具有丢失数据的贝叶斯网络结构学习研究*

王双成+, 苑森淼

(吉林大学 计算机科学与技术学院,吉林 长春 130025)

Research on Learning Bayesian Networks Structure with Missing Data

WANG Shuang-Cheng+, YUAN Sen-Miao

(College of Computer Science and Technology, Jilin University, Changchun 130025, China)

+ Corresponding author: Phn: +86-431-5701777, E-mail: wangsc@nenu.edu.cn, http://www.jlu.edu.cn

Received 2003-01-20; Accepted 2003-09-15

Wang SC, Yuan SM. Research on learning Bayesian networks structure with missing data. *Journal of Software*, 2004,15(7):1042~1048.

http://www.jos.org.cn/1000-9825/15/1042.htm

Abstract: At present, the method of learning Bayesian network structure with missing data is mainly based on the search and scoring method combined with EM algorithm. The algorithm has low efficiency and easily gets into local optimal structure. In this paper, a new method of learning Bayesian network structure with missing data is presented. First, unobserved data are randomly initialized. As a result, a complete data set is got. Based on the complete data set, the maximum likelihood tree is built as an initialization Bayesian network structure. Second, unobserved data are reassigned by combining Bayesian network with Gibbs sampling. Third, on the basis of the new complete data set, the Bayesian network structure is regulated based on the basic dependency relationship between variables and dependency analysis method. Finally, the second and third steps are iterated until the structure goes stable. This method can avoide the exponential complexity of standard Gibbs sampling and the main problems in the existing algorithm. It provides an effective and applicable method for uncertain knowledge representation, inference, and reasoning with missing data.

Key words: Bayesian network; structure learning; missing data; Gibbs sampling; dependency analysis; maximum likelihood tree

摘 要: 目前主要基于 EM 算法和打分-搜索方法进行具有丢失数据的贝叶斯网络结构学习,算法效率较低,而且易于陷入局部最优结构.针对这些问题,建立了一种新的具有丢失数据的贝叶斯网络结构学习方法.首先随机初始化未观察到的数据,得到完整的数据集,并利用完整数据集建立最大似然树作为初始贝叶斯网络结构,然后进行迭代学习.在每一次迭代中,结合贝叶斯网络结构和 Gibbs sampling 修正未观察到的数据,在新的完整数据集的基础上,基于变量之间的基本依赖关系和依赖分析思想调整贝叶斯网络结构,直到结构趋于稳定.该方法既

* Supported by the National Natural Science Foundation of China under Grant No.60275026 (国家自然科学基金); the Natural Science Foundation of Jilin Province of China under Grant No.20030517-1 (吉林省自然科学基金)

作者简介: **王双成**(1958－),男,吉林长春人,博士,副教授,主要研究领域为数据库,数据采掘;**苑森淼**(1943－),男,教授,博士生导师,主要研究领域为数据库,人工智能,计算机网络系统.

解决了标准 Gibbs sampling 指数复杂性问题,又避免了现有学习方法所存在的主要问题,为具有不完整数据的不确定性知识表示、推断和推理提供了有效和可行的方法.

关键词: 贝叶斯网络;结构学习;丢失数据;Gibbs 抽样;依赖分析;最大似然树

中图法分类号: TP18 **文献标识码**: A

由于各种原因(系统外部和内部),在许多现实数据库中存在着丢失数据的现象.数据丢失可能与某些变量的状态有关[1],这时丢失的数据蕴含信息.本文只讨论随机丢失数据的情况.目前,具有完整数据的贝叶斯网络结构学习方法比较成熟,可分为两类,一类是基于打分-搜索的学习方法,另一类是基于依赖分析的学习方法.两类方法各有特点.打分-搜索方法过程简单、规范,但由于搜索空间大,一般要求结点有顺序,并根据打分函数的可分解性进行局部确定或随机搜索(完全搜索是 N-P 困难问题[2]),效率较低,且易于陷入局部最优结构.依赖分析方法过程比较复杂,但在一些假设下学习效率较高,而且能够获得全局最优结构.具有丢失数据的贝叶斯网络结构学习更加困难,现有的研究主要基于打分-搜索方法.数据的丢失导致两方面问题的出现,一方面,打分函数不再具有可分解形式,不能进行局部搜索;另一方面,一些充分统计因子不存在,无法直接进行结构打分.围绕这两个问题相继发展了一些解决的方法.Hecherman 等人[1~6]给出了解决后一个问题的一些方法.这些方法对选择的贝叶斯网络结构首先基于梯度的优化(gradient-based optimization)或 EM(expectation-maximization)算法进行最大后验参数估计,然后使用拉普拉斯近似(Laplace approximation)或贝叶斯信息标准(Bayesian information criterion)等大样本近似方法进行近似结构打分,由于搜索空间大以及存在近似打分的误差,使学习效率较低,且结果不够可靠.Friedman 等人[7,8]改进了上述方法,基于 EM 算法框架进行具有丢失数据的贝叶斯网络结构学习,使用期望充分统计因子代替不存在的充分统计因子,在一些假设下,可使打分函数具有可分解形式(可进行局部搜索),并且在每一次迭代中,结构都有所改进,使结构序列收敛.该方法能够一定程度地提高学习效率,但一般是收敛到局部最优结构.

