文章编号: 1001-0920 (2012) 10-0000-00

基于最大主子图分解的贝叶斯网络等价类学习算法

朱明敏[a], 刘三阳[a,b], 杨有龙[a]

(西安电子科技大学 a. 理学院, b. 综合业务网国家重点实验室, 西安 710071)

摘 要: 针对基于约束方法学习贝叶斯网络 (BN) 结构的不足, 以及随着条件集的增大, 利用统计方法进行条件独立 (CI) 测试不稳定等问题, 提出一种基于最大主子图分解 (MPD) 的 BN 等价类学习算法. 该算法首先通过 MPD 分解技术对 BN 的道德图进行分解; 然后利用 0 阶和 1 阶 CI 测试识别部分子图中的 V 结构, 对于初步未定的 V 结构利用局部评分搜索确定, 从而避免了冗余检验, 有效地减小了条件集的维数, 并且提高了算法的效率; 最后, 理论证明以及实验结果表明了所提出算法的有效性和合理性.

关键词: 贝叶斯网络; 最大主子图分解; 条件独立测试; 结构学习; 马尔科夫等价类

中图分类号: TP18 **文献标识码:** A

Structural learning Bayesian network equivalence classes via maximal prime decomposition

ZHU Ming-min[a], *LIU San-yang*[a,b], *YANG You-long*[a]

(a. School of Science, b. State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China. Correspondent: ZHU Ming-min, E-mail: zmmzhu2010@126.com)

Abstract: To solve the drawbacks of constraint-based method for learning Bayesian networks(BN) and the unreliability of the conditional independence(CI) tests as the conditioning sets become too large, this paper proposes a structural learning algorithm based on maximal prime decomposition(MPD). Firstly, MPD technique is used to transform the moral graph of BN into its sub-graphs. Then, only zero order and first order CI tests are used to identify V-structures in part of sub-graphs and takes scoring function searches to optimize local structure, so that the number of conditional independence tests can be decreased. Redundancy tests can be avoided and the time performance can be greatly enhanced. Finally, theoretical and experimental results show that the new algorithm is effective and reasonable.

Key words: Bayesian network; maximal prime decomposition; conditional independence test; structure learning; Markov equivalence class

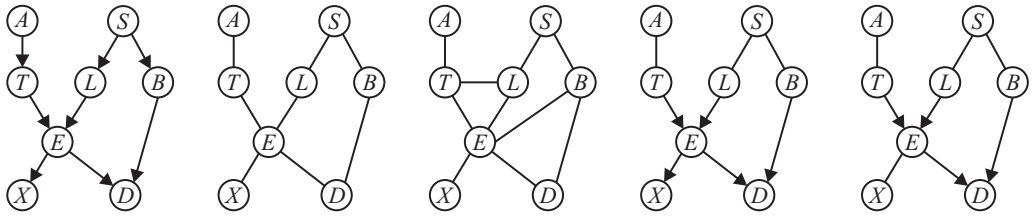
1 引 言

贝叶斯网络 (BN)[1-2], 又称贝叶斯信度网络 (BBN) 或信度网络, 是图论与概率论结合的产物. 经过十几年的发展, 它已广泛地应用于故障检测、医疗诊断、交通管理、军事目标自动识别、数据挖掘、作战意图自动估计, 以及信息融合等方面[3-5]. 近年来, 从数据中学习BN受到了国内外学者们的广泛关注, 人们相继提出了许多种BN结构学习算法[6-9]. 这些算法大致可归为两类: 基于随机搜索机制的评分搜索方法和基于约束的方法. 基于约束的方法将该问题看作约束满足问题, 一般需要两个步骤才能得到所求的部分有向图结构(即贝叶斯网络等价类). 1) 根据所观察到的数据或领域知识通过条件独立 (CI) 测试建立道德图[10-11]; 2) 确定道德图中的 V 结构和不可逆有向边. 为了提高 2) 的计算效率, 本文在深入分析道德图的结构和性质的基础上, 给出了一种基于最大主子图分解的等价类学习算法 (EC-MPD). 该算法不仅证明了确定 BN 的 V 结构等价于确定其所有最大主子图的 V 结构, 并且利用这一性质来压缩搜索空间, 从而将一个高维的网络结构学习问题简化为低维问题. 同时该算法仅执行 0 阶和 1 阶条件独立测试进行初步弧定向, 对于初步未定的 V 结构利用 BDeu 函数[12]进行

收稿日期: 2011-04-07; **修回日期:** 2011-05-20.

