基于模拟退火的贝叶斯网络结构学习算法*)

张少中　王秀坤　丁　华
(大连理工大学计算机科学与工程系　大连116023)

摘　要　贝叶斯网络的学习可分为结构学习和参数学习。基于模拟退火的结构学习算法是一种以搜索最高记分函数为原则的智能优化方法。本文以KL距离、相互信息以及最大相互信息为基础，通过附加合适的约束函数降低学习搜索的复杂度，提出一种附加约束的最大熵优化函数作为模拟退火算法的能量优化函数，并结合贝叶斯网络结构学习的特点设计了适合模拟退火的变量表示和邻近值产生机制。通过与其他用于结构学习的模拟退火算法，以及遗传和进化算法比较分析，结果表明本文中提出的基于模拟退火的贝叶斯网络结构学习算法在时间和精度上都具有较好的效果。

关键词　贝叶斯网络，结构学习，模拟退火算法，最大信息熵，约束函数

An Algorithm for Bayesian Networks Structure Learning Based on Simulated Annealing

ZHANG Shao-Zhong　WANG Xiu-Kun　DING Hua
(Department of Computer Science, Dalian University of Technology, Dalian 116023)

Abstract　The automated creation of Bayesian networks can be separated into two tasks, Structure learning, which consists of creating the structure of the Bayesian networks from the collected data, and parameter learning, which consists of calculating the numerical parameters for a given structure. We present a simulated annealing algorithm for structure learning in Bayesian networks and propose a score function for optimization based on maximum mutual information entropy with oddditional restriction. The entropy is based on KL distance, mutual information and maximum mutual information. We also propose a denotative form for variable and give a mechanism for generating contiguous data. Some experimentation on other simulated annealing and genetic or Evolutionary algorithms are given. The result indicates that the simulated anneáling algorithm based MMI_L that we propose has more efficiency and precisely in cost and precision than others.

Keywords　Bayesian networks, Structure learning, Simulated annealing algorithm, Maximum information entropy, Restrained function

1　引言

贝叶斯网络是根据各个变量之间的概率关系，使用图论方法表示变量集合的联合概率分布的图形模型。该模型是一个有向无环图，结点代表论域中的变量，有向弧代表变量的关系，变量之间的关系强弱由结点与其父结点之间的条件概率来表示。通过贝叶斯网络可以准确地反映实际应用中变量之间的因果和依赖关系[1]。

贝叶斯网络的学习可划分为两个方面：结构学习和参数学习。结构学习是利用训练样本集，尽可能结合先验知识，确定合适的贝叶斯网络拓扑结构；参数学习是在给定贝叶斯网络拓扑结构的情况下，确定各结点处的条件概率密度。

1992年Cooper和Herskovits提出以BDE记分为评价函数的K2算法，该方法是以爬山法为思想的启发式搜索算法[2]；1993年Suzuki和Lam分别提出了基于MDL记分的结构学习算法[3,4]；Suzuki使用branch and bound技术降低计算量，并提出了B&B_MDL启发式算法[5~7]；1995年Chickering提出了模拟退火思想[8]；1996年Larranaga提出了结构学习的遗传算法[9]；1999年Wong和Lam采用演化程序设计(Evolutionary Programming)的方法实现了一个基于MDL记分的学习算法[10]。这些算法通过定义和选择合适的记分函数作为评价结构优化的标准，然后利用爬山法、模拟退火算法、遗传和进化算法等优化方法对该函数进行优化处理。本文从信息论基本理论出发，利用KL距离、相互信息熵准则确立了最大信息熵为评价函数，并通过附加合适的约束函数作为模拟退火的能量优化函数，提出了一种用于贝叶斯网络结构学习的模拟退火算法。最后使用Alarm数据集对典型的Chickering提出的模拟退火算法、Larranaga提出的遗传算法和本文的模拟退火算法进行了实验比较，分析了本文提出的算法在贝叶斯网络结构学习中在时间和精度上的效果。

