

采用约束蚁群优化的贝叶斯网结构学习算法

王翔, 郑建国, 张超群, 阎瑞霞

(东华大学旭日工商管理学院, 200092, 上海)

摘要: 针对贝叶斯网络结构学习问题, 提出了一种约束蚁群优化算法. 该算法根据贝叶斯得分的局部一致性原理设计了一种增边规则, 并且将此规则引入蚁群算法的框架中, 从而实现了在搜索过程中利用启发式信息动态缩减搜索空间、同时减少运行时间的目的. 此外, 还从理论上证明了增边规则的正确性, 而且从实验角度讨论了约束蚁群优化算法的参数敏感性. 实验结果表明, 在解决较大规模的贝叶斯网络结构学习问题时, 约束蚁群优化算法在保证求解精度的条件下比蚁群优化算法的运行时间减少 40% 以上.

关键词: 贝叶斯网络; 约束蚁群优化算法; 增边规则

中图分类号: TP18 **文献标志码:** A **文章编号:** 0253-987X(2011)08-0054-08

A Constrained Ant Colony Optimization Algorithm for Learning Bayesian Networks

WANG Xiang, ZHENG Jianguo, ZHANG Chaoqun, YAN Ruixia

(Glorious Sun School of Business and Management, Donghua University, Shanghai 200092, China)

Abstract: A novel constrained ant colony optimization algorithm is proposed for learning Bayesian networks. An add-edge-rule is designed in the proposed algorithm based on the locally consistent scoring criterion of BDeu metric. The add-edge-rule is embedded into the frame of ant colony optimization so that the heuristic information can be dynamically used to restrict the candidate evaluations during the search process and the running time of the new algorithm can be reduced. In addition, the correctness of the add-edge-rule is proved in theory and the parameter sensitivity of the constrained ant colony optimization is analyzed in experiments. Empirical tests show that without loss of results accuracy, the convergence speed of the proposed algorithm is at least 40% faster than that of the ant colony optimization algorithm for large-scale learning Bayesian networks.

Keywords: Bayesian networks; constrained ant colony optimization algorithm; add-edge-rule

经过 20 年的发展, 贝叶斯网络(BNs)^[1]已经成为不确定性知识表示和推理领域最重要的工具之一. 虽然 BNs 功能强大, 但是人工构建 BNs 是一个非常复杂且耗时的过程, 因此利用样本自动学习 BNs 结构成为该领域最重要的研究课题之一. 通常 BNs 结构学习算法分为如下 3 类: 基于约束的算法, 它的基本思想是用样本中出现的条件独立关系

构建 BNs 结构, 其典型算法有 PC^[2]等; 基于得分搜索的算法, 它的基本思想是利用得分函数 f 寻找最高得分的 BNs 结构, 其典型的算法有 MMHC^[3]、ACOB^[4]等; 混合算法, 它是以上两种算法思想的混合, 其中典型算法有 I-ACOB^[5]等, 因其能克服前两种算法的缺陷, 目前已成为研究热点之一.

ACOB 是一种基于得分搜索的 BNs 结构学习

算法,它虽然精度高,但是时间复杂度也高.因此,一些研究者提出了用 I-ACOB 等混合算法来改进 ACOB,它们均分两个独立阶段执行:针对每个节点利用条件独立性测试缩减候选解集;利用 ACOB 在缩减的候选解集中寻找最优的 BNs 结构.

与以上混合算法思路不同,本文提出了一种能同步执行以上两个独立阶段的约束蚁群优化贝叶斯算法(CACOB).新算法直接执行 ACOB,并且利用搜索每一步计算得到的启发式信息对候选解集进行删减.该算法不需要进行条件独立性测试,而是充分利用与基于 Dirichlet 先验分布的贝叶斯得分(BDEu)相关的启发式信息动态缩减搜索空间,从而加速收敛.仿真实验结果表明,当解决较大规模的 BNs 结构学习问题时,CACOB 可以在不损失求解精度的条件下比 ACOB 的运行时间明显减少.

1 BNs 的 BDEu 及特性

本节主要介绍 BDEu 的可分解形式以及局部一致性,这两者都是本文算法设计的基础.