基金项目: 国家自然科学基金项目 (60974082, 61075055); 国家杰出青年科学基金项目 (11001214); 西安电子科技大学基本科研业务基金项目 (K50510700004).

作者简介: 朱明敏(1985−), 女, 博士生, 从事机器学习与贝叶斯网络优化的研究; 刘三阳(1959−), 男, 教授, 博士生导师, 从事最优化、网络算法等研究.



(a) ASIA网络 (b) ASIA网框架 (c) 道德图 (d) CPDAG (e) PDAG

图 1 ASIA 网络的各种变形

局部评分搜索确定, 避免了冗余检验, 从而大大减少了条件独立测试的计算量. 理论和实验证明了算法的有效性和合理性.

2 贝叶斯网络及其等价类

在论述网络中的节点或概率分布中的随机变量时, 大写字母U, V, W表示节点集合或变量集合; 带下标的字母, 如V_i表示单个节点或变量; $G[U]$表示由变量集U导出的无向子图.

2.1 贝叶斯网络

定义 1 BN是一个二元组, 即$\mathrm{BN}=(G,P)$. 其中: $G=(V,E)$为有向无环图(DAG); P为节点的概率分布, 表示节点之间因果影响强度. 这里: V为节点集, 与领域的随机变量一一对应; E为有向边集, 反映节点变量之间的因果依赖关系.

给定BN, 存在一个离散变量集合$X=\{V_1,V_2,\cdots,V_n\}$上的联合概率分布, 即

$$P(V)=P(V_1,V_2,\cdots,V_n)=\prod_{i=1}^{n}P(V_i|\mathrm{Pa}(V_i)). \quad (1)$$

其中: $\mathrm{Pa}(V_i)$为变量V_i在BN中的父节点集, 展开式意味着对任意变量V_i在给定其父集合$\mathrm{Pa}(V_i)$情况下, V_i独立于除$\mathrm{Pa}(V_i)$之外的所有非子孙节点. 本文用$\mathrm{Ind}(X,Y|Z)$表示给定变量集Z, 变量集X和Y在概率分布P下是相互独立, $X,Y,Z\subset V$.

注 1 若BN中存在有向边$V_i\to V_j$, 则称V_i是V_j的父节点, V_j是V_i的子节点. 若存在有向路$V_i\to\cdots\to V_j$, 则称V_i是V_j的祖先节点, V_j是V_i的子孙节点.

定义 2 有向图$G=(V,E)$的一条链是指一个有限非空序列$\rho=V_1e_1V_2\cdots e_lV_{l+1}$, 其各项交替的是节点和有向边, 其中$e_i$表示有向边$V_i\to V_{i+1}$或$V_i\leftarrow V_{i+1}$, $i=2,3,\cdots,l$. 任意变量子集$Z\subset V$, Z d-分离ρ当且仅当下面两个条件之一成立: 1) ρ包含形如$V_i\to V_j\to V_k$或$V_i\leftarrow V_j\to V_k$的子链, 并且$V_j\in Z$; 2) ρ包含形如$V_i\to V_j\leftarrow V_k$的子链, 并且V_j和V_j的子孙节点都不属于Z.

特别地, G中任意两个不同的子集X, Y被集合Z d-分离当且仅当任意一条经过$X_i\in X$到达$Y_i\in Y$的链被Z d-分离.

对于任意$\mathrm{BN}=(G,P)$, 概率分布P上所有变量之间的条件独立都可通过有向图G上的d-分离表示出来, 则称BN满足忠实性条件. 假设下文所述的BN均满足该条件.

