小数据集的贝叶斯网络结构学习

王双成 [1, 2] 冷翠平 [1] 李小琳 [3]

摘 要 针对直接基于小数据集贝叶斯网络结构学习不可靠, 以及目前对小数据集的处理只强调扩展而忽略对扩展数据的修正等, 提出了将扩展与修正相结合的小数据集处理机制, 以及在此基础上的基于结点排序和局部打分–搜索的贝叶斯网络结构学习方法. 可不需要完全结点顺序的先验知识, 但能够结合专家的部分结点顺序信息. 实验结果显示了这种方法的有效性和可靠性.

关键词 贝叶斯网络, 小数据集, 结构学习, 最大似然树, 吉布斯抽样
中图分类号 TP18

Learning Bayesian Network Structure from Small Data Set

WANG Shuang-Cheng[1, 2] LENG Cui-Ping[1] LI Xiao-Lin[3]

Abstract It is incredible to learn Bayesian network structure directly from small data set. For improving the reliability, many methods of extending small data set have been developed, but the revision of extended data is neglected. In this paper, extending small data set is combined with revising extended data to upswing the data reliability. A directed tree is built from the small data set and variables are sorted according to it. On the basis of the variable order, a Bayesian network structure can be established based on the local search and scoring method. This method dose not need the prior knowledge of the variable order, but the partial order information of expert can be used properly. Experimental results show that this method can effectively learn Bayesian network structure from a small data set.

Key words Bayesian network, small data set, structure learning, maximal likelihood tree, Gibbs sampling

贝叶斯网络 (Bayesian network)[1] 是描述随机变量之间依赖关系的图形模式, 由结构 (有向无环图) 和参数 (条件概率分布) 两部分构成, 分别用于定性和定量不确定性知识表示和推理, 是处理不确定性问题的有力工具, 已成为模式识别、人工智能和机器学习等领域的研究热点之一.

基于贝叶斯网络解决实际问题的基础是贝叶斯网络学习, 包括结构学习和参数学习, 结构学习是研究的核心. 近十几年以贝叶斯网络结构学习为主线, 相继发展了许多著名的算法[2−7], 这些算法已得到了广泛的应用, 但它们主要针对完整或类完整数据集 (具有丢失数据, 但丢失数据被填充后便可得到完整数据集), 对普遍存在的小数据集往往不具有实用性. 由于小数据集所蕴含的信息不充分, 在进行贝叶斯网络结构学习时, 大量充分统计因子为零, 使得直接学习的可靠性无法得到保障. 一些研究者从两个方面对小数据集贝叶斯网络学习进行过探索. 一方面是通过扩展小数据集来提高贝叶斯网络学习的可靠性, 另一方面是在一些假设下直接从小数据集进行贝叶斯网络学习和推理. Friedman 等[8−9] 使用基于数据集本身的 Bootstrap 抽样扩展小数据集, 并基于扩展后的数据集进行贝叶斯网络结构学习. 基于数据集本身的 Bootstrap 抽样是一种以记录为基本单位的可重复性抽样, 由于没有增加额外信息, 使得充分统计因子估计得不到实质性的改进, 因此效果并不理想. Heckerman 等[10] 提出了朴素 (Naive) 贝叶斯结构嵌入的思想, 即在贝叶斯网络结构中用星形结构代替多父结点的汇聚结构, 使用该结构产生模拟数据来扩展小数据集, 但会导致局部和整体结构的不相容 (可能产生环路, 并且星形结构和汇聚结构是两种具有本质差别的结构), 产生的模拟数据一般不是来自于某一联合分布, 可能引入大量的噪声, 因此也很难收到好的效果. 这些研究只注重小数据集的扩展, 却忽略了另一个更重要的环节, 即对扩展部分数据的修正. Heckerman 首先给出了因果影响独立性 (Independence of causal influences, ICI) 假设[11], 基于这一假设, Onisko 和 Baumgartner 等[12−13] 分别建立了 noisy-or-gates 和 noisy-and-gates 等方法. 这些方法使多父结点的

收稿日期 2008-06-23 收修改稿日期 2008-12-31
Received June 23, 2008; in revised form December 31, 2008
国家自然科学基金 (60675036, 60803055), 上海市教委重点学科基金和上海市教委科研创新重点项目 (09zz202) 资助
Supported by National Natural Science Foundation of China (60675036, 60803055), Leading Academic Discipline Project of Shanghai Municipal Education Commission and Innovation Program of Shanghai Municipal Education Commission (09zz202)
1. 上海立信会计学院数学与信息学院 上海 201620 2. 上海立信会计学院开放经济与贸易研究中心 上海 201620 3. 南京大学商学院 南京 210093
1. School of Mathematics and Information, Shanghai Lixin University of Commerce, Shanghai 201620 2. Opening Economy and Trade Research Center, Shanghai Lixin University of Commerce, Shanghai 201620 3. School of Business, Nanjing University, Nanjing 210093
DOI: 10.3724/SP.J.1004.2009.01063

确定可分别进行, 不需要大量的例子数据, 在一些情况下能够取得较好的效果. 但忽略了多父结点之间的联合效应 (父结点的单独影响可能较弱, 联合却提供较强的信息), 而联合效应是多父结点之间所具有的一个重要特征, 因此会造成信息丢失, 从而降低可靠性.