1.1 BDEu 的可分解形式

BNs 是一种表达 n 个随机变量 $X = \{X_1, X_2, \dots, X_n\}$ 之间联合概率分布的工具,与之对应的 BDEu 是在已知样本集合 D 的条件下,对 BNs 结构 G 的一种度量,它是基于得分学习 BNs 结构算法的理论基础.一般情况下,可使用 BDEu 的分解形式将 G 的得分转化成每个节点的得分之和,具体为

$$f_{\text{BDEu}}(G; D) = \log(P(D | G)) = \sum_{i=1}^n f_{\text{BDEu}}(X_i | P_i) \quad (1)$$

其中

$$f_{\text{BDEu}}(X_i | P_i) = \log\left(\prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}\right) \quad (2)$$

式中: $D = \{d^1, d^2, \dots, d^N\}$ 是 N 个独立样本的集合, $d^l = (d_1^l, d_2^l, \dots, d_n^l)$ 是 n 个变量的一组观测值; P_i 有 q_i 种取值可能,表示 G 中 X_i 父节点的集合; X_i 有 r_i 种取值可能; N_{ijk} 表示 D 中 X_i 取第 k 个值且 P_i 取第 j 个值的频数, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ 表示 D 中 P_i 取第 j 个值的频数; $\alpha_{ij1}, \alpha_{ij2}, \dots, \alpha_{ijr_i}$ 服从 Dirichlet

分布, $\alpha_{ijk} = \frac{1}{q_i r_i} (k = 1, 2, \dots, r_i)$ 且 $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$.

基于式(1),BNs 结构学习演变成求最高 BDEu 即 $f_{\text{BDEu}}(G; D)$ 问题. BDEu 的可分解形式的优势在

于:当网络结构从 G 局部改变(增边、删边、反转边)为 G' 时,可分解的 BDEu 形式使得整个 BNs 结构得分的变化演变成一个或两个节点的得分变化,从而极大地提升了算法的执行效率,故本文选择式(1)作为 BNs 结构的评分标准.

1.2 BDEu 的局部一致性

BDEu 具有 3 个特性,即相等性、一致性和局部一致性,其中局部一致性是本文算法的基础,因此详细介绍.

定义 1 得分函数 f 的局部一致性^[6]. D 是某联合概率分布 P 的 m 个独立样本的集合, G 是一个有向无环图,在 G 的基础上增加边 $X_j \rightarrow X_i$ 得到 G' . 当样本量 m 非常大时,如果 f 满足以下两个条件:①如果 $\neg I_p(X_i, X_j | P_i)$,那么 $f(G'; D) > f(G; D)$;②如果 $I_p(X_i, X_j | P_i)$,那么 $f(G; D) > f(G'; D)$,则称 f 具有局部一致性,其中 $I_p(X_i, X_j | P_i)$ 表示分布 P 中两个节点 X_i 和 X_j 关于 X_i 的父节点集合 P_i 条件独立.

命题 1 BDEu 函数 f_{BDEu} 具有局部一致性^[6].

由局部一致性的定义可知,当增加边 $X_j \rightarrow X_i$ 时, f_{BDEu} 的变化能够反映节点 X_i 和 X_j 之间的条件独立关系.需要特别注意的是,由定义 1 可知 BDEu 的局部一致性只有在样本量 m 比较大时才能较好地成立.

2 CACOB

ACOB 是一种基于得分学习 BNs 结构的蚁群算法(ACO)^[7-11],它是目前求解 BNs 结构精度最高的算法之一,但是时间复杂度过高.本文提出的 CACOB 的主要目标是减少 ACOB 的运行时间,其具体实现思路是在 ACOB 的基础上充分利用 BDEu 的局部一致性,在选中边进入 BNs 后,根据启发式信息的变化进行节点间的条件独立性判断,缩减候选边的集合,加速算法收敛.

2.1 CACOB 的 4 个基本操作

CACOB 包含以下 4 个基本操作,其中第 4 个基本操作是新算法的主要创新之处,以下都假设 BNs 中存在 n 个顶点,且样本集合为 D .

(1)信息素矩阵 τ 及其相关操作: τ 是一个 $n \times n$ 的矩阵,其中元素 τ_{ij} 表示边 $X_j \rightarrow X_i$ 被选入 BNs 的概率, τ_{ij} 越大,则该边被选入 BNs 的概率就越大.按照执行的目标和时机,与 τ 相关的操作分 3 种.