2　贝叶斯网络和结构学习

2.1　贝叶斯网络

给定一个有向无环图S和一个在离散变量集合$X=\{x_1, x_2, \cdots, x_n\}$上的联合概率分布$P$，如果$S$可以代$P$，即在$X$中的变量和$S$的结点之间存在有一一对应，使得$P$可以进行如下的递归乘积分解：$P(x_1, \cdots, x_n)=\prod_i P(x_i|Pa(x_i))$，这里$Pa(x_i)$是$S$中$x_i$的直接祖先(父结点)，则我们将图$S$和$P$的联合称为贝叶斯网络。

*)项目基金：国家科技部973专项，2001CCA00700。张少中　博士生，研究方向：数据库的知识发现、决策支持系统；王秀坤　教授，博导，研究方向：数据库系统、决策支持系统；丁　华　博士生，研究方向：数据库的知识发现。

关于一组变量 $X=\{x_1,x_2,\cdots,x_n\}$ 的贝叶斯网络由以下两个部分组成:1)一个表示 X 中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。S 是一个有向无环图,S 中的结点一对一地对应于 X 中的变量,结点之间缺省弧线表示条件独立,S 和 P 定义 X 的联合概率分布。根据条件独立的性质,联合概率分布为:

$P(x_1,x_2,\cdots,x_n|\xi)=\prod_{i=1}^{n}p(x_i|x_1,x_2,\cdots,x_{i-1},\zeta)$,对于每个变量 x_i,令 $Pa(x_i)\subseteq\{x_1,x_2,\cdots,x_{i-1}\}$ 是 x_i 的父结点集,$x_1,x_2,\cdots,x_{i-1}$ 条件独立,则:$P\{x_i|x_1,x_2,\cdots,x_{i-1},\xi\}=p(x_i|Pa(x_i),\xi)$,这里 ξ 为先验知识。

2.2 贝叶斯网络的结构学习

贝叶斯网络的结构学习目标是利用训练样本集 D,$D=\{d_1,d_2,\cdots,d_m\}$,,尽可能结合先验知识 ξ,确定最合适的网络拓扑结构 S。在基于记分搜索的学习方法中,就是寻找使得记分为最佳的贝叶斯网络模型,记为 $S\leftarrow arg\ bestscore(S,D,\zeta)=P(S|D,\zeta)$。

3 最大信息熵记分函数

3.1 KL 距离、信息熵和最大信息熵

关于一个问题域上的两个参数,在同一空间定义两个概率分布,我们可以使用距离函数比较两个参数在该空间的差异。用于两个概率密度函数间评估的方式是采用合式评估函数(Proper Scoring Function)。合式评估函数有多种,其中最广泛使用的是:对数评估函数、平方评估函数和球形评估函数。

在信息论领域,通常采用对数评估函数比较问题域上的两个参数在同一空间定义的两个概率分布的差异,这种评估函数就是 KL(Kullback-Leibler)距离,定义如下:

设 p 是问题域 U 上的一个概率密度函数,另一个概率密度函数 q 是 p 的一种近似,则 p 和 q 之间的 Kullback-leibler 距离表示为:

$$d_{KL}(p\|q)=\sum_{x\in U}p(x)\log p(x)/q(x)$$

KL 距离表示了概率密度函数 p 和 q 之间的接近程度,KL 距离越小,表示 p 越近似于 q,当且仅当 p 等于 q 时,KL 距离等于0。

在信息论中,相互信息用来表示发送或收到某些信息后给某些变量所带来的信息。设 V_i 和 V_j 为问题域 U 上的两个不同变量,则它们之间的相互信息为:

$$MI(v_i,v_j)=\sum\sum p(v_i,v_j)\cdot\log p(v_i,v_j)/[p(v_i)\cdot p(v_j)]$$

这也就是 KL 距离:

$$\sum_{v_j}KL(p(v_i|v_j),p(v_i))$$

$$=\sum\sum p(v_j)\cdot p(v_i|v_j)\log p(v_i|v_j)/p(v_i)$$

V_i 的先验概率为 $p(v_i)$,当观测到 V_j 的值后,V_i 的后验概率为 $p(v_i|v_j)$,变量 V_j 的概率密度函数为 $p(v_j)$,$MI(v_i,v_j)$ 反映了"观测到 V_j 的值"这一事件给变量 V_i 带来的信息。

