好地利用不确定性信息，对 MDL 测度进行改进，提出 SMDL 测度，使之能在学习过程中考虑先验信息的不确定性，使用模拟退火算法对问题进行求解。通过实验对算法的可行性和效率进行验证。

关键词：贝叶斯网络；结构学习；专家知识；模拟退火

Structure Learning Method of Bayesian Network with Uncertain Prior Information

LIU Ming-hui[1], WANG Lei[2], DANG Lin-ge[1], SHI Jing-lan[1]
(1. Unit 63891 of PLA, Luoyang 471003;
2. School of Information System and Management, National University of Defense Technology, Changsha 410073)

【Abstract】This paper presents a structure learning method of Bayesian network to solve the problem of structure learning with uncertain prior information. A description method of the uncertain prior information is given. An improved MDL score method named SMDL is proposed to fit the uncertain prior information in learning process. Simulated annealing method is used to solve the problem. This method is validated by experiments.

【Key words】Bayesian network; structure learning; expert knowledge; simulated annealing

1 概述

贝叶斯网络学习是贝叶斯网络的重要研究内容，也是贝叶斯网络构建中的关键环节。贝叶斯网络的学习，就是要通过某种学习算法来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依赖关系的贝叶斯网络模型。一个完整的贝叶斯网络是由网络拓扑结构和每一个节点上的条件概率表(CPT)组成的，因此，贝叶斯网络学习可以分为结构学习和参数学习 2 个部分。其中，结构学习是贝叶斯网络学习的基础，也是进行贝叶斯网络参数学习的前提条件。

现有的贝叶斯网络结构学习算法大体上可以分为 2 类，即基于条件独立性的方法和基于评分搜索的方法。基于条件独立性的方法主要是通过对训练样本集的条件独立性测试来发现节点之间的依赖关系，再通过节点之间的依赖关系来构建贝叶斯网络；基于评分搜索的方法将贝叶斯网络学习看作最优化问题，首先在网络上定义一个可分解的评分测度，评分测度描述了每个可能的结构对数据样本的拟合程度，通过搜索的方法来寻找评分最高的网络结构。基于评分搜索的方法简单规范，但是随着网络节点的增加，网络结构空间呈指数性增长，因此无法对所有可能的结构进行搜索，为了降低搜索空间，基于评分的方法一般都要求节点有序并使用一些启发式搜索方法。

在贝叶斯网络的学习过程中，先验信息是十分重要的。完全的贝叶斯网络学习是一个 NP 问题，而利用专家知识建立先验贝叶斯网络，可以去除大量的不可能的网络结构，减小结构学习的搜索空间，从而大大降低学习的复杂度。同时，由于数据样本的来源存在可信度不确定的问题，在样本数据不足或者数据样本可信度不够高的情况下，必须依靠专家的先验知识来对贝叶斯网络的学习结果进行综合评判，才能获得与客观实际相符合的结果。

为了将专家先验知识引入贝叶斯网络的学习过程，从而在贝叶斯网络学习过程中，当样本数据不足或者数据样本可信度不够高时可以获得更好的学习效果，本文在学习和研究贝叶斯网络理论的基础上，提出了贝叶斯网络的“非确定先验信息”表示方法，定义了结构不确定度的概念，在此概念的基础上对现有的 MDL 测度和模拟退火算法进行了改进，并通过实验对该算法进行了验证。

2 非确定先验结构信息的贝叶斯网络

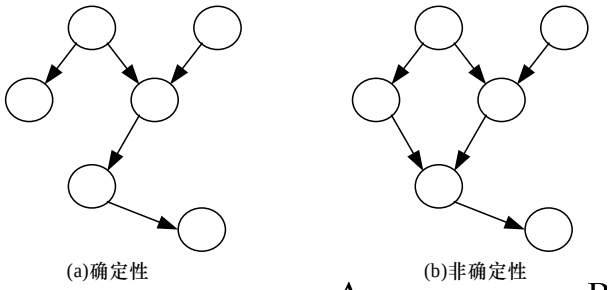
在当前所研究的贝叶斯网络中，先验贝叶斯网络一般是通过专家的领域知识来建立的。一般来说，在先验网络中专家主要根据自身的领域知识确定网络中的因果关系，即贝叶斯网络的结构，而条件概率表则通过数据样本学习来获得。然而，在通常的情况下，专家对贝叶斯网络中的结构往往也是不能完全肯定的。一般来说专家更倾向于给出先验贝叶斯网络中各事件节点存在因果关联的可能性，即网络中各有向边存在的概率。