①初始化操作,它在整个蚁群开始搜索之前按式(3)执行,使 BNs 中每条边被选中的可能性相同

$$\tau_{ij} = \tau_0 = \frac{1}{n | f_{\text{BDEu}}(G_{\text{K2SN}}; D) |} \quad (3)$$

式中: τ_0 称为初始化信息素水平; G_{K2SN} 表示利用 K2SN 算法^[12]求出的 BNs 结构。

②局部更新操作, 当一只蚂蚁选出一条边 $X_j \rightarrow X_i$ 加入 BNs 后, 按照式(4)执行局部更新操作, 它减弱了 $X_j \rightarrow X_i$ 再次被选中的可能, 以防止算法早熟

$$\tau_{ij} = (1 - \psi)\tau_{ij} + \psi\tau_0 \quad (4)$$

式中: ψ 称为局部挥发系数。

③全局更新操作, 当整个蚁群迭代一次之后, 按照式(5)执行全局更新操作, 增加了目前最优解的每条边被选中的概率

$$\begin{aligned} \tau_{ij} &= (1 - \rho)\tau_{ij} + \rho\Delta\tau_{ij} \quad (5) \\ \Delta\tau_{ij} &= \begin{cases} 1/f_{\text{BDEu}}(S^+; D), & \text{if } \{X_i \rightarrow X_j\} \in S^+ \\ \tau_{ij} & \end{cases} \end{aligned}$$

式中: ρ 称为全局性挥发系数; S^+ 表示到目前为止最优的 BNs 结构。

(2)启发式信息矩阵 η 及其相关操作: η 是一个 $n \times n$ 的矩阵, 其中元素 η_{ij} 表示边 $X_j \rightarrow X_i$ 加入 BNs 的启发式信息。按照目的和执行时机, 与 η 相关的操作分 2 种。

①初始化操作。它在一只蚂蚁开始寻找最优解之前按照式(6)执行, 初始化每条边的启发式信息

$$\eta_{ij} = f_{\text{BDEu}}(X_i | \{X_j\}) - f_{\text{BDEu}}(X_i | \Phi) \quad (6)$$

②更新操作。当 BNs 增加一条边 $X_j \rightarrow X_i$ 后, 首先利用 $\eta_{ij} = -\infty$ 使边 $X_j \rightarrow X_i$ 不能再次选中, 其次为了防止出现环, 使从 X_j 的祖先到 X_i 的子孙的边不能被选中, 最后按照式(7)更新所有以 X_i 为终点的边 $X_k \rightarrow X_i$ 的启发式信息

$$\eta_k = f_{\text{BDEu}}(X_i | P_i \cup \{X_k\}) - f_{\text{BDEu}}(X_i | P_i) \quad (7)$$

(3)状态转移规则 p : 它是一种选择边进入 BNs 的准则, 它在当前状态下依据 τ 和 η 从候选边集合 F_G 中选择一条边 $X_l \rightarrow X_r$ 进入 BNs。状态转移准则的思想是在开发和探索之间寻找一个平衡, 以概率 q_0 按照式(8)开发一条拥有最大被选中可能的候选边, 以概率 $1 - q_0$ 按照式(9)利用轮盘赌法从 F_G 中探索一条边

$$r, l = \begin{cases} \arg \max_{i, j \in F_G} \{[\tau_{ij}][\eta_{ij}]^\beta\}, & \text{if } q \leq q_0 \\ I, J & \text{if } q > q_0 \end{cases} \quad (8)$$

$$\begin{aligned} p_k(i, j) &= \\ \begin{cases} [\tau_{ij}][\eta_{ij}]^\beta / \sum_{u, v \in F_G} [\tau_{uv}][\eta_{uv}]^\beta, & \text{if } i, j \in F_G \\ 0 & \text{其他} \end{cases} \quad (9) \end{aligned}$$

式中: β 是状态转移规则中启发式信息相对于信息素信息的权重; q 是 $[0, 1]$ 之间均匀分布的随机数。

(4)增加一条边的操作: 它是指增加一条边进入 BNs 后, 缩减 BNs 候选解集空间的一系列操作。它根据 BDEu 的局部一致性, 利用启发式信息的变化得出节点间的条件独立性, 从而动态缩减候选解集。在 4 个基本操作中, 前 3 个是 ACOB 和 CACOB 都拥有的, 只有增加一条边的操作是 CACOB 特有的, 故以下对它的理论依据和详细实现进行解释。