本文建立了一种新的具有丢失数据的贝叶斯网络结构学习方法——BN-GS(Bayesian network & Gibbs sampling).该方法使用 Gibbs sampling[9,10]修复丢失的数据,基于依赖分析方法进行贝叶斯网络结构学习和调整.首先随机初始化丢失的数据,并建立最大似然树[11]作为初始贝叶斯网络结构,然后进行数据集和贝叶斯网络的迭代修正-调整,直到结构趋于稳定或满足给定的终止条件为止.每一次数据集修正后进行贝叶斯网络结构调整,使调整后的贝叶斯网络适合于当前的数据集,并且不会陷入局部最优结构.Gibbs sampling 迭代收敛到平稳分布[9,10],因此结构序列将收敛到平稳分布的贝叶斯网络结构.联合概率可按贝叶斯网络结构进行分解,对一个变量的抽样只需考虑对应的条件概率因子即可,而条件概率因子中条件变量的数量与所有变量的数量没有联系,解决了满条件分布(full conditional distribution)所带来的问题,从而能够显著提高抽样效率.

用 $X_1,...,X_n$ 表示离散随机变量,简称为变量,$x_1,...,x_n$ 为其值.数据库 D 中具有 N 个记录,假设数据是独立地产生于某一概率分布 P,数据丢失是随机的.在概率模式中的变量和表示概率模式的图形模式中的结点有时不加区分.

1 数据集和贝叶斯网络的初始化

首先随机初始化丢失的数据,把得到的数据集作为初始数据集($D^{(0)}$).然后按某一顺序依次计算两个变量之间的互信息,并由大到小排序,依据不产生环路的原则按顺序依次添加边,直到添加 $n-1$ 条边为止.选择一个结点作为根结点,由根结点向外的方向为边定向,并计算最大似然树参数,把最大似然树作为初始贝叶斯网络($G^{(0)}$, $\theta^{(0)}$).

丢失的数据被随机初始化后,数据中所蕴含的变量之间依赖关系可能比较混乱,直接进行贝叶斯网络结构学习会使得到的贝叶斯网络结构过于复杂,由复杂的贝叶斯网络结构收敛到平稳分布的贝叶斯网络结构效率较低.最大似然树是和贝叶斯网络拟合得最好的树形结构,其结构简单且稳定,选择作为初始贝叶斯网络将使贝叶斯网络结构序列由最大似然树收敛到平稳分布的贝叶斯网络结构,能够提高学习效率.

2 迭代数据修正与结构调整

每一次迭代包括两个部分:一部分是利用贝叶斯网络进行抽样,修正待修正的数据和参数;另一部分是使用修正后的数据集调整贝叶斯网络.迭代产生 3 个序列,分别是数据集序列($D^{(k)}$)、结构序列($G^{(k)}$)和参数向量序列($\theta^{(k)}$).在每一次迭代中,首先利用贝叶斯网络($G^{(k)}$,$\theta^{(k)}$)进行抽样,修正数据集 $D^{(k)}$得到 $D^{(k+1)}$,在修正数据集的同时修正参数.然后使用 $D^{(k+1)}$调整贝叶斯网络结构 $G^{(k)}$得到 $G^{(k+1)}$,并根据新结构和数据集调整参数,由 $D^{(k+1)}$和 $G^{(k+1)}$得到$\theta^{(k+1)}$实现一次迭代,当结构不发生变化或满足终止条件时结束迭代.