给定有向图$G=(V,E)$, 任意V_i, V_j, $V_k\in V$, 如果V_i, V_j在G中不相邻, 且有向边$V_i\to V_k$, $V_j\to V_k\in E$, 则称$V_i\to V_k\leftarrow V_j$是G中的V结构. 将有向图G的有向边变为无向边, 得到的无向图称为图G的框架. 有向图G的道德图是通过将G中的有向边转变为无向边并在具有共同子节点的节点之间添加无向边得到的, 这些被添加的边称为道德边. 以图1(a)中ASIA网络为例, 图1(b)表示ASIA网络的框架, 图1(c)表示ASIA网络的道德图, 其中边T-L和E-B是道德边.

2.2 马尔科夫等价

BN结构学习就是在给定一个样本数据集合D的前提下, 寻找一个与训练集D匹配最好的网络结构, 该结构反映了节点变量间潜在的条件独立关系. 如果任意两个不同的BNs定义了同一联合概率分布, 则称这两个BNs是马尔科夫等价的. 对于大部分评分函数, 例如: AIC(Akaike information criterion)[13], BIC(Bayesian information criterion)[14]和BDe(Bayesian dirichlet equivalent)[12]等, 等价的BNs具有相同的函数得分[15], 因此, 对BN结构的学习可简化为学习其等价类. 两个BNs之间是否马尔科夫等价可通过以下方法进行判断[16].

定理 1 贝叶斯网络$\mathrm{BN}_1=(G_1,P_1)$和$\mathrm{BN}_2=(G_2,P_2)$马尔科夫等价, 当且仅当G_1和G_2具有相同的框架和V结构.

定理1表明, 与有向图G等价的图结构是不唯一的并且不一定都是有向图. 一般地, 一个有向图G的马尔科夫等价类是由一系列结构等价的部分有向图组成的.

定义 3 设$P=(V,E_p)$是一个图, 若边集E_p中包含有向边和无向边, 则称P是一个部分有向图. 若部分有向图P中不存在有向圈, 则称P是一个部分有向无环图(PDAG).

给定部分有向无环图 $P=(V,E_p)$, 任意有向边 $V_i \to V_j \in E_p$, 若存在图 $P'=(V,E_{p'})$ 与 P 等价, 且 $V_j \to V_i \in E_{p'}$, 则称有向边 $V_i \to V_j$ 在 P 中是可逆的, 否则是不可逆的. 同理, 对任意无向边 $V_i - V_j \in E_p$, 若存在 $P_1=(V,E_{p_1})$ 和 $P_2=(V,E_{p_2})$ 均与 P 等价, 且 $V_i \to V_j \in E_{p_1}$, $V_j \to V_i \in E_{p_2}$, 则称无向边 $V_i - V_j$ 在 P 中是可逆的, 否则是不可逆的.

定义 4　设 $P=(V,E_p)$ 是一个部分有向无环图, 若 E_p 中的有向边都是不可逆的, 并且 E_p 中的无向边都是可逆的, 则称 P 是一个完全部分有向无环图 (CPDAG).

任意 BN 的马尔科夫等价类都存在唯一的完全部分有向无环图与之等价[17]. 因此, CPDAG 可作为贝叶斯网络等价类的图形化表示, 根据 CPDAG 的定义可知, 对于任意弧 $V_i - V_j$, 若在 CPDAG 中有向, 则在所有与之等价的有向图中也是有向的, 且方向与 CPDAG 中相同; 若弧在 CPDAG 中无向, 则在所有与之等价的有向图中会出现不同方向. 图 1(d) 和图 1(e) 是与 ASIA 网络等价的部分有向图. 图 1(d) 中所有有向边都是不可逆的并且所有无向边都是可逆的, 因此图 1(d) 是完全的; 而图 1(e) 中的无向边 E-X 是不可逆的, 因此图 1(e) 是不完全的部分有向图.