命题 2 在样本量比较大的情形下, 设 η_{ij} 是算法的启发式信息, $\eta_{ij} = f_{\text{BDEu}}(X_i | P_i \cup \{X_j\}) - f_{\text{BDEu}}(X_i | P_i)$ 。若 $\eta_{ij} < 0$, 则 $I_p(X_i, X_j | P_i)$ 。

反证法: 由命题 2 不成立可知, 当启发式信息 $\eta_{ij} < 0$ 时存在 $\neg I_p(X_i, X_j | P_i)$, 根据命题 1, 从而

$$f_{\text{BDEu}}(G'; D) - f_{\text{BDEu}}(G; D) > 0$$

又由定义可知, $\eta_{ij} = f_{\text{BDEu}}(X_i | P_i \cup \{X_j\}) - f_{\text{BDEu}}(X_i | P_i) = f_{\text{BDEu}}(G'; D) - f_{\text{BDEu}}(G; D) < 0$, 矛盾。命题 2 得证。

由命题 2 可知, 从启发式信息 η_{ij} 可以推断节点 X_i 和 X_j 之间是否存在条件独立性关系。

定义 2 增边规则。当样本量比较大时, 在算法运行过程中, 若 $\eta_{ij} < 0$, 则 $F(X_i) = \{F(X_i), X_j\}$, $F(X_j) = \{F(X_j), X_i\}$ 。其中, $F(X_i)$ 称为禁止父节点集合, 表示不能成为 X_i 父节点的集合。

增边规则是一个对命题 2 的实现, 基本思想是在构建 BNs 的过程中若出现 $\eta_{ij} < 0$, 则 $I_p(X_i, X_j | P_i)$, 从而将 X_j 加入 $F(X_i)$, 又考虑到条件独立的对称性, 故将 X_i 加入 $F(X_j)$ 。

2.2 CACOB 的伪码

CACOB 包含两个重要组成部分: ①CAntB 算法, 它表示一只蚂蚁利用 τ 和 η 寻找 BNs 结构的流程, 需要特别指出的是它将增边规则引入寻优过程, 从而加速了算法收敛; ②CACOB 主控程序, 它表示整个蚁群基于 CAntB 算法寻找最优 BNs 结构的流程。在这两部分中, CAntB 算法是 CACOB 的基础。

图 1、2 分别是 CAntB 算法和 CACOB 主控程序的伪码, 本文用黑体标出 CACOB 与 ACOB 的区别。两者相比, CACOB 修改了 3 处: 当 η 更新时, 只要出现负值, CAntB 算法就调用增边规则; 将一只蚂蚁的搜索算法从 AntB 替换成 CAntB; 局部寻优算法从爬山算法(HillClimbing)^[13]改变为二次迭代约束爬山算法(2iConstrainedHillClimbing, 2iCHC)^[14]。

需要特别指出的是, 2iCHC 是一种基于命题 1 的改进 HillClimbing 算法, 与 CACOB 相同, 它也充分利

```
输入: 信息素矩阵 $\tau$ , 样本 $D$ 
输出: 父节点集合 $P_i (i=1,\cdots,n)$ 

//初始化
1. 父节点集合 $P_i (i=1,\cdots,n)$ : 初始化 $n$ 个节点的父节点集合 $P_i=\Phi$ ;
2. 启发式信息矩阵 $\eta$ : 初始化每条边  $X_j \rightarrow X_i$  的启发式信息
    $\eta_{ij}=f_{\text{BDEu}}(X_i|\{X_j\})-f_{\text{BDEu}}(X_i|\Phi)$ ;
3. for all  $i,j \in \{1,2,\cdots,n\} \ \&\& \ i \neq j$  do
   //增边规则
4.   if  $\eta_{ij} < 0$  then
5.      $F(X_j)=F(X_j) \cup \{X_j\}$ ,
      $F(X_j)=F(X_j) \cup \{X_i\}$ ,
      $\eta_{ij}=\eta_{ji}=-\infty$ ;
6.   end
7. end
//搜索过程
8. while  $\exists i,j (\eta_{ij} > 0)$ 
9.   利用状态转移准则 $p$  的 式(8)和式(9)选出一条
   增边  $X_j \rightarrow X_i$ ;
10.  if  $\eta_{ij} > 0$  then
11.     $P_i=P_i \cup \{X_j\}$ ;
12.  end
13.   $\eta_{ij}=-\infty$ ;
14.  for all  $X_a \in \text{Ancestors}(X_j) \cup \{X_j\}$  and  $X_b \in \text{Descendants}$ 
     $(X_i) \cup \{X_i\}$  do
15.     $\eta_{ab}=-\infty$ ; //防止出现环
16.  end
17.  for  $k=1$  to  $n$  do
18.    if  $\eta_{ik} > -\infty$  then
19.       $\eta_{ik}=f_{\text{BDEu}}(X_i|P_i \cup \{X_k\})-f_{\text{BDEu}}(X_i|P_i)$ ;
      //增边规则
20.      if  $\eta_{ik} < 0$  then
21.         $F(X_j)=F(X_j) \cup \{X_k\}$ ;
22.         $F(X_k)=F(X_k) \cup \{X_i\}$ ;
23.         $\eta_{ik}=\eta_{ki}=-\infty$ ;
24.      end
25.    end
26.  end
//信息素局部更新
27.  利用式(4)局部更新边  $X_j \rightarrow X_i$  的信息素 $\tau_{ij}$ ;
28. end
29. return 父节点集合  $P_i (i=1,\cdots,n)$ ;
```