针对上述情况, 本文提出一种有效实用的小数据集贝叶斯网络结构学习方法, 该方法主要包括四个部分: 1) 从小数据集中建立最大似然树 (Maximal likelihood tree)[14]; 2) 对小数据集进行扩展, 并将最大似然树与 Gibbs 抽样相结合修正扩展数据, 得到完整的数据集 (包含小数据集和修正后的扩展数据集); 3) 基于小数据集为最大似然树的边定向获得有向树, 并使用有向树和完全有向无环图法依次进行块间排序和块内结点排序; 4) 在完整数据集和结点顺序的基础上, 采用局部打分–搜索方法, 通过发现一个结点的父结点来进行贝叶斯网络结构学习. 最后通过实验验证方法的有效性和可靠性.

用 $X_1,\cdots,X_n$ 表示离散随机变量, 简称变量, $x_1,\cdots,x_n$ 为其值. 数据集 D 中具有 N 个记录 (或例子), 并假设数据是独立地随机产生于概率分布 P. 在概率模式中的变量以及表示概率模式的图形模式中的结点有时不加区分.

1 最大似然树学习

虽然不能直接从小数据集中可靠地进行贝叶斯网络结构学习, 但却往往能够比较可靠地学习最大似然树. 因为建立最大似然树只使用互信息由大到小排序后的顺序信息, 互信息计算也不需要大量的数据, 而顺序信息又是相对比较稳定的信息, 其参数最多只是一阶条件概率.

分别用 $I(X_i,X_j)$ 和 $I(X_i,X_j|X_k)$ 表示变量 X_i 和 X_j 之间的互信息和以 X_k 为条件的条件互信息. 用 L_0 表示边表, 其中的元素是三元组 (i,j,δ), $j>i$, $i=1,\cdots,n$, $j=2,\cdots,n$, $\delta=0,1$. 当 $\delta=1$ 和 $\delta=0$ 时分别表示边 X_i-X_j 存在与否.

下面给出经典的 Chow 和 Liu[14−15] 两步建树算法, 算法时间复杂度是 $\mathrm{O}(n^2)$: 1) 把互信息 $I(X_i,X_j|D)$ 作为边 X_i-X_j 的权重, 计算所有边的权重, 并按权重由大到小排序边; 2) 遵循不产生环路的原则, 依据边权重的大小, 由大到小依次选择边建树, 并修改边表, 直到选择 $n-1$ 条边为止, 得到最大似然树 T.

选择一个结点作为根结点, 由根结点向外的方向为边定向, 定向的目的是便于分解联合概率, 这种方向不具有因果语义. 也可以采用不需要定向的基于 Clique (无向图中的最大完全子图) 结合树[1] 的方法进行联合概率分解, 两种方法等价, 但定向的方法更为简单.

2 小数据集的扩展与扩展数据的修正

由于小数据集中所蕴涵的依赖识别信息不充分, 导致直接从小数据集中进行贝叶斯网络结构学习可靠性得不到保障. 通过对小数据集的扩展和对扩展数据的修正, 来提高学习的有效性和可靠性.

2.1 小数据集的扩展

使用基于数据集本身的 Bootstrap 抽样扩展小数据集 (不易引入噪声), 得到具有 N^* 个记录的扩展数据集 D^*. N^* 的大小与小数据集中例子数量, 变量的最大取值数量和变量之间的依赖复杂程度有关, 一般是依据能够比较可靠地进行充分统计因子计算 (基于打分–搜索方法) 或高阶条件概率估计 (基于依赖分析方法) 来大致确定 N^*. 在 D^* 中分布均匀地产生具有一定比例的位置 $E=\{l_1,l_2,\cdots,l_Q\}$ (一般占扩展数据的 30 %∼60 %), l_i 表示第 i 个待修正数据的位置, 并对这些位置上的数据进行随机初始化.

2.2 基于最大似然树的扩展数据修正

最大似然树是与贝叶斯网络具有最好拟合的树形结构[14, 16], 往往构成贝叶斯网络结构的主体骨架, 所确定的联合概率分布也是产生数据联合概率分布的良好近似. 基于最大似然树的修正将使扩展数据在记录内部发生实质性的变化, 充分统计因子估计得到改进. 将最大似然树与 Gibbs 抽样[17] 相结合对扩展数据进行修正, 其过程是一个迭代. 在迭代中, 依次修正 E 中位置上的数据, 每修正一个数据, 如果发生变化, 调整相关的参数, 修正完所有位置上的数据实现一次迭代, 直到满足终止条件结束迭代.

2.2.1 修正方法

设变量 X_i 和它的父结点 $\Pi(x_i)$ 可能的取值为 $x_i^1,\cdots,x_i^{r_i}$ 和 $\pi^1(x_i),\cdots,\pi^{q_i}(x_i)$, 用 $\theta_{ijk}=p(x_i^k|\pi^j(x_i),T)$ 表示最大似然树 T 的参数, $\sum_{k=1}^{r_i}\theta_{ijk}=1$. 记 $\boldsymbol{\theta}=\bigcup_{i=1}^{n}\{\boldsymbol{\theta}_i\}$, $\boldsymbol{\theta}_i=\bigcup_{j=1}^{q_i}\{\boldsymbol{\theta}_{ij}\}$, $\boldsymbol{\theta}_{ij}=\bigcup_{k=1}^{r_i}\{\boldsymbol{\theta}_{ijk}\}$, 那么, 由 $(T,\boldsymbol{\theta})$ 确定的联合概率分布为

$$p(x_1,\cdots,x_n|\boldsymbol{\theta},T)=\prod_{i=1}^{n}p(x_i|\pi(x_i),\boldsymbol{\theta}_i,T) \quad (1)$$