2.1 使用$G^{(k)}$和$\theta^{(k)}$修正数据集$D^{(k)}$

由 $G^{(k)}$和$\theta^{(k)}$所决定的联合分布为

$$p^{(k)}(x_1,...,x_n)=\prod_{i=1}^{n}p^{(k)}(x_i\mid\pi_{x_i},\theta^{(k)},G^{(k)}) \tag{1}$$

其中 π_{x_i} 为变量 X_i 父结点集 Π_{X_i} 的配置,参数 $\theta_{ijh}^{(k)}=p^{(k)}(x_i^h\mid\pi_{x_i}^j,\theta^{(k)},G^{(k)})$, x_i^h 为 X_i 取第 h 个值, $\pi_{x_i}^j$ 为 Π_{X_i} 的配置索引后的第 j 个配置.按照结构 $G^{(k)}$所决定的变量顺序和数据库中记录的顺序依次对具有丢失数据的变量进行抽样,并用抽样值修正待修正的值.每一次迭代修正数据集的次数不宜过多(可在 1~10 之间),过多的修正次数会使数据集和贝叶斯网络过度拟合,导致结构序列收敛到局部最优结构.

设 X_i 在第 m 个记录具有待修正值 x_{im},修正后的值为 $\hat{x}_{im}$,变量 X_i 的可能取值为 $x_i^1,...,x_i^{r_i}$.用 $D^{(k)}=D_{(1,1)}^{(k)}$ 表示修正前的数据集, $D_{(i,m)}^{(k)}$ 表示在修正数据 x_{im} 之前的最新数据集, $D^{(k+1)}=D_{(1,N+1)}^{(k)}$ 表示修正后的数据集.修正一次数据集的时间复杂性是 $O(nN)$.

(1) 修正数据集

不存在零概率:对任何 x_i 都有 $\hat{p}^{(k)}(x_i\mid\pi_{x_{im}},D_{(i,m)}^{(k)})>0$, $\pi_{x_{im}}$ 表示在第 m 个记录变量 X_i 父结点集的配置,生成随机数λ,则

$$\hat{x}_{im}=\begin{cases}x_i^1, & 0<\lambda\le\hat{p}^{(k)}(x_i^1\mid\pi_{x_{im}},D_{(i,m)}^{(k)})\\ \cdots\cdots & \\ x_i^h, & \sum_{j=1}^{h-1}\hat{p}^{(k)}(x_i^j\mid\pi_{x_{im}},D_{(i,m)}^{(k)})<\lambda\le\sum_{j=1}^{h}\hat{p}^{(k)}(x_i^j\mid\pi_{x_{im}},D_{(i,m)}^{(k)})\\ \cdots\cdots & \\ x_i^{r_i}, & \lambda>\sum_{j=1}^{r_i-1}\hat{p}^{(k)}(x_i^j\mid\pi_{x_{im}},D_{(i,m)}^{(k)})\end{cases} \tag{2}$$

存在零概率:设

$$\hat{p}^{(k)}(x_i^u\mid\pi_{x_{im}},D_{(i,m)}^{(k)})=0,$$
$$\hat{p}^{(k)}(x_i^v\mid\pi_{x_{im}},D_{(i,m)}^{(k)})>0,$$
$$u\in\{u_1,...,u_s\},v\in\{v_1,...,v_t\},$$
$$\{u_1,...,u_s\}\cup\{v_1,...,v_t\}=\{1,...,r_i\},s+t=r_i.$$

对 $\hat{p}^{(k)}(x_i^u\mid\pi_{x_{im}},D_{(i,m)}^{(k)})$ 进行拉普拉斯修正(Laplace-corrected)[12],

$$\hat{p}^{(k)}(x_i^u\mid\pi_{x_{im}},D_{(i,m)}^{(k)})=(1/N)/(N(\pi_{x_{im}})+N(x_i^u)(1/N)),$$