图 1 CAntB 算法伪码

用了 BDEu 的局部一致性在搜索过程中逐渐缩减搜索空间,从而加快收敛速度,故本文选用它作为局部搜索算法。

2.3 CACOB 的特点

与 ACOB 相比,CACOB 具有以下两个特点。

(1)CACOB 搜索空间较小. 由算法伪码及定义 2 可知,CACOB 是在 ACOB 的基础上利用 BDEu 的局

```
输入: 样本 $D$ , 最大迭代 $t_{\text{max}}$ 次,  $m$ 只蚂蚁, 局部寻优的频率 $t_{\text{step}}$ 
输出: 图 $G_+$ 

//初始化
1. 利用K2SN算法得到图 $G_{\text{K2SN}}$ ;
2. 根据式(3)得到 $\tau_0$ ;
3. 初始化信息素矩阵 $\tau$ : 初始化每条边 $X_j \rightarrow X_i$ 的信息素 $\tau_{ij}=\tau_0$ ;
4. 初始化最优解 $G_+$ :  $G_+=G_{\text{K2SN}}$ ;
//搜索过程, 最大迭代 $t_{\text{max}}$ 次,  $m$ 只蚂蚁
5. for  $t=1$  to  $t_{\text{max}}$  do
   // $m$ 只蚂蚁进行搜索
6.   for  $k=1$  to  $m$  do
7.      $G_k=\text{CAntB}()$ ; /*调用CAntB算法*/
8.     if  $\text{mod}(t,t_{\text{step}})=0$  then
       /*调用约束爬山算法*/
9.        $G_k=2\text{iConstrainedHillClimbing}(G_k)$ ;
10.    end
11.  end
12.   $G_b=\arg \max_{k:1,\cdots,m} f_{\text{BDEu}}(G_k;D)$ ;
13.  if  $f_{\text{BDEu}}(G_b;D) > f_{\text{BDEu}}(G_+;D)$ 
14.     $G_+=G_b$ ;
15.  end
//信息素 $\tau$ 全局性更换
16.  利用  $f_{\text{BDEu}}(G_+;D)$ , 根据式(6)更新信息素矩阵 $\tau$ ;
17. end
18. return  $G_+$ ;
```

图 2 CACOB 主控程序伪码

部一致性原理进一步减小了搜索空间,故CACOB比 ACOB 搜索空间小。

(2)CACOB 在样本较大时表现较好. 一方面 CACOB 缩减搜索空间的理论基础是 BDEu 的局部一致性,另一方面由定义 1 可知局部一致性在样本量较大时才有效,因此当样本量较大时 CACOB 表现较好。

2.4 CACOB 的时间复杂度

CACOB 中最耗时的基本操作是计算增加一条边 $X_j \rightarrow X_i$ 的启发式信息 η_{ij} , 故度量时间复杂度时以计算一次 η_{ij} 为基本单位. CACOB 拥有 m 只蚂蚁,同时进行 t_{max} 次迭代,从而执行了 $m \times t_{\text{max}}$ 次 CAntB 算法,同时最坏的情形下 CAntB 算法需要计算 $O(n^3)$ 次启发式信息,故 CACOB 需要计算 $O(m \times t_{\text{max}} \times n^3)$ 次启发式信息。