如图 1 所示，在图 1(a)中，专家所提供的先验结构的信息是确定的，如节点 A 和 D 之间存在确定的有向边，而 C 和 E 之间不存在有向边；而在图 1(b)中，专家认为先验结构的信息是非确定的，如节点 A 和 D 之间存在有向边的概率为

作者简介：刘明辉(1980－)，男，工程师、硕士，主研方向：作战效能评估，电子系统建模仿真，体系对抗仿真；王 磊，博士研究生；党林阁、石景岚，工程师

收稿日期：2009-10-17 **E-mail：**wang.laye@gmail.com

0.7；节点 C 和 E 之间存在有向边的概率不大，但仍有 0.2 的可能性。



(a)确定性　　(b)非确定性

图 1　先验结构贝叶斯网络

由此可见，在传统的贝叶斯网络研究中，所提供的先验信息是一种约减之后的粗糙信息。这种先验信息忽略掉了专家所提供的先验信息的部分细节，从而造成了对专家领域知识的部分浪费。这部分细节信息在经过处理之后，是可以用于优化先验贝叶斯网络结构的，为了描述先验网络结构中这种有向边存在的不确定性，本文提出用结构不确定度(Structure Uncertainty, SU)来描述先验贝叶斯网络结构中有向边的不确定性大小。

先验贝叶斯网络的结构不确定度定义如下：

一个非确定先验结构信息贝叶斯网络的结构不确定度 $SU(BN)$ 是其各有向边的结构不确定度 $SU(E_{ij})$ 之积，即：

$$SU(BN)=\prod_i\prod_j SU(E_{ij}) \tag{1}$$

其中，

$$SU(E_{ij})=\frac{1}{2}[1-\mathrm{lb}P(E_{ij})] \tag{2}$$

由式(2)可以看出，贝叶斯网络的结构不确定度实际上是其有向边存在概率的负对数函数，在专家所确定的先验贝叶斯网络中，有向边可能存在的概率越大，则结构不确定度越小，若网络中某一有向边 E_{ij} 信息未知，则可规定其存在概率为 0.5(即存在与否可能性均等)，对应的 $SU(E_{ij})=1$。

在定义结构不确定度之后，可以将一个先验贝叶斯网络表示为三元组 $BN_P(S_P,T,S_{SU})$ 的形式，其中，S_P 为先验网络的结构；T 为先验网络的条件概率表；S_{SU} 为先验网络的结构不确定度矩阵。

对于非确定先验信息的贝叶斯网络结构学习方法，目前的研究还比较少，文献[1]曾提出一种知识和数据融合的贝叶斯网络结构建模方法。该方法首先利用证据理论融合多位专家的意见，通过专家知识确定节点之间的因果联系，去除大量无意义的拓扑结构，缩小搜索空间，然后再利用学习算法来搜索最好的网络结构。这种方法考虑了依靠数据对根据专家先验知识所建的贝叶斯网络进行缩减，但是没有对依靠数据对专家所建立的贝叶斯网络进行完善丰富的情况。同时，其仅在构建先验网络的过程中用到了专家知识，而没有将专家所提供的先验信息应用到贝叶斯网络学习过程之中。

3　基于结构不确定度的改进MDL测度

在基于评分搜索的贝叶斯网络结构学习方法中，评分测度是网络学习的基础，现有的评分测度主要有 2 种：MDL 测度和 BDe 测度。然而，这 2 种测度对贝叶斯网络的先验信息描述都不够完善。MDL 测度是完全基于数据的，没有用到先验知识，其学习结果的正确性完全依赖于实例数据集合。而 BDe 测度虽然考虑了先验信息，但要求专家给出的先验信息必须是确定性的，并且要给出每一个可能的先验结构 B_S^i 的概率值，当可能的结构空间很大时，先验结构的概率分布很难计算。文献[2]采用了简单的估计方法，假设所有的先验结构 B_S^i 可能出现的概率相同，即网络结构的先验概率是一个均匀分布，这个假设较为武断，对网络结构的先验信息描述性较弱；文献[3]也给出了另一种先验分布的估计方法，在 Herckerman 的方法中，首先要求构造一个完整的先验贝叶斯网络 B_{PS}，并通过式(3)来计算一个可能的先验贝叶斯网络结构的概率：

$$P(B_S^i\mid\xi)=ck^{\delta_i} \tag{3}$$

其中，ξ 为先验信息；c 为归一化常数；$0\leqslant k\leqslant 1$；δ_i 为网络结构 B_S^i 与完整先验贝叶斯网络 B_{PS} 结构中不同的有向边的数目。显然，这是一种近似的估计方法，可信度不高。此外，BDe 测度中没有明确地包含结构复杂性指标，在实际的网络学习过程中，容易向结构复杂的网络方向倾斜。