最大相互信息原则是 Chow 和 Liu 的 MWST(Maximum Weight Spanning Tree)算法、Rebane 和 Pearl 的 Polytree Construction 算法、Wallace 的 WKD 算法及 Sakar 和 Murthy 算法的推广和总结,文[11]对此进行了详细的推导并得出结论:

设研究的问题域为 U,$U=\{v_1,v_2,\cdots,v_n\}$,$p(v_1,v_2,\cdots,v_n)$为训练样本在 U 上的联合概率分布,$pa(v_i)$为贝叶斯网络 S 中结点 v_i 的父结点集合,记 $\hat{p}(v_1,v_2,\cdots,v_n)$为 U 上的实际联合概率分布,则有:$\hat{p}(v_1,v_2,\cdots,v_n)=\prod_{i=1}^{n}p(v_i|Pa(v_i))$

$p(v_1,v_2,\cdots,v_n)$与 $\hat{p}(v_1,v_2,\cdots,v_n)$的 KL 距离反映了 $p(v_1,v_2,\cdots,v_n)$与实际联合概率分布的接近程度,且等式:

$$KL(p(v_1,v_2,\cdots,v_n),\hat{p}(v_1,v_2,\cdots,v_n)=\sum_{v_1,v_2,\cdots,v_n}p(v_1,v_2,\cdots,v_n)\cdot\log p(v_1,v_2,\cdots,v_n)-\sum_{i=1}^{n}MI(v_i,Pa(v_i))-\sum_{i=1}^{n}\sum_{v_1,v_2,\cdots,v_i}p(v,v,\cdots,v)\cdot\log p(v_i)$$

成立。

上式中 $\sum_{v_1,v_2,\cdots,v_n}p(v_1,v_2,\cdots,v_n)\cdot\log p(v_1,v_2,\cdots,v_n)$是概率密度分布函数 $p(v_1,v_2,\cdots,v_n)$的熵,$\sum_{i=1}^{n}\sum_{v_1,v_2,\cdots,v_i}p(v,v,\cdots,v)\cdot\log p(v_i)$为 $p(v_1,v_2,\cdots,v_n)$分布下各结点的边缘概率分布熵的和,它们与网络拓扑结构无关。因此 $p(v_1,v_2,\cdots,v_n)$与 $\hat{p}(v_1,v_2,\cdots,v_n)$之间的 KL 距离取决于 $\sum_{i=1}^{n}MI(v_i,Pa(v_i))$,当 $\sum_{i=1}^{n}MI(v_i,Pa(v_i))$值越大,则 $p(v_1,v_2,\cdots,v_n)$与 $\hat{p}(v_1,v_2,\cdots,v_n)$之间的 KL 距离越小。

这就是重要的最大相互信息原则,由于单纯利用最大相互信息原则将导致搜索终止于完全图[11],因此有必要对该原则进行复杂度约束。

3.2 复杂度约束函数

贝叶斯网络的复杂度与网络结点的维数和网络结构密切相关,我们从这两方面出发对最大相互信息原则进行复杂度约束。

网络模型的维数:记为 $Dim(S)$,维数 $Dim(S)$可以通过每个结点不同状态的数目 S_i 来计算,表示为:

$$Dim(S)=\sum_{i=1}^{n}(S_i-1)\prod_{v_j\in Pa(v_i)}S_j$$

式中 n 为网络中结点的数目。

令 m 为数据库中数据集的数目,这样每个参数具有一个协方差$\sqrt{m}$,对于网络中的每个参数,其需要计算的次数为:

$$d=\log\sqrt{m}\Rightarrow d=(\log m)/2$$

网络结构的复杂度记为:$DL(S)$,网络结构的复杂度与结点的父结点的数目密切相关,也就是有向边的数目,由下式计算得到:

$$DL(S)=\sum_{i=1}^{n}k_i\log_2(n),k_i\text{ 为结点 }v_i\text{ 父结点的数目。}$$

这样关于复杂度约束函数就由维数和网络复杂度两部分构成,即:

$$L_{NETWORK}=\frac{\log m}{2}Dim(S)+DL(S)$$

$$=\frac{\log m}{2}[\sum_{i=1}^{n}(S_i-1)\prod_{v_j\in Pa(v_i)}S_i]+\sum_{j=1}k\log_2(n)$$