3　基于最大主子图分解的结构学习算法

如上文所述, BN 是随机变量概率依赖关系的图形化表示, 因此 BN 结构可通过对样本数据进行有效的 CI 测试, 并判断所有变量间的连接关系来确定, 这也恰是基于约束方法学习 BN 结构的基本思想. 一般地, 该类方法分为两个阶段实现. 第 1 阶段, 根据所观察到的数据利用条件独立测试或领域知识建立 BN 的道德图 (最小无向独立图); 第 2 阶段, 根据道德图确定 BN 的 V 结构和其余的不可逆有向边, 并删除道德边, 得到一个 CPDAG, 即为所求. 针对第 2 阶段的学习, 本文首先利用图论中的最大主子图分解技术对道德图进行分解; 然后, 通过初步识别每个子图中的 V 结构来确定 BN 中的不可逆有向边, 因为 V 结构只包含在形如 $X-Y-Z-X$ 结构 (简称为三元组 $\mathrm{Tr}(X,Y,Z)$) 的无向子图中, 所以这一过程不需要对所有子图进行测试, 只需测试那些包含在三元组 $\mathrm{Tr}(X,Y,Z)$ 中的边, 从而极大地减少了冗余弧和变量引起的统计因子的计算, 提高了算法的效率; 最后, 利用 BDeu 评分函数对初步未定的 V 结构进行局部搜索识别, 并确定不在 V 结构中的其余不可逆有向边. 由于第 1 阶段学习道德图结构不是本文讨论的重点, 在此不再赘述, 详见文献 [10-11]. 下面给出与算法相关的基本概念, 并给出算法的具体描述及其理论证明.

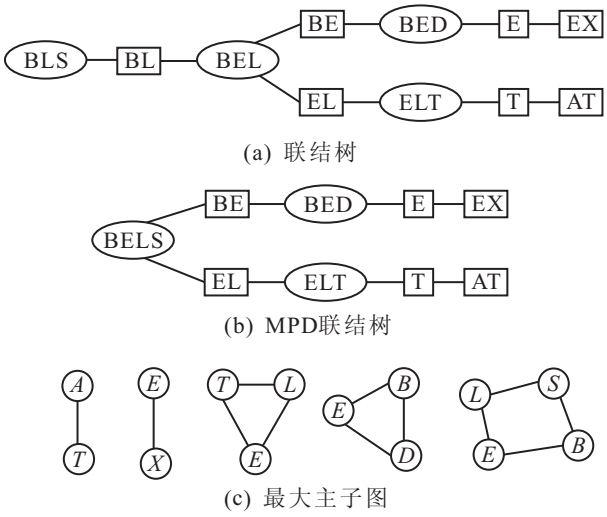


图 2　构建 ASIA 网络最大主子图的过程

3.1　最大主子图分解

定义 5　设 $G=(V,E)$ 是一个图, $V' \subseteq V$. 以 V' 为节点集, 以 G 中两端点均属于 V' 的所有边作为边集所构成的子图, 称为 G 的由节点集 V' 导出的子图, 记为 $G[V']$.

定义 6　设 $G=(V,E)$ 是一个无向图, 若存在 V 的非空子集 V', S 和 V'', 且 $V' \bigcup S \bigcup V'' = V$, 使 G 中每一条经过点 $Y' \in V'$ 到达点 $Y'' \in V''$ 的无向路上至少存在一个点在 S 中, 则称 S 分离子集 V' 和 V''; 如果 $G[S]$ 在 G 中是完全的, 则称 G 是可分解的, 且分解为子图 $G[V' \bigcup S]$ 和 $G[V'' \bigcup S]$.

若 $G[U]$ 是 $G=(V,E)$ 的一个导出子图且 $G[U]$ 是不可分解的, 则称 $G[U]$ 是 G 的主子图. 若对任意子集 $W \supset U$, $G[W]$ 都是可分解的, 则称 $G[U]$ 是 G 的最大主子图.

定义 7　BN 的联结树 (JT) 是一个二元组, 即 $\mathrm{JT}=(C,S)$. 任意 $C_i \in C$, $G[C_i]$ 为 G^m 三角化过程中产生的最大完全子图, 在 JT 上 C 中的元素满足联结树特性, 即给定联结树上的任意两个簇节点 C_i 和 C_j, 在 C_i 和 C_j 之间的路径上所有簇节点包含 $C_i \bigcap C_j$; S 为 JT 中的边集, 任意相邻两个簇节点 C_i 和 C_j 之间的边 $S_{ij}=C_i \bigcap C_j$.