其中 $N(\pi_{x_{im}})$ 为 X_i 的父结点集 Π_{X_i} 具有配置 $\pi_{x_{im}}$ 的例子数量, $N(x_i^u)$ 为 $X_i=x_i^u$ 的例子数量,并进行归一化处理,记 $w(h)=\dfrac{\hat{p}^{(k)}(x_i^h\mid\pi_{x_{im}},D_{(i,m)}^{(k)})}{\sum_{j=1}^{s}\hat{p}^{(k)}(x_i^j\mid\pi_{x_{im}},D_{(i,m)}^{(k)})+\sum_{j=1}^{t}\hat{p}^{(k)}(x_i^j\mid\pi_{x_{im}},D_{(i,m)}^{(k)})}$, $h\in\{1,...,r_i\}$,生成随机数λ,则

$$\hat{x}_{im}=\begin{cases}x_i^1, & 0<\lambda\le w(1)\\ \cdots\cdots & \\ x_i^h, & \sum_{j=1}^{h-1}w(j)<\lambda\le\sum_{j=1}^{h}w(j)\\ \cdots\cdots & \\ x_i^{r_i}, & \lambda>\sum_{j=1}^{r_i-1}w(j)\end{cases} \tag{3}$$

(2) 局部概率(参数)修正

如果 $x_{im}\neq\hat{x}_{im}$,需要修正对应的参数,利用修正后的参数进行抽样,修正下一个待修正的数据.参数修正方法如下:

$$\hat{p}^{(k)}(x_{im}\mid\pi_{x_{im}},D^{(k)}_{(i+1,m)})=\hat{p}^{(k)}(x_{im}\mid\pi_{x_{im}},D^{(k)}_{(i,m)})-1/N \tag{4}$$

$$\hat{p}^{(k)}(\hat{x}_{im}\mid\pi_{x_{im}},D^{(k)}_{(i+1,m)})=\hat{p}^{(k)}(\hat{x}_{im}\mid\pi_{x_{im}},D^{(k)}_{(i,m)})+1/N \tag{5}$$

2.2 使用修正后的数据集调整贝叶斯网络结构

利用贝叶斯网络的信息管道模型[13]描述变量之间存在的 3 种基本依赖关系(边)[14]:(1) 及物依赖(transitive dependencies),表示变量的结点之间存在直接的信息流动,而且信息流不能被其他结点所阻塞,即结点所表示的变量之间条件不独立;(2) 非及物依赖(non-transitive dependencies),结点之间不存在直接的信息流动,而是由连接两结点之间的开路(不含碰撞结点[14]的链路)产生的信息流,能被切割集[13]中的结点所阻塞,即以切割集中结点表示的变量为条件时,两个结点所表示的变量之间条件独立;(3) 诱发依赖(induced dependencies),这种依赖是由 V 结构[14,15]所导致的,结点之间不存在直接的信息流动,是 V 结构中的碰撞结点诱发的信息流,结点所表示的变量之间无条件独立,以切割集(不包括诱发结点)中结点表示的变量为条件时,两个结点所表示的变量之间也条件独立.建立贝叶斯网络结构就是在错综复杂的依赖关系中,在有效地保留第 1 种依赖关系的同时去除第 2 种和第 3 种依赖关系.

使用互信息和条件互信息进行变量之间定量条件独立性检验,分别用 $I(X_i,X_j)$ 和 $I(X_i,X_j\mid X_{n_1},...,X_{n_s})$ 表示变量 X_i 和 X_j 之间的互信息以及以 $X_{n_1},...,X_{n_s}(n_h\neq i,j,h=1,...,s)$ 为条件的条件互信息,对给定的小正数 ε,如果 $I(X_i,X_j\mid X_{n_1},...,X_{n_s})<\varepsilon$,就认为 X_i 和 X_j 之间条件独立. 使用修正后的数据集 $D^{(k+1)}$ 对贝叶斯网络结构 $G^{(k)}$ 按如下 3 个步骤进行调整,得到 $G^{(k+1)}$.调整贝叶斯网络结构的时间复杂性是 $O(n^3)$.

2.2.1 第 1 种边的存在性调整

把 $G^{(k)}$ 作为初始贝叶斯网络结构,设结点的顺序为 $X_1^{(k)},...,X_n^{(k)}$,对不存在弧的结点对依次进行条件独立性检验.对选择的结点对 $X_i^{(k)},X_j^{(k)}(i<j)$,用 $D_{X_j^{(k)}}(X_i^{(k)},X_j^{(k)})$ 表示 $X_j^{(k)}$ 的父结点集中在 $X_i^{(k)}$ 和 $X_j^{(k)}$ 链路上的结点集合.如果 $I(X_i^{(k)},X_j^{(k)}\mid D_{X_j^{(k)}}(X_i^{(k)},X_j^{(k)}),D^{(k+1)})>\varepsilon$,增加弧 $X_i^{(k)}\to X_j^{(k)}$.这一过程最多需要 $n(n-1)/2$ 次条件互信息计算,调整后的结构记为 M_1.