在第 h 次迭代中, 用 $D_f^{(h)}$ 和 $D_b^{(h)}$ 表示修正前后的数据集 (其中包含已知数据和扩展数据, 但只修正扩展数据), $\boldsymbol{\theta}_f^{(h)}$ 和 $\boldsymbol{\theta}_b^{(h)}$ 是对应的参数向量, 每修正一个数据便产生一个新的数据集, 这样形成一个数据集序列 $D_f^{(h)}=D_0^{(h)}$, $D_1^{(h)}$, $\cdots$, $D_Q^{(h)}$, $D_{Q+1}^{(h)}=D_b^{(h)}$, 对应地也产生一个参数向量序列 $\boldsymbol{\theta}_f^{(h)}=\boldsymbol{\theta}_0^{(h)}$,

$\boldsymbol{\theta}_1^{(h)}, \cdots, \boldsymbol{\theta}_Q^{(h)}, \boldsymbol{\theta}_{Q+1}^{(h)} = \boldsymbol{\theta}_b^{(h)}$. 下面给出在 $D_v^{(h)}$ $(0 \le v \le Q)$ 基础上通过修正 l_v 位置上的数据得到 $D_{v+1}^{(h)}$, 以及由 $\boldsymbol{\theta}_v^{(h)}$ 得到 $\boldsymbol{\theta}_{v+1}^{(h)}$ 的方法.

当 $p(x_i|\pi(x_i),\boldsymbol{\theta}_i,D_v^{(h)},T)=0$ 但 $p(\pi(x_i)|D_v^{(h)},T)\neq 0$ 时, 对 $p(x_i|\pi(x_i),\boldsymbol{\theta}_i,D_v^{(h)},T)$ 进行拉普拉斯修正 (Laplace-corrected), 令

$$p(x_i|\pi(x_i),D_v^{(h)},T)=\frac{\frac{1}{N+N^*}}{N'(\pi(x_i),D_v^{(h)})+N'(x_i,D_v^{(h)})(\frac{1}{N+N^*})}$$

其中, $N'(x_i,D_v^{(h)})$ 和 $N'(\pi(x_i),D_v^{(h)})$ 分别表示在 $D_v^{(h)}$ 中 $X_i=x_i$ 和 $\Pi(X_i)=\pi(x_i)$ 的例子数量. 当 $p(\pi(x_i)|D_v^{(h)},T)=0$ 时, $p(x_i|\pi(x_i),\boldsymbol{\theta}_i,D_v^{(h)},T)$ 采用均匀分布.

标准的 Gibbs sampling 采用满条件分布进行抽样[17], 其抽样复杂程度随变量增加指数增长, 基于最大似然树分解联合概率能够有效降低抽样的复杂性 (将 $n-1$ 阶条件概率转化为 n 个 1 阶条件概率的乘积). 根据贝叶斯公式可得

$$\begin{aligned}&p(x_i|x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n,D_v^{(h)},T)=\\&\frac{p(x_1,\cdots,x_n,D_v^{(h)},T)}{p(x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n,D_v^{(h)},T)}=\\&\alpha p(x_1,\cdots,x_n,D_v^{(h)},T)=\\&\alpha\prod_{i=1}^{n}p(x_i|\pi(x_i),D_v^{(h)},T)\end{aligned}\tag{2}$$

其中, α 是与 x_i 无关的量. 对其进行归一化处理, 记

$$\omega(k)=\frac{\prod_{i=1}^{n}p(x_i^k|\pi(x_i),D_v^{(h)},T)}{\sum_{k=1}^{r_i}\prod_{i=1}^{n}p(x_i^k|\pi(x_i),D_v^{(h)},T)}$$

对产生的随机数 λ, 变量 X_i 的修正值为

$$\hat{x}_i=\begin{cases}x_i^1, & 0<\lambda\le\omega(1)\\ x_i^k, & \sum_{j=1}^{k-1}\omega(j)<\lambda\le\sum_{j=1}^{k}\omega(j)\\ x_i^{r_i}, & \lambda>\sum_{j=1}^{r_i-1}\omega(j)\end{cases}\tag{3}$$

修正一个数据后, 调整与其有关的参数, 再修正下一个数据, 直到修正完所有待修正的数据. 修正一次数据集的时间复杂度是 O(Q).

2.2.2 参数更新

设位置 l_v 在第 i 列第 m 行, 此位置上的数据用 x_{im} 表示, 如果 $x_{im}\neq\hat{x}_{im}$ 需要调整对应的参数, 参数调整方法如下

$$\hat{p}^{(h)}(x_{im}|\pi(x_{im}),D_{v+1}^{(h)},T)=\hat{p}^{(h)}(x_{im}|\pi(x_{im}),D_v^{(h)},T)-\frac{1}{N+N^*}\tag{4}$$

$$\hat{p}^{(h)}(\hat{x}_{im}|\pi(x_{im}),D_{v+1}^{(h)},T)=\hat{p}^{(h)}(\hat{x}_{im}|\pi(x_{im}),D_v^{(h)},T)+\frac{1}{N+N^*}\tag{5}$$

2.3 迭代终止条件

设相邻两次迭代 (第 h 和 $h+1$ 次) 所得到的修正数据序列为 $x_{n_1}^{(h)},x_{n_2}^{(h)},\cdots,x_{nQ}^{(h)}$ 和 $x_{n_1}^{(h+1)},x_{n_2}^{(h+1)},\cdots,x_{nQ}^{(h+1)}$, 采用两个序列的一致性检验进行迭代终止判断. 设

$$\mathrm{sig}(x_{n_i}^{(h)},x_{n_i}^{(h+1)})=\begin{cases}0, & x_{n_i}^{(h)}=x_{n_i}^{(h+1)}\\ 1, & x_{n_i}^{(h)}\neq x_{n_i}^{(h+1)}\end{cases}$$

对给定的阈值 $\eta>0$, 如果 $\frac{1}{Q}\sum_{i=1}^{Q}\mathrm{sig}(x_{n_i}^{(h)},x_{n_i}^{(h+1)})\le\eta$, 则结束迭代.