对一个无向图进行最大主子图分解, 实质上是寻找这个图的所有最大主子图, 一个图的最大主子图分解是唯一的[18]. BN 上的最大主子图是指 BN 的道德图 G^m 的所有最大主子图. 本文利用联结树算法求解 G^m 的所有最大主子图[18]. 以 ASIA 网络为例, 由 ASIA 网的道德图 G^m 得到如图 2(a) 所示的联结树; 考察任意两个簇节点之间的边, 由于边 $\{B,L\}$ 导出的无向子图在图 1(c) 中是不完全的, 合并簇节点 $\{B,L,S\}$ 和 $\{B,E,L\}$, 得到如图 2(b) 所示的 MPD 联结树; 由图 2(b) 中所有簇节点导出的无

向子图即为ASIA网的最大主子图, 如图2(c)所示.

3.2 结构学习算法及其理论证明

在给出本文的主要算法之前, 首先给出两个定理, 作为算法的理论依据.

定理 2 设$G^m = (V, E^m)$表示有向图$G = (V, E)$的道德图, 任意V_i, $V_j \in V$, 如果存在变量子集S在G中d-分离V_i和V_j, 当且仅当S在G^m中分离V_i和V_j.

定理2的证明可参阅文献[19]. 定理2表明, 有向图中变量间的d-分离关系等价于其对应道德图中变量间的分离关系.

定理 3 设$G^m = (V, E^m)$表示$G = (V, E)$的道德图, 假设G^m可分解成它的最大主子图${G_1}^m$, ${G_2}^m, \cdots, {G_k}^m$, 对于任意V_i, V_k, $V_j \in V$, 如果$V_i \to V_k \leftarrow V_j$是$G$中的V-结构, 则至少存在$G^m$的一个最大主子图包含三元组$\mathrm{Tr}(V_i, V_j, V_k)$.

证明 根据G^m的连通性, 一定存在G^m的子图包含边$V_i - V_k$或$V_j - V_k$. 下面用反证法证明. 假设边$V_i - V_k$和$V_j - V_k$分别属于G^m的两个不同子图${G_p}^m$和${G_q}^m$, 由最大主子图分解的定义可知, 一定存在子集$S \subset V$且$V_k \in S$使得S分离${G_p}^m$和${G_q}^m$, 从而V_k在G^m中分离V_i和V_j, 根据定理2, V_k在G中d-分离V_i和V_j, 这与$V_i \to V_k \leftarrow V_j$是$G$中的V-结构矛盾. □

定理3表明, 确定BN中的所有V结构等价于将其道德图进行最大主子图分解, 对包含三元组的子图进行CI测试. 因此, 定理3将一个高维的结构学习问题转化为低维的CI测试问题, 基于这一思想, 下面给出本文的主要算法.

3.3 CI测试以及局部结构搜索

与其他基于约束的算法一样, EC-MPD算法需要运用少量的0阶和1阶CI测试确定BN中的V结构. 本文采用χ^2统计量进行CI测试. 给定H_0假设: 在给定数据集D条件下, $\mathrm{Ind}(V_i, V_j|V_k)$成立. 设${N_{ijk}}^{abc}$表示给定样本数据集中$V_i = a$, $V_j = b$, $V_k = c$的样本数目. 变量V_i, V_j和V_k在H_0假设下的充分统计量为

$$G^2 = 2\sum_{a,b,c} {N_{ijk}}^{abc} \log \frac{{N_{ijk}}^{abc} {N_k}^c}{{N_{ik}}^{ac} {N_{jk}}^{bc}}. \tag{2}$$

定理证明[20], 如果在给定样本数据集条件下, H_0假设成立, 则统计量G^2近似服从自由度为$(r_i - 1)(r_j - 1)r_k$的χ^2分布, 其中r_i表示变量V_i的取值个数. 因此, 可以采用χ^2检验进行BN弧的定向, χ^2检验的置信度设为99.5%.

算法 1 基于最大主子图分解的等价类学习算法(EC-MPD)如下:

Input: Data set D; Moral graph $G^m = (V, E^m)$.