命题 1. 在 M_1 中包括所有的第 1 种边.

证明:设在 M_1 中 $X_i^{(k)}$ 和 $X_j^{(k)}$ 之间不存在边,那么 $D_{X_j^{(k)}}(X_i^{(k)},X_j^{(k)})$ 能够 d-separate[15] $X_i^{(k)}$ 和 $X_j^{(k)}$,已知 $X_i^{(k)}$ 和 $X_j^{(k)}$ 之间不存在第 1 种边,因此 M_1 中包括所有的第 1 种边. □

2.2.2 弧的方向与第 3 种边的调整

(1) 碰撞识别调整方向

按顺序依次检验每一对结点,对满足 $I(X_i^{(k)},X_i^{(k)})<\varepsilon$ 的结点对进行碰撞识别检验.设与 $X_i^{(k)}$ 和 $X_j^{(k)}$ 所有可能形成 V 结构的结点为 $X_{m_1}^{(k)},...,X_{m_t}^{(k)}(m_h\neq i,j,h=1,...,t)$.对给定的阈值 $\delta>0$,如果 $\dfrac{I(X_i^{(k)},X_i^{(k)}\mid X_{m_h}^{(k)})}{I(X_i^{(k)},X_i^{(k)})}>(1+\delta)$,则 $X_i^{(k)},X_j^{(k)}$ 和 $X_{m_h}^{(k)}$ 形成 V 结构,定向为 $X_i^{(k)}\to X_{m_h}^{(k)}$ 和 $X_j^{(k)}\to X_{m_h}^{(k)}$,并删除 $X_i^{(k)}$ 和 $X_j^{(k)}$ 之间的弧(如果存在).当碰撞识别结束后,依据 V 结构和无环性特征再调整部分弧的方向,不能使用 V 结构定向的弧标记为待定向弧.这一

过程最多需要 $n(n-1)(n-2)/2$ 次条件互信息计算.

(2) MDL 打分调整方向

对于不能使用碰撞识别调整方向的待定向弧,使用 MDL(minimal description length)标准[16]进行局部搜索-打分确定方向.调整结束后,如果有方向的变化,对结点重新排序.这一过程最多需要 $n(n-1)/2$ 次 MDL 计算.对 M_1 中的弧进行方向调整后得到的有向无环图记为 M_2.

2.2.3 第 2 种边的存在性调整

仍用 $X_1^{(k)},...,X_n^{(k)}$ 表示结点顺序,对存在弧的结点对依次进行条件独立性检验.对选择的结点对 $X_i^{(k)},X_j^{(k)}(i<j)$,用 $D_{X_j^{(k)}}(X_i^{(k)},X_j^{(k)})$ 表示 $X_j^{(k)}$ 的父结点集中在 $X_i^{(k)}$ 和 $X_j^{(k)}$ 链路上的结点集,如果 $I(X_i^{(k)},X_j^{(k)}\mid D_{X_j^{(k)}}(X_i^{(k)},X_j^{(k)}),D^{(k+1)})<\varepsilon$,删除 $X_i^{(k)}$ 和 $X_j^{(k)}$ 之间的弧.这一过程最多需要 $n(n-1)/2$ 次条件互信息计算,对 M_2 调整后的有向无环图记为 M_3.

命题 2. 在 M_3 中不存在第 2 种边,也不丢失第 1 种边.

证明:由 $X_j^{(k)}$ 不是 $X_i^{(k)}$ 的祖先结点可知, $X_j^{(k)}$ 的父结点集能够 d-separate $X_i^{(k)}$ 和 $X_j^{(k)}$, $X_j^{(k)}$ 的父结点集在 $X_i^{(k)}$ 和 $X_j^{(k)}$ 链路上的子集也是如此,因此在 M_3 中不存在第 2 种边.由于 $D_{X_j^{(k)}}(X_i^{(k)},X_j^{(k)})$ 不能 d-separate 具有第 1 种依赖的结点对 $X_i^{(k)},X_j^{(k)}$,故 M_3 中不会丢失第 1 种边. □

命题 3. 假设 V 结构是可识别的,则 M_3 的骨架[15](不考虑弧方向的结构)是 $D(k+1)$ 的贝叶斯网络结构的骨架.