3.3 附加约束的最大信息熵记分函数

为了避免单纯利用最大相互信息原则导致搜索终止于完全图,本文结合上面的复杂度约束规则,对最大相互信息附加一个关于网络拓扑结构与变量维数复杂度约束函数,得到附加约束的最大信息熵记分函数如下:

$$(MMI-L)=\sum_{i=1}^{n}MI(v_i,Pa(v_i))-L_{NETWORK}$$

$$=\sum_{i=1}^{n}MI(v_i,Pa(v_i))-[\frac{\log m}{2}Dim(S)+DL(S)]$$

$$=\sum_{i=1}^{n}MI(v_i,Pa(v_i))-\{\frac{\log m}{2}[\sum_{i=1}^{n}(S_i-1)\prod_{v_j\in pa(v_j)}S_i]+\sum_{j=1}^{n}k_i\log_2(n)\}$$

定理 设训练样本集 D 中蕴含的问题域，U，$U=\{v_1,v_2,\cdots,v_n\}$上的联合概率密度函数为 $p(v_1,v_2,\cdots,v_n)$，则在所有问题域上的贝叶斯网络拓扑结构中，满足：

$$Score_{MMI-L}=\sum_{i=1}^{n}MI(v_i,Pa(v_i))-\{\frac{\log m}{2}[\sum_{i=1}^{n}(S_i-1)\prod_{v_j\in pa(v_j)}S_i]+\sum_{j=1}^{n}k_i\log_2(n)\}$$

最大的拓扑结构是对 $p(v_1,v_2,\cdots,v_n)$在逼近程度和复杂度之间取折衷的最合适的拓扑结构。

4 基于模拟退火的贝叶斯网络结构学习算法

4.1 模拟退火算法的主要思想

模拟退火算法得益于材料统计力学的研究成果。统计力学表明材料中粒子的不同结构对应于粒子的不同能量水平。在高温情况下，粒子的能量较高，可以自由运动和重新排序；在低温条件下，粒子的能量较低。如果从高温开始，非常缓慢地降温，粒子就可以在每个温度下达到热平衡，这个过程称为退火。当系统完全被冷却时，最终形成处于低能状态的晶体。

如果使用粒子的能量定义材料的状态，假定材料在状态 i 下的能量为 $E(i)$，那么材料在温度 T 时从状态 i 进入状态 j 遵循如下规律：

1)如果 $E(j)\leqslant E(i)$，接收该状态被转换；

2)如果 $E(j)>E(i)$，则状态转换以如下概率被接收：$e^{\frac{E(i)-E(j)}{KT}}$

K 为物理学中的波尔兹曼常数，T 为材料的温度。

在贝叶斯网络结构学习问题中，给定问题域 $U=\{v_1,v_2,\cdots,v_n\}$和训练样本集 D，以记分函数为评价原则，找出问题域中每个结点 v_i 的父结点的集合 $Pa(v_i)$。

4.2 结构学习的模拟退火算法描述

在本算法中，确定网络拓扑结构是通过搜索问题域中每个结点的父结点集实现的。搜索结点 v_i 的父结点集 $Pa(v_i)$的算法描述如下：

对每个结点 v_i 搜索其父结点集，$i=1$到 n 执行循环：

1)初始化父结点集 $Pa(v_i)\leftarrow\phi$；

2)计算评价函数：$Score_{MMI-L}(S)=\sum_{i=1}^{n}MI(v_i,Pa(v_i))-\{\frac{\log m}{2}[\sum_{i=1}^{n}(S_i-1)\prod_{v_j\in pa(v_i)}S_i]+\sum_{j=1}^{n}k_i\log_2(n)\}$；$m$ 为数据库中数据集的数目，n 为网络中结点的数目，k_i 为结点 v_i 的父结点的数目，S_i 为每个结点不同状态的数目；