Initialization: $E^c = E^m$, $T = \varnothing$;

Construct JT $= (C, S)$from G^m;

for $S_{ij} \in S$ do

  if $G(S_{ij})$ is not complete then

    $C = C \cup \{C_i \cup C_j\} \setminus \{C_i, C_j\}$, $S = S \setminus \{S_{ij}\}$;

  end if

end for

Set $G_l^m = G[C_l], l = 1, 2, \cdots, |C|$;

repeat

  for $\mathrm{Tr}(V_i, V_k, V_j) \subset G_l^m, l = 1, 2, \cdots, |C|$ do

    if $\mathrm{Ind} = (V_i, V_j)$ then

      $E^c = E^c \cup \{V_i \to V_k, V_j \to V_k\} \setminus \{V_i - V_j\}$;

      else if $\exists t \neq k, \mathrm{Ind}(V_i, V_j|V_t)$ then

      $E^c = E^c \cup \{V_i \to V_k, V_j \to V_k\} \setminus \{V_i - V_j\}$;

      else

        $T = T \cup \mathrm{Tr}(V_i, V_k, V_j)$;

    else if

  end for

until E^c has not changed;

Orient the undirected edges in E^c by using local BDeu score search method if they appear in the list T;

Orient other edges in E^c if each opposite of them creates either a directed cycle or a new V structure;

return $P^c = (V, E^c)$;

Output: CPDAG $P^c = (V, E^c)$.

EC-MPD算法首先利用联结树算法对道德图G^m进行最大主子图分解, 并检测分解后的每个子图是否包含三元组; 然后利用0阶和1阶CI测试初步确定BN中的V结构, 具体包括以下两个准则: 1) 若G^m的任意子图包含三元组$\mathrm{Tr}(V_i, V_k, V_j)$且$\mathrm{Ind}(V_i, V_j)$, 则有$V_i \to V_k \leftarrow V_j$; 2) 若$G^m$的任意子图包含三元组$\mathrm{Tr}(V_i, V_k, V_j)$且存在$t \neq k$使得$\mathrm{Ind}(V_i, V_j|V_t)$, 则有$V_i \to V_k \leftarrow V_j$. 容易证明, 在CI测试可靠的条件下, 由准则1)和2)得到的局部结构是BN的V结构.

经过CI测试阶段若仍存在包含无向边的三元组, 即集合T非空, 则需要运用评分搜索方法确定T中的三元组$\mathrm{Tr}(V_i, V_j, V_k)$是否包含V结构. 本文采用BDeu函数作为评分标准, BDeu函数是基于Bayesian Dirichlet先验分布的贝叶斯评分函数的一种特殊形式, 如果取先验等价样本量为10, 结构的先验信息为0.001^σ, σ表示网络中自由参数的数目, 则BDeu函数可表示为

$$\mathcal{F}_{\mathrm{BDeu}}(\mathrm{BN}, D) = \log \prod_{i=1}^{n} 0.001^{(r_i-1)q_i} \times \prod_{j=1}^{q_i} \frac{\Gamma(10/q_i)}{\Gamma(10/q_i + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(10/(r_i \cdot q_i) + N_{ijk})}{\Gamma(10/(r_i \cdot q_i))}. \quad (3)$$

其中: q_i 表示 V_i 父节点的取值数目; $\Gamma(\cdot)$ 表示 Gamma 函数; N_{ijk} 表示 V_i 取第 k 个值, 同时 V_i 的父节点取第 j 个值的样本数目, $N_{ij} = \sum_k N_{ijk}$. 在局部搜索定向过程中, 根据 BDeu 函数的可分解性和评分等价性, 本文只需要对包含在 T 中的任意三元组 $\mathrm{Tr}(V_i, V_j, V_k)$ 的以下 4 种结构进行局部评分, 即计算子结构 $V_i \to V_j \leftarrow V_k$, $V_j \to V_k \leftarrow V_i$, $V_k \to V_i \leftarrow V_j$ 和 $V_i - V_j - V_k - V_i$ 的评分增益. 因此, 最多计算 $4^{|T|}$ 次得分函数, 其中 $|T|$ 表示 T 中包含三元组的数目. 事实上, 根据网络的无圈约束和第 1 阶段的 CI 测试结果, 以上 4 种情况中可能存在无效的子结构, 从而实际计算评分函数的次数远远小于 $4^{|T|}$.