3 贝叶斯网络结构学习

首先基于小数据集为最大似然树的边因果定向得到有向树, 然后使用有向树和局部完全有向无环图分别进行块间排序和块内结点排序, 得到所有结点的序列, 最后在结点顺序的基础上采用局部打分−搜索方法进行贝叶斯网络结构学习.

3.1 有向树学习

碰撞识别 (Collider identification)[1, 14] 是普遍采用的确定因果关系的方法, 具有操作简单、效率和可靠性高等特点, 但其只能发现部分 (具有多父结点) 边的方向. 在边表 L_0 中查询, 选择不存在边的结点对 X_i, X_j, 设在 T 中与 X_i 和 X_j 可能形成 V 结构[14] 的结点为 $X_{m_1},\cdots,X_{m_t}$, 对每一个可能的 V 结构进行碰撞识别. 对给定的阈值 $\delta>0$, 如果 $I(X_i,X_j|X_{m_h})/I(X_i,X_j)>(1+\delta)$, $1\le h\le t$, 则 X_i, X_j 和 X_{m_h} 形成 V 结构, 定向为 $X_i\to X_{m_h}$ 和 $X_j\to X_{m_h}$, 这样可为部分边定向. 根据链图 (Chain graph)[18] 定义 (在一个由有向边和无向边构成的混合图中, 如果对每一个回路中的无向边任意定向都不会产生有向环, 称这个混合图为链图), 所得到的是一个链图, 使用基于链图的打分方法为其他边定向. 设链图中没有定向的边为 $e_1,\cdots,e_s$, 用 G_C^{i+} 和 G_C^{i-} 分别表示边 $e_1,\cdots,e_{i-1}$ 已定向, 而边 e_{i+1},

$\cdots$, e_s 还没定向, 由边 e_i 具有不同方向所构成的链图. 按照 Buntine[18] 给出的基于链图分解联合概率的方法, 便能计算一个链图的 MDL (Minimal description length) 打分. 根据 $MDL(G_C^{i+}|D)$ 和 $MDL(G_C^{i-}|D)$ 的大小确定边 e_i 的方向, 如此下去可得到有向树, 用 T_D 表示学习得到的有向树.

3.2 结点排序

使用有向树 T_D 可对结点进行拓扑排序, 结点序列不唯一, 但结点块序列唯一. 首先基于有向树对结点块排序, 然后再采用局部完全有向无环图法对快内结点排序, 最终得到所有结点的序列.

3.2.1 结点块排序

在有向树 T_D 中, 选择没有父结点的结点构成第一个结点块, 然后从 T_D 中删除被选择的结点, 以及与这些结点相联接的弧, 再以同样方式选择和删除结点, 便可得到结点块序列, 可以证明, 在不考虑块内结点顺序的情况下块序列是唯一的.

3.2.2 块内结点排序

块内结点排序可依据专家知识、排序算法以及二者的结合. 下面给出一种块内结点排序的完全有向无环图法. 该方法分为两个步骤: 首先根据两个变量之间的条件相对平均熵确定块内所有结点之间弧的方向, 建立完全有向图, 可以证明所构建的有向图不存在环路; 然后依据完全有向无环图对块内结点进行拓扑排序 (排序结果唯一).

分别用 $H(X_i)$、$H(X_i|X_j)$ 和 $R(X_j \to X_i)$ 表示 X_i 的熵、条件熵和 X_j 对 X_i 的条件相对平均熵, $H(X_i) = -\sum_{x_i} p(x_i)\log p(x_i)$, $H(X_i|X_j) = -\sum_{x_i,x_j} p(x_i,x_j)\log p(x_i|x_j)$, $R(X_j \to X_i) = H(X_i|X_j)/(H(X_i)\times|X_i|)$. 对任意的 X_i, X_j, 如果 $R(X_j \to X_i) > R(X_i \to X_j)$, 则定向为 $X_j \to X_i$, 否则定向为 $X_i \to X_j$. 设最终得到的结点序列为 $X_{d_1}, \cdots, X_{d_n}$.

3.3 贝叶斯网络结构学习

在结点序列的基础上, 使用局部打分–搜索方法, 通过发现一个结点的父结点来进行贝叶斯网络结构学习.

3.3.1 打分函数

对于数据集 $D \cup D^*$ 和贝叶斯网络结构 G_B, 用 $MDL(G_B|D\cup D^*)$ 表示 G_B 的 MDL 打分值, 打分最小的网络结构便是最好的结构.

$$MDL(G_B|D\cup D^*) = \frac{\log N}{2}|G_B| - LL(G_B|D\cup D^*)$$

$$\frac{\log N}{2}|G_B| = \frac{\log N}{2}\sum_{i=1}^{n}|X_i||\Pi_i|$$

$$LL(G_B|D\cup D^*) = \sum_{m=1}^{N+N^*}\log(p(u_m|G_B)) = \sum_{m=1}^{N+N^*}\log\left(\prod_{i=1}^{n}p(x_{im}|\pi_{im},G_B)\right)$$

其中, $|G_B|$ 表示网络参数数量, $|X_i|$ 是 X_i 的取值数量, $|\Pi_i|$ 是 X_i 的父结点集 Π_i 的配置情况数量, u_m 表示数据集中的第 m 个纪录, x_{im} 是 X_i 在 u_m 中的取值, π_{im} 是 Π_{im} 的配置.