3)对附加约束后的最大相互信息取负值，转化为能量表示：$E_{old}=-Score_{MMI-L}(S)$；

4)如果没有达到结束条件，执行下一步；否则执行步骤15；

5)迭代记数器初始化 $l=1$；$l_c=L(x)$；$L(x)$为一个控制步数函数；

6)如果 $l<l_c$ 则执行下一步；否则执行步骤14；

7)产生邻近值：$Pa_{new}(v_i)\leftarrow v_j$；$v_j$ 是 $U=\{v_1,v_2,\cdots,v_n\}$的一个随机结点；

8)计算评价函数：$Score_{MMI-L}(S)=\sum_{i=1}^{n}MI(v_i,Pa_{new}(v_i))-\{\frac{\log m}{2}[\sum_{i=1}^{n}(S_i-1)\prod_{v_j\in pa(v_j)}S_i]+\sum_{j=1}^{n}k_i\log_2(n)\}$；

9)$E_{new}=-Score_{MMI-L}(S)$；对附加约束后的最大相互信息取负值；

10)判断：如果新能量函数小于原能量函数：$E_{new}<E_{old}$，则执行下一步；否则执行步骤12；

11)接受结点加入父结点集 $Pa(v_i)=Pa_{new}(v_i)$，转步骤13；

12)以概率 $P\leftarrow\exp((E_{old}-E_{new})/T_i)$接受结点加入父结点集；

13)计数器：$l=l+1$，转步骤6；

14)温度下降计算：$T_r=rT_i$；

15)结束。

4.3 特征组合的表示

给定包含 n 个属性的问题域 $U=\{v_1,v_2,\cdots,v_n\}$，则定义在问题域的贝叶斯网络拓扑结构可以表示为：阵$\lfloor s_{ij}\rfloor_{n\times n}$，其中 $s_{ij}=\begin{cases}1,v_j\in Pa(v_i)\\0,v_j\notin Pa(v_i)\end{cases}$，即，当矩阵中元素 s_{ij}为1时，表示变量 v_j 是变量 v_i 的父结点，反映到拓扑结构中为存在 v_j 到 v_i 的有向边，表示为：$v_j\rightarrow v_i$；当矩阵中元素 s_{ij}为0时，表示变量 v_j 不是变量 v_i 的父结点，反映到拓扑结构中为 v_j 到 v_i 之间不存在有向边。

4.4 邻近值的产生

理论上邻近值的产生是随机的，但是邻近值的产生机制直接影响模拟退火算法的效率。本文产生邻近值分为三个部分：交换结点、加入新结点和删除旧结点。交换结点部分是将父结点集中的结点与父结点集外的结点进行随机交换操作，然后计算其能量函数，例如：初始父结点集为(v_1,v_j)，其余结点为$(v_k,\cdots,v_n)$，交换后为父结点集为(v_1,v_k)，其余结点为$(v_j,\cdots,v_n)$；加入和删除部分是将新结点随机加入父结点集或将结点从父结点集合中随机删除，例如：初始父结点集为(v_1,v_j)，其余结点为$(v_k,\cdots,v_n)$，加入后是父结点集为(v_1,v_j,v_k)，其余结点为$(v_{k+1},\cdots,v_n)$，删除后是父结点集为(v_j)，其余结点为$(v_{j+1},\cdots,v_n)$。

4.5 能量优化函数

能量优化函数就是需要进行优化计算的目标函数，其最小点为所求的最优解。由于模拟退火算法的最优解是全局最小值，在本文求最大相互信息函数中只需将优化函数取附加约束的最大相互信息准则函数的负值即可，即 $E=-Score_{MMI-L}(S)$。若 $E_{new}<E_{old}$则接受 S_{new}为新的状态并更新温度参数 T_i 的值，否则以概率 $P\leftarrow exp((E_{old}-E_{new})/T_i)$接受。具体做法是产生0到1之间的随机数 d，如 $P>d$ 则接受 $Pa_{new}(v_i)$为结点 v_i 新的父结点集，能量状态为 E_{new}，否则仍保留原来变化前的 $Pa(v_i)$为结点 v_i 的父结点集，能量状态为 E_{old}。

5 实验结果及分析

本文对 Chickering 等人提出的模拟退火算法、Larranaga 等提出的遗传算法进行了对比实验。实验中，本文提出的结构学习的模拟退火算法采用基于附加约束的最大信息熵原则