3 实验结果及讨论

3.1 评价网络及评价指标

本文分别选择规模较大的 ALRAM 网络(37 个顶点、46 条边)以及规模较小的 ASIA 网络(8 个顶点、8 条边)来评价算法的性能。

为度量学习得到的 BNs 质量和学习算法的时间复杂度,本文选用了文献[3]的 3 个指标: f_{BDEu} 、结构海明距离 d_{SHD} 和统计调用次数 n_{NSC} . 其中: f_{BDEu} 是度量结构优劣的标准; d_{SHD} 是度量学习得到的结构 G' 与真实结构 G 之间差异的指标,它表示原网络 G 与学习得到的网络 G' 完全匹配时需要执行的最小操作步骤数; n_{NSC} 是度量算法时间复杂度的指标,一次统计调用表示计算一次节点的 $f_{\text{BDEu}}(X_i|P_i)$.

3.2 CACOB 和 ACOB 的参数设置

为公平起见,两算法的参数设置完全相同,具体设置如下:式(4)中 $\psi=0.4$;式(5)中 $\rho=0.4$;式(8)中 $q_0=0.7, \beta=2$;蚁群大小 $m=10$;最大迭代次数 $t_{\text{max}}=50$;局部搜索的频数 $t_{\text{step}}=5$.

需要特别指出的是,在以上 7 个参数中, q_0 和 t_{step} 对算法影响最大,3.5 节将详细讨论它们对算法性能的影响,并给出了设置原则.

3.3 CACOB 和 ACOB 的 3 个指标的实验比较及讨论

为了比较 CACOB 与 ACOB,本文针对 ALARM 网络和 ASIA 网络的 3 个不同样本尺寸 (1 000、3 000、5 000) 分别执行了 20 次独立实验. 为了具有直观可比性,3 个指标都利用 CACOB 的相应指标进行了标准化. 需要注意的是,由于 f_{BDEu} 小于 0, 故标准化后不是越大越好,而是越小越好.

本部分用图和表显示每个指标比较的结果. 图 3 ~5 以直观的方式比较相应指标 20 次实验的均值;表 1~3 以更具体的方式展示了相应指标 20 次实验的均值和标准差.

3.3.1 CACOB 和 ACOB 的 f_{BDEu} 比较 f_{BDEu} 与 d_{SHD} 都是度量结构优劣的指标,不同的是 f_{BDEu} 还是两种算法设计的基础,故更重要. 图 3 和表 1 展示了两种算法针对 ALARM 网络和 ASIA 网络的 3 个不同样本尺寸的 f_{BDEu} 比较. 为了评估两种算法是否存在显著性差异,表 1 中还展示了配对 Wilcoxon-Ranksum (WR) 检验的 p 值,该检验设定显著性水平为 0.05.

由图 3a 可知,针对规模较大的 ALARM 网络,在样本尺寸 1 000 和 3 000 上 ACOB 求解 BNs 的质量略优于 CACOB,在样本尺寸 5 000 上正好相反.

由图 3b 可知,针对规模较小的 ASIA 网络,在 3 个样本尺寸上 ACOB 求解 BNs 的质量都优于 CACOB.

由表 1 可知,两种算法的 f_{BDEu} 非常接近. WR 检验对应的 p 值显示:针对规模较大的 ALARM 网络,样本尺寸为 5 000 时,CACOB 显著优于 ACOB;样本

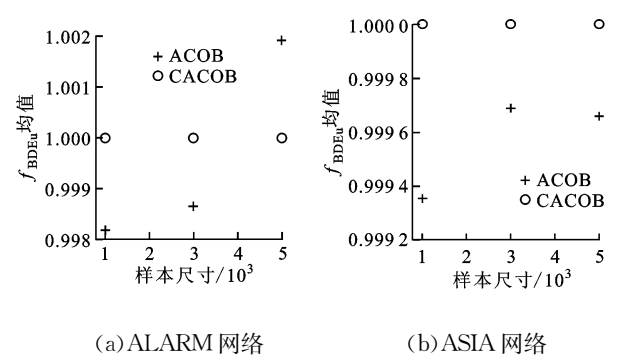


图 3 两种算法针对两个网络标准化的 BDEu

尺寸为 3 000 时,ACOB 显著优于 CACOB;样本尺寸为 1 000 时,两种算法无明显差异.