局部打分定理. 在两个贝叶斯网络结构 G_1 和 G_2 中, 设 X_i 在两个结构中的父结点集分别为 $\Pi_i^{(1)}$ 和 $\Pi_i^{(2)}$, 如果只有 X_i 的父结点集不同, 那么 $MDL(G_1|D\cup D^*) - MDL(G_2|D\cup D^*) = MDL(X_i\cup\Pi_i^{(1)}|D\cup D^*) - MDL(X_i\cup\Pi_i^{(2)}|D\cup D^*)$.

局部打分定理保证了可在结点有序的情况下进行局部打分, 相对于整体打分–搜索可有效提高结构学习效率和可靠性.

3.3.2 搜索算法

用 $G(X_{d_i}|\Pi_{d_i})$ 表示结点 X_{d_i} 和它的父结点所构成的局部结构. 如果 $MDL(G(X_{d_i}|X_{d_i}^{(1)})|D\cup D^*) = \min_{X_{d_k}\in\{X_{d_1},\cdots,X_{d_i-1}\}}\{MDL(G(X_{d_i}|X_{d_k})\mid D\cup D^*)\} < MDL(X_{d_i}|D\cup D^*)$, 选择 $X_{d_i}^{(1)}$ 作为 X_{d_i} 的第 1 个父结点, 否则结束 X_{d_i} 的父结点搜索, 该结点没有父结点. 假设已经选择了 w 个父结点 $X_{d_i}^{(1)}, \cdots, X_{d_i}^{(w)}$, 如果

$$MDL(G(X_{d_i}|X_{d_i}^{(1)},\cdots,X_{d_i}^{(w)},X_{d_i}^{(w+1)})|D\cup D^*) = \min_{X_{d_k}\in\{X_{d_1},\cdots,X_{d_i-1}\}-\{X_{d_i}^{(1)},\cdots,X_{d_i}^{(w)}\}}\{MDL(G(X_{d_i}|X_{d_i}^{(1)},\cdots,X_{d_i}^{(w)},X_{d_k})|D\cup D^*)\} < MDL(G(X_{d_i}|X_{d_i}^{(1)},\cdots,X_{d_i}^{(w)}|D\cup D^*)$$

选择 $X_{d_i}^{(w+1)}$ 作为 X_{d_i} 的第 $w+1$ 个父结点, 否则结束搜索. 这样便可依次确定每一个结点的父结点集, 最终得到贝叶斯网络结构. 算法最坏情况下 (学习得到的是每两个结点之间都存在一条弧的有向完全图) 的时间复杂度是 $\mathrm{O}(n^3)$. 由于确定每一个结点的父结点是独立进行的, 因此适合于设计并行算法来提高学习效率.

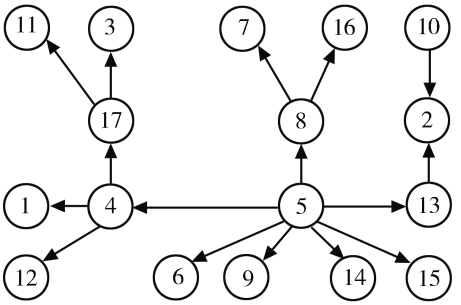
3.4 结构学习举例

从 UCI 机器学习数据仓库[19] 中选择数据集 Voting_records, 将数据集中的字段依次记为 X_1, $\cdots, X_{17}$, 排序结果使用下标数字简记, 使用上述方法进行贝叶斯网络结构学习.

从数据集 Voting_records 中学习得到的有向树如图 1 (a) 所示, 对有向树拓扑排序产生的结点块序

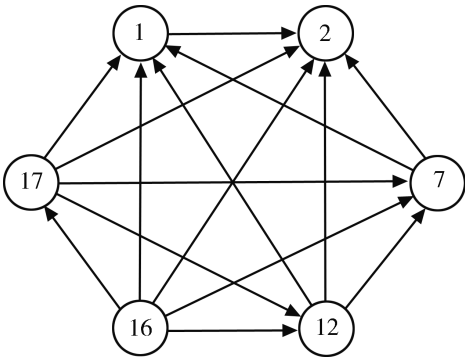
列为: $\{5, 10\}$, $\{4, 6, 8, 9, 13, 14, 15\}$, $\{1, 2, 7, 12, 16, 17\}$, $\{3, 11\}$.

第三个块内结点构成的完全有向无环图如图 1 (b) 所示. 对完全有向无环图进行拓扑排序而得到的块内结点序列为: 16, 17, 12, 7, 1, 2. 所有结点的序列为: 5, 10, 6, 14, 15, 4, 8, 13, 9, 16, 17, 12, 7, 1, 2, 11, 3. 基于这一序列, 使用局部打分–搜索方法得到的贝叶斯网络结构如图 1 (c) 所示.