(MMI-L),简称为(SA-MMI-L);Chickering 等人提出的模拟退火算法采用贝叶斯记分(SA-BDE)、Larranaga 的遗传算法采用了 MDL 记分(GA-MDL)。实验选取 Alarm 数据集作为训练样本集,选取以下5种指标来评价结构学习的性能:

1)20次运行结果的基于附加约束的平均相互信息 AMMI-L,该指标反映的是学习结果与实际训练样本的接近程度。MMI-L 值越大,表示网络拓扑结构越接近实际样本。

2)20次运行结果的平均相互信息 AMI(Average Mutual Information),该指标反映的是学习结果与实际训练样本的接近程度。AMI 值越大,表示网络拓扑结构越接近实际样本。

3)20次运行结果的平均最小描述长度 AMDL(Average MDL),该指标反映的是学习结果与实际训练样本的接近程度。AMDL 值越小,表示网络拓扑结构越接近实际样本。

4)20次运行结果的平均出错边数 AMA,其中出错边数为缺失边数、错误方向边数、多余边数的总和。

5)20次运行结果中找到正确网络拓扑结构的次数 TNN (True Network Numbers)。

Alarm 数据集来源于 Beinlich,见 http://www2.sis.pitt.edu/~genie。该数据集表示了一个用于病人监测的医学诊断系统,由37个变量组成,每个变量可以有2到4个状态,共20000条记录,正确拓扑结构如图1所示,对应于该拓扑结构的 MDL 记分为:1384176.35、MI 记分为17.801443、MMI-L 为:17.22348。实验中我们选取1000个记录作为训练样本。

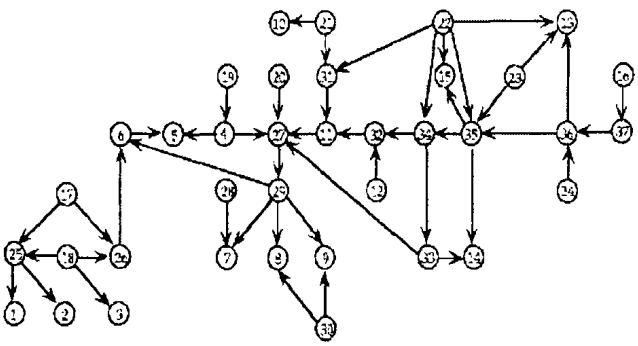

图1 Alarm 网络拓扑结构

5.1 学习算法的精确度比较

采用 SA-MMI-L 算法在训练样本数目分别为200、500和1000时,不同初始温度和不同温度下降系数条件下的学习实验结果如表1所示(20次运行的平均值)。

表1 不同样本时的结构学习实验结果比较

样本数目	$T_0=1, r=0.80$		$T_0=5, r=0.85$	
	MMI-L 记分	出错边数	MMI-L 记分	出错边数
200	17.20747	3.0	17.20892	2.7
500	17.21455	2.7	17.21543	2.6
1000	17.21871	2.5	17.21981	2.5
样本数目	$T_0=10, r=0.90$		$T_0=20, r=0.95$	
	MMI-L 记分	出错边数	MMI-L 记分	出错边数
200	17.20958	2.7	17.20958	2.7
500	17.21581	2.6	17.21581	2.6
1000	17.22157	2.45	17.22157	2.45

使用 SA-MMI-L 算法与其它算法比较,其中 SA-MMI-L 算法初始温度分别取1、5、10,下降系数 r 分别取0.7、0.8、0.9,训练样本为1000个。不同的算法结果比较如表2。