表 1 f_{BDEu} 的统计结果

测试数据集	算法	f_{BDEu} 均值(标准差)		
		样本尺寸 1 000	样本尺寸 3 000	样本尺寸 5 000
ALARM	CACOB	1.000(0.004)	1.000(0.001)	1.000(0.001)
	ACOB	0.998(0.002)	0.999(0.002)	1.002(0.012)
	WR 检验	$p=0.156$	$p=0.033^*$	$p=0.000^*$
ASIA	CACOB	1.000 (1.95×10^{-4})	1.000 (3.84×10^{-4})	1.000 (3.29×10^{-4})
	ACOB	0.999 (2.80×10^{-4})	1.000 (4.91×10^{-5})	1.000 (6.94×10^{-5})
	WR 检验	$p=0.454$	$p=0.108$	$p=0.568$

注:加*的 p 值表示 WR 检验显著.

针对规模较小的 ASIA 网络,WR 检验结果显示在 3 个样本尺寸上两种算法均无明显差异.

3.3.2 CACOB 和 ACOB 的 d_{SHD} 比较 d_{SHD} 的功能类似, d_{SHD} 也是对 BNs 结构优劣评判的一个指标,差异之处在于它是从距离角度度量优劣. 显然,这个指标应该越小越好. 图 4 和表 2 展示了两种算法针对 ALARM 网络和 ASIA 网络的 3 个不同样本尺寸的 d_{SHD} 比较.

表 2 d_{SHD} 的统计结果

测试数据集	算法	d_{SHD} 均值(标准差)		
		样本尺寸 1 000	样本尺寸 3 000	样本尺寸 5 000
ALARM	CACOB	1.000(0.212)	1.000(0.161)	1.000(0.721)
	ACOB	1.346(0.433)	1.024(0.576)	1.183(0.648)
ASIA	CACOB	1.000(0.323)	1.000(0.318)	1.000(0.495)
	ACOB	0.557(0.410)	0.278(0.256)	0.119(0.291)

由图 4a 和表 2 可知,针对规模较大的 ALARM 网络,在 3 个样本尺寸上 CACOB 的 d_{SHD} 都小于

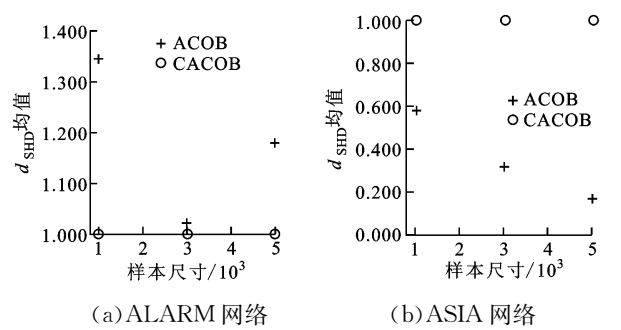


图 4 两种算法针对两个网络标准化的结构海明距离

ACOBS的 d_{SHD} ,故CAOB的求解质量优于ACOB. 由图4b和表2可知,针对规模较小的ASIA网络,在3个样本尺寸上CACOB的 d_{SHD} 都大于ACOB的 d_{SHD} ,因此CACOB的求解质量劣于ACOB.

3.3.3 CACOB和ACOB的 n_{NSC} 比较 n_{NSC} 是对算法收敛速度的一种度量,其值越小表示算法收敛速度越快,反之则越慢.图5和表3展示了两种算法针对ALARM网络和ASIA网络的3个不同样本尺寸的 n_{NSC} 比较.

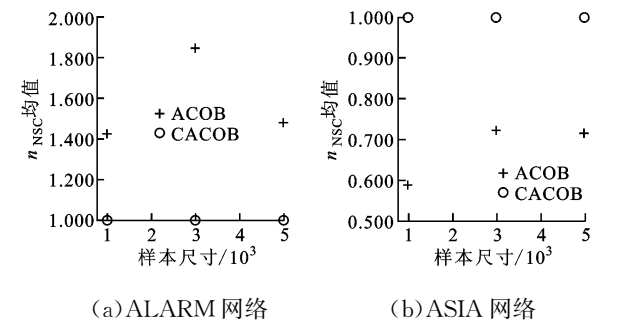


图 5 两种算法针对两个网络标准化的统计调用次数

表 3 n_{NSC} 的统计结果

测试数据集	算法	n_{NSC} 均值(标准差)		
		样本尺寸 1 000	样本尺寸 3 000	样本尺寸 5 000
ALARM	CACOB	1.000(0.314)	1.000(0.323)	1.000(0.229)
	ACOB	1.426(0.788)	1.844(1.281)	1.482(0.955)
ASIA	CACOB	1.000(0.865)	1.000(1.541)	1.000(1.238)
	ACOB	0.586(0.817)	0.723(0.914)	0.714(1.237)