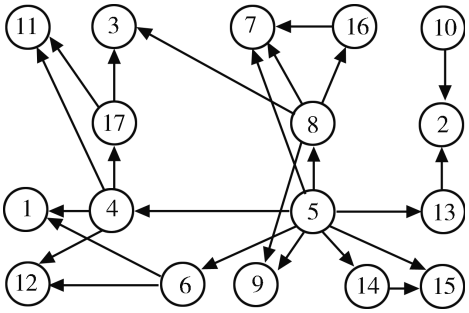
(a) 有向树

(a) Directed tree



(b) 完全有向无环图

(b) Perfect directed acyclic graph



(c) 贝叶斯网络结构

(c) Bayesian network structure

图 1　Voting_records 有向树、块内完全有向无环图和贝叶斯网络结构

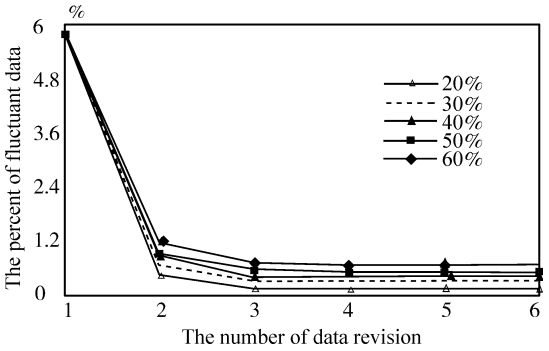
Fig. 1　The directed tree, perfect directed acyclic graph, and Bayesian network structure on Voting_records data set

4　实验与分析

根据网站 http://www.norsys.com 提供的 ALARM 网 (国际标准贝叶斯网络, 具有 37 个变量, 变量最多取值为 4, 最多父结点数为 4), Car-Diagnosis 网 (具有 18 个变量, 变量最多取值为 3, 最多父结点数为 4), 以及人工 Adata 网 (具有 10 个变量, 变量最多取值为 3, 最多父结点数为 5) 的概率分布表生成具有 5 000 个例子的模拟数据作为标准数据集, 分别从迭代收敛性、结点排序和贝叶斯网络结构学习效果比较三个方面进行实验与分析. 为表述方便, 对三个网络中的结点进行了编号, 并使用编号表示对应的结点.

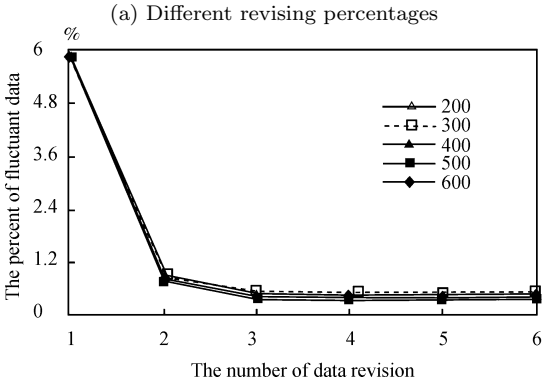
4.1　迭代收敛性

从 ALARM 网的模拟数据集中依次选择 200, 300, 400, 500 和 600 个记录建立五个小数据集, 从这五个数据集中学习得到的最大似然树结构完全一致. 利用 Bootstrap 抽样将五个小数据集扩展为具有 5 000 个例子的数据集, 首先选择具有 400 个例子的数据集, 修正数据比例分别为 20 %, 30 %, 40 %, 50 % 和 60 % 进行修正迭代收敛性实验, 然后选择五个数据集和 40 % 的修正比例再进行迭代收敛性实验, 取 $\eta = 1\,\%$, 情况如图 2 所示.



(a) 不同修正比例

(a) Different revising percentages



(b) 不同大小数据集

(b) Different size data sets

图 2　修正扩展数据的迭代情况

Fig. 2　The iteration of modifying extended data

从图 2 中可以看出, 无论是不同修正比例, 还是数据集的不同大小, 当修正迭代 3 次后均收敛, 显示了修正迭代具有较高的效率, 而且没有出现波动现象, 也表明了最大似然树与贝叶斯网络能够实现良好的拟合.

4.2 结点排序

分别使用标准的 ALARM 网和有向树对结点进行块排序, 然后再使用完全有向无环图法对块内结点排序, 并对两个得到的结点序列进行比较. 基于标准网的块序列: $\{12,16,17,18,19,20,21,22,23,24,28,30\}$, $\{3,4,10,25,26,31,37\}$, $\{1,2,36\}$, $\{13,35\}$, $\{15,34\}$, $\{32,33\}$, $\{11,14\}$, $\{27\}$, $\{29\}$, $\{6,7,8,9\}$, $\{5\}$. 基于有向树的块序列: $\{12,16,17,18,19,20,21,22,23,24,33,28,30\}$, $\{3,4,10,25,26,31,37\}$, $\{1,2,36\}$, $\{13,35\}$, $\{15,34\}$, $\{11,32\}$, $\{14\}$, $\{27\}$, $\{29\}$, $\{6,7,8,9\}$, $\{5\}$. 两个块序列之间存在一定的差异, 这些差异是由于结点 33 的位置变化所导致. 得到的最终两个结点序列非常接近, 只有结点 33 的位置不同, 这会影响结点 33 的父结点学习, 但对其他结点的父结点学习不会产生太大的影响.

4.3 学习效果比较

使用具有 400 个例子的 ALARM 网、Car-Diagnosis 网和人工 Adata 网的模拟数据集, 选择全部具有 3 个以上父结点的结点、部分具有 2 个父结点的结点. 分别从小数据集、扩展数据集和标准数据集中进行局部父结点学习, 具体情况如表 1 所示, 其中第 1 列中的数字表示局部学习的结点, 后 4 列中的数字是不同情况学习得到的和标准网中的对应父结点集, 括号中 "+" 和 "−" 分别表示相对于标准网增加和减少父结点.