基于上述原因可知，在非确定先验信息的贝叶斯网络结构描述下，利用现有的评分测度来进行网络结构学习，其效果难以保证，因此，本文提出了一种基于结构不确定度的改进 MDL 测度，具体定义如下：

设有贝叶斯网络结构图 G，$V=\{V_1,V_2,\cdots,V_n\}$ 为贝叶斯网络的节点集，E 为贝叶斯网络的有向边集，则贝叶斯网络的改进 SMDL 测度可以表示为

$$SMDL(G)=MDL(G)\otimes SU(E) \tag{4}$$

令 E_{ij} 为节点 $V_i\to V_j$ 的有向边，$MDL(V_j,Pa(V_j))$ 为贝叶斯网络某节点及其父节点集的局部 MDL 测度，n 为网络中的节点数，k 为节点 V_j 的父节点数，$SU(E_{ij})$ 为先验贝叶斯网络有向边的结构不确定度，由测度的可分解性，可求得贝叶斯网络的改进 SMDL 测度为

$$SMDL(G)=\sum_{j=1}^{n}MDL(V_j,Pa(V_j))\times SU(V_j)=\sum_{j=1}^{n}MDL(V_j,Pa(V_j))\times\prod_{i=1}^{k}SU(E_{ij}) \tag{5}$$

与 MDL 测度相比，改进的 SMDL 测度由于引入了专家知识的非确定先验信息表示，使得利用 SMDL 测度学习的网络与专家领域知识有着更高的拟合度。利用该测度进行学习，可以对网络结构的合理性有更好的解释。

4　基于SMDL测度的改进模拟退火算法

在基于评分搜索的贝叶斯网络结构学习中，模拟退火算法(SA)是一种常用的启发式算法。该算法最早是由 Metropolis 于 1953 年提出来的，其核心思想是通过模拟热力学中经典粒子系统的降温过程来求解规划问题的极值。

设组合优化问题的一个解 m 及其目标函数 $f(m)$ 分别与固体的一个微观状态 m 及其能量 E_m 等价，随机选择某粒子并使其位移产生微小变化，得到一个新状态 n 及其能量 E_n，在模拟退火算法中，采用 Metropolis 准则来决定更新系统状态的概率 p，Metropolis 准则如下式所示：

$$p(m\Rightarrow n)=\begin{cases}1 & f(n)\leqslant f(m)\\ \exp[\dfrac{f(m)-f(n)}{T}] & f(n)>f(m)\end{cases} \tag{6}$$

其中，$T\in R^+$ 表示控制参数，开始令 T 取较大的初值，在进行足够多转移之后，缓慢减小 T 的值，最终即可获得系统的最优解。

文献[4]给出了完整的采用模拟退火算法进行贝叶斯网

络结构学习的步骤:

(1)初始化网络结构，确定一个初始的网络结构，作为初始网络的拓扑结构，并确定一个较高的 T 值。

(2)对该结构进行修改(如添加一条有向边，删除一条有向边，改变一个有向边的方向)，要求修改后所产生的新结构不能含有有向环。用集合 B_S' 表示对当前结构所有可能的修改构成的集合。随机地从集合 B_S' 中选取一个对当前结构可能的修改 e，修改后结构测度值的变化量为 Δe，求出 $p=\exp(-\Delta e/T)$ 的值。

(3)如果 $p>1$，则采纳修改 e，修改当前结构，如果 $p<1$，则以概率值 p 采纳修改 e，修改当前结构。

(4)重复第(2)步和第(3)步 α 次。如果在 α 次重复中没有修改网络结构，则停止算法，此时的结构即为所求的网络结构；否则，循环次数加 1。如果 $i>\gamma$，则算法停止，此时的结构即为所求的网络结构。如果 $i<\gamma$，按照降温步长 $\beta(0<\beta<1)$ 降低 T 值，即 $T=\beta\times T$，然后转入第(2)步，继续该算法。

与传统的随机搜索方法算法不同，模拟退火算法不但引入了适当的随机因素，还引入了物理系统退火的自然机理，在寻求系统最优解的过程中，还以一定的概率接受系统的恶化解，因此，该算法不像其他一些算法那样易于陷入系统的局部最优解中，并且对系统初值的依赖较小。