表2 不同算法的结构学习实验结果比较

采用的算法		A-MMI-L	AMMI	AMDL	AMA	TNN
SA-MMI-L	$r=0.9$ $T_0=10$	17.22157	17.8013	1384188.4	0.1	0.6
	$r=0.9$ $T_0=10$	17.22106	17.7661	1384189.1	0.1	0.6
	$r=0.9$ $T_0=5$	17.22074	17.6191	1384199.9	0.15	0.5
	$r=0.8$ $T_0=10$	17.22122	17.6211	1384191.0	0.1	0.6
	$r=0.8$ $T_0=5$	17.21898	17.6187	1384205.1	0.2	0.65
	$r=0.8$ $T_0=1$	17.21871	17.6012	1384212.5	0.25	0.5
	$r=0.7$ $T_0=10$	17.21859	17.6176	1384213.0	0.3	0.7
	$r=0.8$ $T_0=5$	17.21377	17.6133	1384221.6	0.3	0.6
	$r=0.7$ $T_0=1$	17.20159	17.5333	1384229.8	0.3	0.8
SA-BDE		17.15323	17.4732	1384248.9	0.9	1.25
GA-MDL		16.73267	17.0194	1384269.4	2.5	3.3

5.2 时间复杂度比较

学习算法的复杂度可以通过处理相应数目的结点所进行的循环次数和所需 CPU 时间来衡量。循环次数是指计算评价函数所进行的计算次数。采用三种算法在训练样本数目分别为200、500和1000时,算法的时间复杂度比较如表3~5所示。

表3 200个训练样本时的计算量

V	SA_BDE 算法		GA_MDL		SA_MMI_L	
	循环次数	时间	循环次数	时间	循环次数	时间
1	1	1	1	1	1	1
5	25	1	25	1	25	1
10	1382	1	435	1	311	1
15	4367	1	989	1	1345	1
20	12354	2	5324	1	3524	1
25	25668	3	9956	1	8751	1
27	31432	3	17543	2	14372	2
29	39045	4	24593	2	21981	2
31	47865	5	32890	3	26347	3
33	68353	6	45738	4	34468	4
35	97467	8	68230	7	58463	6
37	114253	10	84356	9	72357	7

5.3 实验结果分析

从表1可以看出,在相同训练样本数时,初始温度和下降系数选取得越高,则出错边数就越少,相应的记分越接近实际值;随着样本数目的增加,出错边数减少,相应的记分也越接近实际值;同时,初始温度和下降系数选取的越高、样本数目越大,则计算量也越大。

从表2可以看出,当训练样本数目相同情况下,SA-MMI-L 算法在所有列出初始温度和下降系数下均较其它两种算法要精确,出错边数少于 SA-BDE 算法和 GA-MDL 算法,平均输出正确拓扑结构的次数也多于两者;在样本数目为1000时,

(下转第208页)

度偏差抽样技术可以根据用户的要求快速地从一个大规模数据集中抽取出准确性很高的偏差样本，对孤立点检测算法的快速高效执行提供了充分的保证。

本文证明了密度偏差抽样技术可以应用于孤立点检测这项具体的数据挖掘任务，同时，探索密度偏差抽样技术在其他数据挖掘技术如分类、决策树、关联规则等方面的应用也具有积极的指导作用，这是我们下一步将要研究的课题。

参 考 文 献

1 Han J, Kamber M. Data Mining: Concepts and Techniques. Copyright by Morgan Kaufmann Publishers, Inc. 2001

2 Palmer C R, Faloutsos C. Density biased sampling: An improved method for data mining and clustering. In: Proc. Of the ACM SIGMOD'2000, 2000

3 Guha S, Rastogi R, Shim K. CRUE: An Efficient Clustering Algorithm for Large Database. In: Proc. ACM SIGMOD, June 1998. 73～84

4 Knorr E, Ng R. Algorithms for Mining Distance Based Outliers in Large Databases. In: Proc. Very Large Data Bases Conf., Aug. 1998. 392～403

5 Barnett Y, Lewis T. Outliers in Statistical Data. John Wiley & Sons, 1994

6 Scott D. Multivariate Density Estimation: Theory, Practice and Visualization. Wiley and Sons, 1992

7 Wand M P, Jones M C. Kernel Smoothing. Monographs on Statistics and Applied Probability, Chapman and Hall, 1995

8 Knorr E, Ng R. Algorithms for Mining Distance-based Outliers in Large Datasets. In: Proc. 1998 Int. Conf. Very Large Data Base(VLDB98), New York, 1998(8):392～403