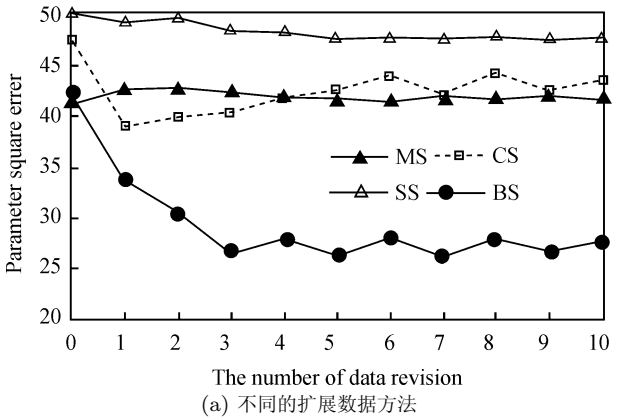
表 1 多父结点局部学习效果比较

Table 1 Comparison of local learning effects on multi-father nodes

局部学习结点	小数据集 (400)	扩展数据集 (5 000)	标准数据集 (5 000)	标准贝叶斯网络
ALARM	$(-12,+0)$	$(-4,+0)$	$(-2,+0)$	$(-0,+0)$
13	$36\,(-2,+0)$	$22,36\,(-1,+0)$	$22,23,36\,(-0,+0)$	$22,23,36$
14	$33\,(-1,+0)$	$33,35\,(-0,+0)$	$33,35\,(-0,+0)$	$33,35$
15	$35\,(-2,+0)$	$35\,(-1,+0)$	$35\,(-1,+0)$	$22,35$
27	$11,20\,(-2,+0)$	$4,11,20\,(-1,+0)$	$4,11,20\,(-1,+0)$	$4,11,20,33$
31	$22\,(-1,+0)$	$22,21\,(0,+0)$	$22,21\,(-0,+0)$	$22,21$
32	$34\,(-1,+0)$	$12,34\,(-0,+0)$	$12,34\,(-0,+0)$	$12,34$
34	$35\,(-1,+0)$	$22,35\,(-0,+0)$	$22,35\,(-0,+0)$	$22,35$
35	$36\,(-2,+0)$	$23,36\,(-1,+0)$	$22,23,36\,(-0,+0)$	$22,23,36$
Car-Diagnosis	$(-10,+0)$	$(-4,+0)$	$(-4,+0)$	$(-0,+0)$
7	$4\,(-2,+0)$	$4,5\,(-1,+0)$	$4,6\,(-1,+0)$	$5,6$
4	$2\,(-1,+0)$	$2,3\,(-0,+0)$	$2,3\,(-0,+0)$	$2,3$
9	$4\,(-2,+0)$	$4,5\,(-1,+0)$	$4,5\,(-1,+0)$	$4,5,8$
18	$12,13\,(-3,+0)$	$12,13,15\,(-2,+0)$	$12,13,15\,(-2,+0)$	$12,13,14,15,17$
Adata	$(-11,+0)$	$(-4,+0)$	$(-2,+0)$	$(-0,+0)$
6	$2\,(-4,+0)$	$2,3,5\,(-2,+0)$	$2,3,4,5\,(-1,+0)$	$1,2,3,4,5$
8	$7\,(-2,+0)$	$4,6,7\,(-0,+0)$	$4,6,7\,(-0,+0)$	$4,6,7$
9	$7\,(-2,+0)$	$6,7\,(-1,+0)$	$5,6,7\,(-0,+0)$	$5,6,7$
10	$8\,(-3,+0)$	$5,6,8\,(-1,+0)$	$5,6,8\,(-1,+0)$	$5,6,7,8$

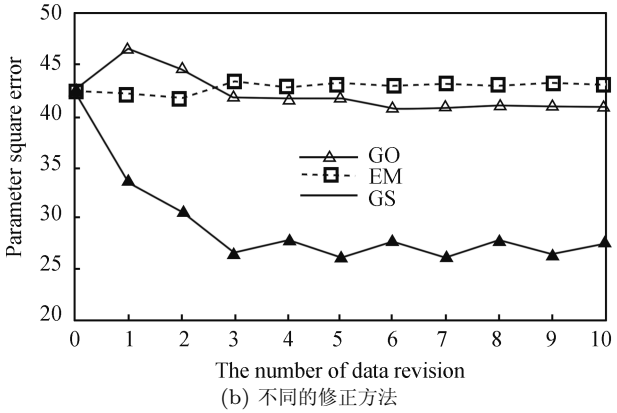
从表 1 中可以看到, 通过对小数据集的扩展以及对扩展数据的修正, 在使用具有扩展数据的数据集进行局部父结点学习时, 相对于小数据集丢失父结点的数量显著减少, 与标准数据集情况非常接近. 这表明了对小数据集的扩展以及对扩展数据的修正能够显著提高贝叶斯网络结构学习的有效性和可靠性.

分别采用最大似然树模拟、因果独立性模拟、星形结构嵌入模拟和 Bootstrap 抽样 (分别简记为 MS、CS、SS 和 BS) 来扩展 ALARM 网具有 400 个例子的数据集, 并对扩展数据进行修正, 修正位置是扩展数据的 40 %, 用于打分函数计算的充分统计因子平方误差变化情况如图 3 (a) 所示, 分别基于梯度优化、EM 算法和 Gibbs 抽样 (分别简记为 GO、EM 和 GS) 修正扩展数据的比较见图 3 (b).



(a) 不同的扩展数据方法

(a) Different extending methods



(b) 不同的修正方法

(b) Different revising methods

图 3　修正扩展数据情况比较

Fig. 3　Comparison of different situations on revising extended data

图 3 (a) 显示, 使用 Bootstrap 抽样扩展的数据集, 在修正迭代过程中充分统计因子误差显著下降, 并很快趋于稳定, 而采用其他方法得到的扩展数据集, 在修正迭代过程中充分统计因子误差没有明显的改进, 可见扩展数据的方法对修正效果影响较大. 从图 3 (b) 中可以看出, 基于梯度的优化方法和 EM 算法均不能达到修正扩展数据的目的, 其主要原因是局部最优性使得修正后的数据趋于极端化, 因此, 充分统计因子估计得不到良好的改进.