证明:经过第 1 步增加了所有的第 1 种边,由 V 结构的可识别性假设,第 2 步去除了所有的第 3 种边,第 3 步删除了所有的第 2 种边,且不丢失第 1 种边,因此经过 3 步调整得到的 M_3 的骨架是贝叶斯网络结构的骨架. □

V 结构是贝叶斯网络结构中最重要和最具特色的基本结构,是打分-搜索和依赖分析方法的基础,许多文献中具有 V 结构可识别性假设,如文献[13~15],V 结构一般是可识别的.不可识别的特殊情况非常少见,这时弧的方向对贝叶斯网络的影响较小,关于 V 结构的详细论述参见文献[14].

命题[15]. 两个贝叶斯网络结构等价(表示同一个概率分布)的充分必要条件是它们具有相同的 V 结构和骨架.

命题 4. 经过反复迭代修正-调整而得到的贝叶斯网络结构序列收敛到平稳分布的贝叶斯网络结构.

证明:由于每次数据修正后要对贝叶斯网络结构进行调整,根据命题 1~命题 3 和命题,经过调整可以得到当前数据集的贝叶斯网络结构,则按照贝叶斯网络结构所决定变量的顺序依次使用 $p(x_i\mid\pi_{x_i},D^{(k)})$ 进行抽样等价于使用满条件分布 $p(x_i\mid\{x_1,...,x_n\}-\{x_i\})$ 进行抽样,Gibbs sampling 迭代收敛到平稳分布,因此贝叶斯网络结构序列收敛到平稳分布的贝叶斯网络结构. □

2.3 调整贝叶斯网络参数

每一个待修正数据修正结束后,如果有变化,对应的参数已被修正,因此,只需根据 $D(k+1)$ 重新计算父结点集发生变化变量的参数即可,得到新的参数向量 $\theta(k+1)$,实现一次迭代.

3 实 验

根据网站 http://www.norsys.com 提供的 ALARM 网概率分布表生成用于实验的模拟数据.

生成具有 4 000 个例子的 4 个模拟数据集,并随机产生具有 10%,20%,30%,40%的丢失数据.分别选择 25 对具有相对较弱第 1 种依赖的变量对和随机选择 25 对具有第 2 种依赖的变量对.进行 10 次迭代,每一次迭代修正数据集 5 次.随机初始化和迭代修正-调整后变量之间依赖关系的变化情况如图 1 所示.

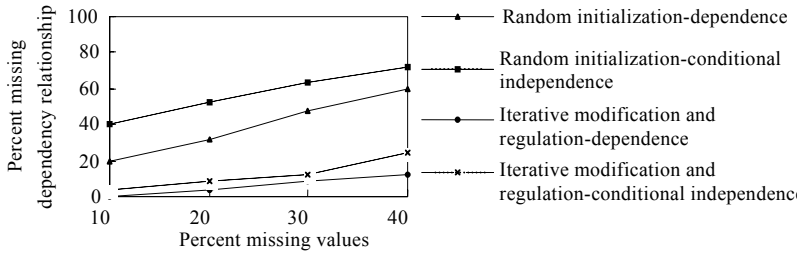

Fig.1 The influence of missing data to dependency relationship between variables

图 1 丢失数据对变量之间依赖关系的影响

由图 1 可以看出,丢失数据对变量之间依赖关系的影响随丢失数据比例的增加而增大,从不经过修正的数据集学习得到的贝叶斯网络结构将是不可靠的,而经过迭代修正-调整得到的贝叶斯网络结构能够很好地反映变量之间的依赖关系.其主要原因是,在迭代修正-调整过程中,观察到的数据中所蕴含的变量之间的依赖信息能够得到有效的利用,并使变量之间的依赖关系不断得到改进,直到结构趋于稳定(平稳分布的贝叶斯网络结构).

生成具有 500,1 000,2 000 例子的训练数据和 10 000 个例子的测试数据.在训练集中分别随机产生具有 10%,20%,30%的丢失数据.随机初始化丢失数据,进行与上例相同的修正-调整,实验结果与 MS-EM[8]算法的比较如图 2 所示.