（上接第199页）

较好的初始温度和下降系数为：$T_0=5$和$r=0.9$。从整体上看SA-MMI-L算法具有较好的学习效果。

表4 500个训练样本时的计算量

V	SA-BDE 算法		GA-MDL		SA-MMI-L	
	循环次数	时间	循环次数	时间	循环次数	时间
1	1	1	1	1	1	1
5	30	1	30	1	30	1
10	2483	1	1345	1	1124	1
15	41783	4	12753	1	9637	1
20	81308	8	61635	5	45684	4
25	162593	16	84138	7	58852	5
27	190949	19	121937	11	81351	7
29	220835	22	152018	15	121756	11
31	312638	31	220957	21	173564	16
33	371904	37	278769	26	214683	19
35	414473	41	318546	29	235429	22
37	568256	56	369354	34	307538	27

表5 1000个训练样本时的计算量

V	SA-BDE 算法		GA-MDL		SA-MMI-L	
	循环次数	时间	循环次数	时间	循环次数	时间
1	1	1	1	1	1	1
5	135	1	87	1	65	1
10	7592	1	5247	1	3867	1
15	73756	7	13793	1	11354	1
20	111536	9	51785	5	43659	4
25	164565	17	81178	7	77860	7
27	363745	35	109935	11	94681	9
29	561085	58	153426	15	138213	13
31	648362	64	269928	27	238156	22
33	719620	73	348667	35	338907	33
35	873247	87	418546	41	407679	40
37	1175585	109	574356	57	527640	51

从表3～5可以看出，当样本数目较小时，三种算法所用时间基本相同，当样本数量增大时，SA-MMI-L算法和GA-MDL算法速度较快，SA-MMI-L算法较GA-MDL算法稍好。

结论 近年来，基于相互信息熵理论应用于贝叶斯网络结构学习得到了广泛关注，附加网络复杂度约束的最大熵记分原则能够很好地用于贝叶斯网络结构的评估，最大相互信息保证了变量间信息量最大，附加的约束函数使得网络拓扑结构不过于复杂，是逼近程度和复杂度之间取折衷的最合适的评价方法；模拟退火算法是传统的智能优化方法，通过设计不同的优化函数、邻近值产生机制、温度下降策略以及算法终止和结束条件，可以将模拟退火算法应用于贝叶斯网络的结构学习，实验表明该算法在计算复杂度和学习精度上都具有较好效果。

参 考 文 献

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

2 Cooper G F, Herskovits E. A Bayesian Method for the induction of probabilistic networks from data. Machine Learning, 1992, 9: 309～347

3 Suzuki J. A Construction of Bayesian Networks from Databases Based on an MDL Principle. In: Proc. of the 9th Conf. on Uncertainty in Artificial Intelligence. Washington D. C., 1993. 266～273

4 Lam W, Bacchus F. Using Causal Information and Local Measures to Learning Bayesian Networks. In: Proc. of the 9th Conf. on Uncertainty in Artificial Intelligence. Washington D. C., 1993. 243～250

5 Suzuki J. Learning Bayesian belief networks based on the MDL principle: an efficient algorithm using the branch and bound technique. In: Proc. of the Intl. Conf. on Machine Learning, Bally, Italy, 1996

6 Suzuki J. Learning Bayesian Belief Networks Based on the MDL Principle: An Efficient Algorithm Using the Branch and Bound Technique TIEICE. IEICE Transactions on Communications/Electronics/Information and Systems, 1998. 1～12

7 Suzuki J. Learning Bayesian Belief networks Based on the Minium Description Length Priciple: Basic Properties. IEEE Transactions on Fundamentals, 1999, E82-A(9)

8 Chickering D, Geiger D, Heckerman D. Learning bayesian networks: Search methods and experimental results. In: Proc. of the fifth conf. on Artificial Intelligence and Statistics, IEEE Press, 1995. 112～128

9 Larranaga P, et al. Structure Learning of Bayesian Networks by Genetic Algorithms: Aperformance analysis of Control parameters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18b(9): 912～926

10 Wong M L, Lam W, Leung K-S. Using Evolutionary Programming and Minimum Description Length Principle for Data Mining of Bayesian Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(2): 174～178

11 李刚. 知识发现的图模型方法. 中国科学院软件研究所：[博士学位论文]. 2001