5　结论

本文在小数据集扩展的基础上引入了对扩展数据的修正机制, 并给出了使用小数据集进行结点排序以及基于结点序列的局部打分–搜索算法, 从而实现了比较可靠的小数据集贝叶斯网络结构学习, 为贝叶斯网络广泛用于解决实际问题又提供了一种有效适用的方法, 同时也为普遍存在的一般小数据集问题开拓了新的思路和可借鉴的方法.

References

1 Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. California: Morgan Kaufmann, 1988. 383−408

2 Heckerman D, Geiger D, Chickering D M. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 1995, **20**(3): 197−243

3 Cheng J, Greiner R, Kelly J, Bell D, Liu W R. Learning Bayesian networks from data: an information theory based on approach. *Artificial Intelligence*, 2002, **137**(1-2): 43−90

4 Zgurovskii M Z, Bidyuk P I, Terent′ev A N. Methods of constructing Bayesian networks based on scoring functions. *Cybernetics and Systems Analysis*, 2008, **44**(2): 219−224

5 de Campos L M, Castellano J G. Bayesian network learning algorithms using structural restrictions. *International Journal of Approximate Reasoning*, 2007, **45**(2): 233−254

6 Martínez-Rodríguez A M, May J H, Vargas L G. An optimization-based approach for the design of Bayesian networks. *Mathematical and Computer Modelling*, 2008, **48**(7-8): 1265−1278

7 Fan Min, Huang Xi-Yue, Shi Wei-Ren, Xian Xiao-Dong. Improved Bayesian networks structure learning algorithm. *Journal of System Simulation*, 2008, **20**(17): 4613−4617
(范敏, 黄席樾, 石为人, 鲜晓东. 一种改进的贝叶斯网络结构学习算法. 系统仿真学报, 2008, **20**(17): 4613−4617)

8 Friedman N, Goldszmidt M, Wyner A. Data analysis with Bayesian networks: a bootstrap approach. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. Stockholm, Sweden: Morgan Kaufmann, 1999. 196−205

9 Borchani H, Amor N B, Khalfallah F. Learning and evaluating Bayesian network equivalence classes from incomplete data. *International Journal of Pattern Recognition and Artificial Intelligence*, 2008, **22**(2): 253−278

10 Heckerman D, Meek C. Embedded Bayesian Network Classifiers, Technical Report MSR-TR-97-06, Microsoft Research, USA, 1997

11 Heckerman D, Breese J S. Causal independence for probability assessment and inference using Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 1996, **26**(6): 826−831

12 Onisko A, Druzdzel M J, Wasyluk H. Learning Bayesian network parameters from small data sets: an application of Noisy-OR gates. *International Journal of Approximate Reasoning*, 2001, **27**(2): 165−182

13 Baumgartner K, Ferrari S, Palermo G. Constructing Bayesian networks for criminal profiling from limited data. *Knowledge-based Systems*, 2008, **21**(7): 563−572

14 Wang Shuang-Cheng, Yuan Sen-Miao. Learning decomposable Markov network structure with missing data. *Chinese Journal of Computers*, 2004, **27**(9): 1221−1228
(王双成, 苑森淼. 具有丢失数据的可分解马尔科夫网络结构学习. 计算机学报, 2004, **27**(9): 1221−1228)

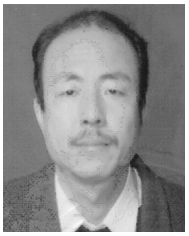
15 Chow C K, Liu C N. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 1968, **14**(3): 462−467

16 Perrier E, Imoto S, Miyano S. Finding optimal Bayesian network given a super-structure. *Journal of Machine Learning Research*, 2008, **9**(10): 2251−2286

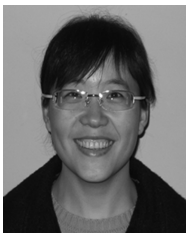
17 Geman S, Geman D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984, **6**(6): 721−741

18 Buntine W L. Chain graphs for learning. In: Proceedings of the 11th Conference on Uncentainty in Artificcial Intelligence. Montreal, Canada: Morgan Kaufmann, 1995. 46−54

19 Murphy P M, Aha D W. UCI repository of machine learning databases [Online], available: http: //www.ics.uci.edu/mlearn/MLRepository. Html, October 20, 2007



王双成 上海立信会计学院数学与信息学院教授. 主要研究方向为人工智能, 机器学习, 数据采掘及在经济领域中的应用. 本文通信作者.
E-mail: wangsc@lixin.edu.cn
(**WANG Shuang-Cheng** Professor at the School of Mathematics and Information, Shanghai Lixin University of Commerce. His research interest covers artificial intelligence, machine learning, and data mining and their application in economic domain. Corresponding author of this paper.)



冷翠平 上海立信会计学院数学与信息学院讲师. 主要研究方向为智能控制和数据采掘.
E-mail: aleng2001@sina.com
(**LENG Cui-Ping** Lecturer at the School of Mathematics and Information, Shanghai Lixin University of Commerce. Her research interest covers intelligence control and data mining.)

李小琳 南京大学商学院讲师. 主要研究方向为机器学习和数据采掘.
E-mail: lixl_126@126.com
(**LI Xiao-Lin** Lecturer at the School of Business, Nanjing University. Her research interest covers machine learning and data mining.)