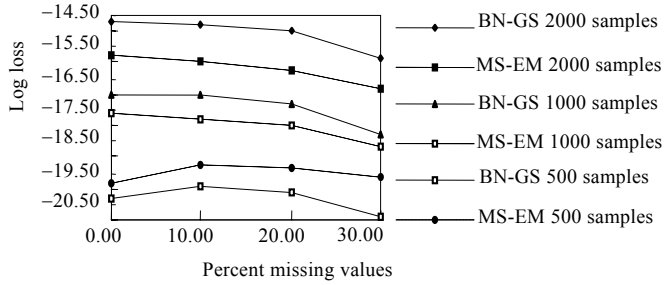

Fig.2 The comparison of experimental result on BN-GS algorithm and MS-EM algorithm

图 2 BN-GS 算法与 MS-EM 算法实验结果比较

图 2 显示出,当训练数据集较小时,局部极值的影响并不明显,变量之间的依赖和条件独立性信息不够可靠,因此,MS-EM 算法更准确.随着训练集的增大,局部极值的影响逐渐增大,变量之间的依赖和条件独立性信息的可靠性逐渐增加,BN-GS 算法的优势也逐渐明显.可见,BN-GS 算法适合于大数据集具有丢失数据的贝叶斯网络结构学习.

4 结 语

文中建立的具有丢失数据贝叶斯网络结构学习方法,把 Gibbs sampling 和数据集修正与贝叶斯网络结构调整有机地结合在一起.一方面,由 Gibbs sampling 过程的收敛性保证了贝叶斯网络结构序列的收敛性;另一方面,每一次根据联合概率的分解式依次独立地进行抽样,而不是使用满条件分布进行抽样,能够显著提高抽样效率.由于 BN-GS 方法能够有效地修复缺损数据,因此也可用于其他具有丢失数据的数据采掘问题.根据修正后的数据集调整贝叶斯网络结构可能存在一些重复计算,优化对贝叶斯网络结构的调整是进一步的研究课题.

References:

[1] Heckerman D. Bayesian networks for data mining. Technical Report, MSR-TR-97-02, Microsoft Research, Redmond, 1997.

[2] Chickering DM. Learning Bayesian networks is NP-Hard. Technical Report, MSR-TR-94-17, Microsoft Research, Redmond, 1994.

[3] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Stiatistical Society, 1977,39(1):1~38.

[4] Binder J, Koller D, Russell S, Kanazawa K. Adaptive probabilistic networks with hidden variables. Machine Learning, 1997,29 (2-3):213~244.

[5] Chickering DM, Heckerman D. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. Machine Learning, 1997,29(2-3):181~212.

[6] Liu DY, Wang F, Lu YN, Xue WX, Wang SX. Research on learning Bayesian network structure based on genetic algorithm. Journal of Computer Research and Development, 2001,38(8):916~922 (in Chinese with English abstract).

[7] Friedman N. Learning belief networks in the presence of missing values and hidden variables. In: Proc. of the 14th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997. 125~133.

[8] Friedman N. The Bayesian structural EM algorithm. In: Proc. of the 14th Int'l Conf. on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1998. 129~138.

[9] Geman S, Geman D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1984,6(6):721~742.

[10] Mao SS, Wang JL, Pu XL. Advanced Mathematical Statistics. Beijing: Higher Education Press, Berlin: Springer-Verlag, 1998. 401~459 (in Chinese).

[11] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning, 1997,29(2-3):131~161.

[12] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 1997,29(2-3): 103~130.

[13] Cheng J, Bell D, Liu WR. Learning Bayesian networks from data: An efficient approach based on information-theory. Artificial Intelligence, 2002,137(1-2):43~90.

[14] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo: Morgan Kaufmann Publishers, 1988. 117~133.

[15] Chickering DM. Learning equivalence classes of Bayesian network structures. Machine Learning, 2002,2(2):445~498.

[16] Lam W, Bacchus F. Learning Bayesian belief networks: An approach based on the MDL principle. Computational Intelligence, 1994,10(4):269~293.

附中文参考文献:

[6] 刘大有,王飞,卢奕南,薛万欣,王松昕.基于遗传算法的 Bayesian 网结构学习研究.计算机研究与发展,2001,38(8):916~922.

[10] 茆诗松,王静龙,濮晓龙.高等数理统计.北京:高等教育出版社,柏林:施普林格出版社,1998.401~459.