对于已经给出了部分先验信息的贝叶斯网络结构，由于各有向边的存在概率已知，一个很朴素的想法是：在增加有向边的时候，要尽量提高存在可能性大的边被添加的概率；在删除有向边的时候，要尽量提高那些存在可能性小的边被删除的概率。通过这种方法，可以提高有先验信息的贝叶斯网络结构学习的效率，由此，可以将上文中所提出的结构不确定度引入模拟退火算法，重新定义模拟退火算法中的 Metropolis 准则为

$$p'(m\Rightarrow n)=\alpha\times p(m\Rightarrow n)=\begin{cases}\min(\alpha,1) & f(n)\leqslant f(m)\\ \min\{\alpha\cdot\exp[\dfrac{f(m)-f(n)}{T}],1\} & f(n)>f(m)\end{cases}\tag{7}$$

其中，

$$\alpha=\begin{cases}2\cdot 2^{1-2SU(E_{ij})} & \text{(增加边)}\\ 2\cdot[1-2^{1-2SU(E_{ij})}] & \text{(删除边)}\end{cases}\tag{8}$$

当专家未提供某有向边的先验信息时，$P(E_{ij})=0.5$, $SU(E_{ij})=1$，在这种情况下 $\alpha=1$，此时的改进模拟退火算法与原算法完全相同。

改进的模拟退火算法由于在搜索过程中，依据结构不确定度动态地调整了各有向边的搜索概率，因而，在限制一定精度的前提条件下，有可能比传统算法更快地收敛到最优解，当退火时间充分长的时候，由于任何微小可能性的事件均已发生，因此改进的模拟退火算法与传统算法获得的结果趋向一致，即两者具有一致收敛性。

5 算法验证

为了验证本文中提出的基于SMDL测度的非确定先验信息贝叶斯网络结构学习算法，本文利用该算法对 Alarm 网络进行学习。

Alarm 网络是一个用于病人监测的医疗诊断系统，它由 37 个节点、46 条有向边组成，每个节点的不同取值数在 2~4 之间，整个网络共有 20 000 条数据记录。其拓扑结构如图 2 所示。

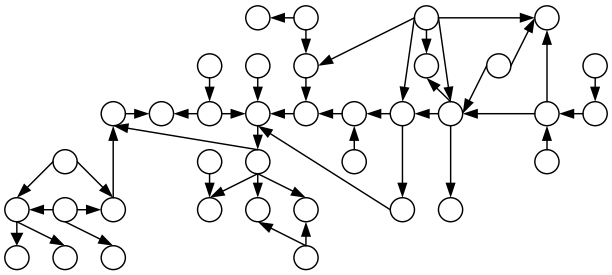


图 2　Alarm 网络拓扑结构

在实验中，随机选取 50%的已知有向边作为先验结构信息，对这些有向边加入先验信息，采用 SMDL 测度和改进的模拟退火算法对其进行学习，并将学习的结果与采用 MDL 进行结构学习的结果进行对比，所得数据如表 1 所示。

表 1　MDL 测度与 SMDL 测度学习 Alarm 网络结果对比

测度类型	网络分值	丢失边数	多余边数
MDL 测度	81 343.6	2	1
SMDL 测度	55 675.3	1	0

由表 1 的数据可以看出，由于加入了专家提供的先验信息，网络的测度值有所减小，同时网络的学习结果也有所改善。在实验过程中，未考虑专家提供错误信息的情况。若需要增加专家先验信息的可靠性，则可以通过专家群决策的办法，融合多位专家提供的先验信息，从而尽量减少错误先验信息带来学习结果的不稳定性。

6 结束语

传统的贝叶斯网络学习算法对数据样本的依赖性较强，当数据样本不足或者可信度不高时，往往不能获得令人满意的学习结果，同时，由于专家知识固有的不确定性，已有的结构学习算法没有对专家知识加以较好的利用，在学习过程中存在效率低下、可信度较低的问题。本文提出的非确定先验信息的贝叶斯网络结构学习方法，通过改进 MDL 测度，将具有不确定性的先验信息融入了贝叶斯网络结构学习过程中，充分提高了贝叶斯网络结构学习的可信度。

参考文献

[1] 胡笑旋. 贝叶斯网建模技术及其在决策中的应用[D]. 合肥: 合肥工业大学, 2006.

[2] Cooper G F, Herskovits E. A Bayesian Method for the Induction of Probabilistic Networks Form Data[J]. Machine Learning, 1992, 9(4): 309-348.

[3] Heckerman D, Geiger D, Chickering D M. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data[J]. Machine Learning, 1995, 20(3): 197-243.

[4] 邢永康, 沈一栋. 学习信度网的结构[J]. 计算机科学, 2000, 27(10): 83-88.

编辑　任吉慧