4　实　　验

为了测试 EC-MPD 算法的性能, 本文采用通用的 Benchmark 数据集 ALARM 网络[21]来完成下面的实验, 并将测试结果与基于等价类空间的 PC 算法[22]和 TPDA 算法[23]进行比较. 实验的运行环境为: 操作系统 Windows XP, CPU 为 Pentium4 2.8 GHz, 内存为 512 MB.

表 1　3 种算法在 ALARM 数据集上的实验结果

Data size	EC-MPD			PC			TPDA		
	ex	mi	SHD	ex	mi	SHD	ex	mi	SHD
1 000	2.8	3.6	15.4	1.3	8.1	24.1	64.6	2.5	84.3
2 000	2.9	2.4	15.2	0.0	6.3	22.3	62.2	3.3	83.5
5 000	2.9	1.4	15.2	0.2	3.4	20.9	65.5	2.5	88.5
8 000	2.2	0.7	14.8	0.1	2.5	21.0	54.8	2.5	74.5
10 000	2.4	0.6	12.8	0.3	1.7	20.7	51.5	2.5	73.6

为了保证 PC 算法和 TPDA 算法与本文算法在完全相同的实验条件下进行, 本文采用原网络的道德图作为初始候选图, 在不同样本数据集上分别独立运行 10 次进行测试, 测试结果如表 1 所示. 其中: mi 和 ex 分别表示与真实网络的 CPDAG 相比, 算法未确定出的平均边数目和算法确定出的平均冗余边数目; SHD 表示实验得到的网络与真实网络的结构海明距离, 即将算法确定的最优图结构转化为真实网络的 CPDAG 所需要的平均运算总数目 (包括添加边、删除边、反转边).

由表 1 可知, TPDA 算法得到的冗余边数目最多, PC 算法得到的冗余边数目最少而丢失的边数目最多, 而本文算法虽然丢失的边数目多于 PC 算法, 但结构海明距离是最小的, 并且随着样本量的增加, 学习精度比前两种算法有明显变化.

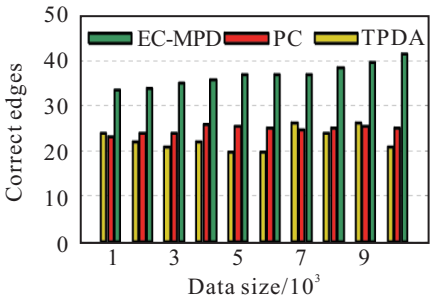


图 3　3 种算法在 ALAM 数据集上得到的正确边数目

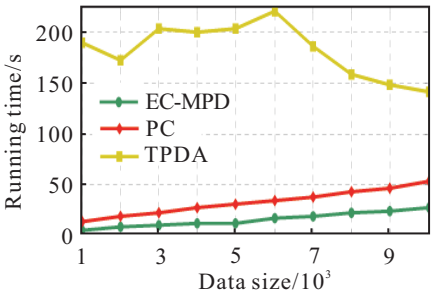


图 4　3 种算法的时间性能比较

图 3 为 PC 算法、TPDA 算法和本文算法在不同样本数据集上分别独立运行 10 次得到的平均正确边数目, 图 4 为相对应的运行时间比较. 由实验结果可知, 本文算法的运算时间在整个样本容量的范围上都优于其他两种算法, 并且在样本容量小的情况下得到的正确边数目最多, 同时在样本容量较大时, 本文算法的时间性能优势更加明显. 究其原因, 一方面, 本文算法利用评分函数对 CPDAG 中的不可逆有向边进行局部搜索, 提高了算法的学习精度; 另一方面, 该算法通过最大主子图分解技术分解约束空间, 极大地减少了大量冗余弧和变量引起的统计因子的计算, 同时, 对分解后的道德图仅利用 0 阶和 1 阶条件独立测试确定 V 结构, 从而有效减小了条件集的维数和独立性计算的次数, 具体表现在时间性能随样本容量的增加缓慢地增长, 这也表明本文算法具有良好的处理较大数据集的能力.

5　结　　论

本文提出了基于最大主子图分解的贝叶斯网络等价类学习算法. 该算法通过将一个大的网络结构分解为其最大主子图, 从而对 BN 结构学习进行分解, 有效降低了计算复杂度并提高了条件独立测试的效率. 同时本文从理论上严格证明了该算法的正确性和有效性, 且给出了算法的时间复杂度分析及实验分析, 为贝叶斯网络广泛用于解决实际问题提出了一种新的方案. 该方法适用于节点较多的大型网络和复杂型贝叶斯网络. 下一步将研究如何将该方法推广到链图的结构学习中, 并进行网络的局部优化学习.

参考文献(References)

[1] Cai Z Q, Sun S D, Si S B, et al. Identifying product failure rate based on a conditional Bayesian network classifier[J]. Expert Systems with Applications, 2011, 38(5): 5036-5043.

[2] Hsieh N C, Hung L P. A data driven ensemble classifier for credit scoring analysis[J]. Expert Systems with Applications, 2011, 37(1): 534-545.

[3] Sun Y, Tang Y Y, Ding S X, et al. Diagnose the mild cognitive impairment by constructing Bayesian network with missing data[J]. Expert Systems with Applications, 2011, 38(1): 442-449.

[4] Ben-Gal I, Shani A, Gohr A, et al. Identification of transcription factor binding sites with variable-order Bayesian networks[J]. Bioinformatics, 2005, 21(11): 2657-2666.

[5] Aquaroa V, Bardoscia M, Bellotti R, et al. A Bayesian Networks approach to Operational Risk[J]. Physica A, 2010, 389(8): 1721-1728.

[6] Bouchaala L, Masmoudi A, Gargouri F, et al. Improving algorithms for structure learning in Bayesian Networks using a new implicit score[J]. Expert Systems with Applications, 2010, 37(7): 5470-5475.

[7] Perrier E, Imoto S, Miyano S. Finding optimal bayesian network given a super-structure[J]. J of Machine Learning Research, 2008, 9(2): 2251-2286.

[8] Borgelt C. A conditional independence algorithm for learning undirected graphical models[J]. J of Computer and System Sciences, 2010, 76(1): 21-33.

[9] Ji J Z, Zhang H X, Hu R B, et al. A Bayesian network learning algorithm based on independence test and ant colony optimization[J]. Acta Automatica Sinica, 2009, 35(3): 281-288.

[10] Pena J M, Nilsson R, Bjorkegren J, et al. Towards scalable and data efficient learning of Markov boundaries[J]. Int J of Approximate Reasoning, 2007, 45(2): 211-232.

[11] Pellet J P, Elisseef A. Using Markov blankets for causal structure learning[J]. J of Machine Learning Research, 2008, 9(2): 1295-1342.

[12] 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.

[13] Akaike H. A new look at the statistical model identification [J]. IEEE Trans on Automatic Control, 1974, 19(6): 716-723.

[14] Schwarz G. Estimating the dimension of a model[J]. Annals of Statistics, 1978, 6(2): 461-464.

[15] Chickering D M. A transformational characterization of Bayesian network structures[C]. Proc of the Eleventh Conf on Uncertainty in Artificial Intelligence. Morgan Kaufmann, 1995: 87-98.

[16] Verma T, Pearl J. Equivalence and synthesis of causal models[C]. Proc of the 6th Conf on Uncertainty in Artificial Intelligence. New York: USA Elsevier Science Inc, 1990: 255-270.

[17] Chickering D M. Learning equivalence classes of Bayesian-network structures[J]. J of Machine Learning Research, 2002, 2(1): 445-498.

[18] Olesen K, Madsen A. Maximal prime subgraph decomposition of Bayesian networks[J]. IEEE Trans on Systems, Man and Cybernetics, Part B, 2002, 32(1): 21-31.

[19] Lauritzen S L, Graphical Models[M]. Oxford: Clarendon Press, 1996.

[20] Kullback S. Information Theory and Statistics[M]. Dover Publication, 1968.

[21] Beinlich I, Suermondt G, Chavez R, et al. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks[C]. Proc of the Second European Conf on Artificial Intelligence in Medicine. London, 1989: 247-256.

[22] Sprites P, Glymour C, Scheines R. Causation, prediction and search[M]. 2nd ed. The MIT Press, 2000.

[23] Cheng J, Greiner R, Kelly J, et al. Learning Bayesian networks from data: An information theory based approach[J]. Artificial Intelligence, 2002, 137(1/2): 43-90.