由图5和表3可知,针对规模较大的ALARM网络,在3个样本尺寸上CACOB的 n_{NSC} 无论是均值还是标准差都显著低于ACOB的 n_{NSC} .这表示在解决较大规模的BNs结构学习问题时,CACOB运行时间稳定,且明显优于ACOB.

与之相反,针对规模较小的ASIA网络,在3个样本尺寸上CACOB的 n_{NSC} 无论是均值还是标准差都显著高于ACOB的 n_{NSC} .这表示在解决较小规模

的BNs结构学习问题时,ACOB运行时间稳定,且明显优于CACOB.

3.3.4 三指标比较结果及讨论 综合以上两个网络3个指标的比较结果,可得到以下两个结论.

(1)当解决规模较小的BNs(例如ASIA网络)结构学习问题时,CACOB无论是求解质量还是收敛速度都不如ACOB,这种现象是合理的.首先,BNs结构学习问题是典型的NP难问题,当问题规模较小时,可行解空间也较小,故在实验中ACOB可以很快找到全局最优解,从而达到收敛状态;其次,由2.3节的讨论可知CACOB的搜索空间小于ACOB,而且这种缩减可能会把全局性最优解排除在外,从而使CACOB由于找不到全局最优解而一直搜索下去,因此新算法求解精度较差,同时收敛速度也较慢.

(2)当解决规模较大的BNs(例如ALARM网络)结构学习问题时,CACOB比ACOB收敛速度快,而且两者求解质量相当,同时此结论在样本比较大时更明显.首先,由2.3节的讨论可知CACOB的搜索空间小于ACOB,故CACOB的收敛速度快于ACOB;其次,由2.3节的讨论可知CACOB在样本量较大时表现较好,故以上结论在样本较大时更明显;再次,由于BNs结构学习问题是典型的NP难问题,从而当问题规模较大时可行解空间呈现指数增长,故在实验中两种算法都无法遍历所有的局部极值状态,故两种算法求解质量相当.

3.4 CACOB与ACOB的 n_{NSC} 与 f_{BDEu} 的相关性分析

由于CACOB只对较大规模的BNs有效,故本文只讨论ALARM网络的情形.图6a、6b针对样本尺寸为5 000的ALARM网络,分别展示了20次独立实验中随着CACOB和ACOB的运行,学习得到BNs的进化过程.

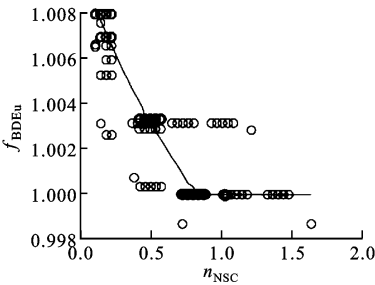
比较图6a和6b可知,当算法收敛时,CACOB的 n_{NSC} 比ACOB的 n_{NSC} 小,故其收敛速度快于ACOB.

3.5 CACOB算法参数 q_0 和 t_{step} 的敏感性分析

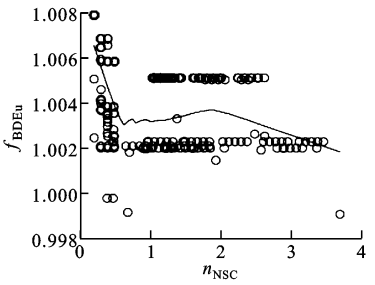
CACOB有多个参数,但对其效果影响最大的是状态转移准则的开发概率 q_0 和局部搜索算法的频率 t_{step} .上文CACOB选用 $q_0=0.7$ 、 $t_{step}=5$.为了研究这两个参数对CACOB的影响,在其他实验参数保持不变的前提下,本部分针对拥有5 000个样本的ALRAM网络进行了一系列的实验,其中 q_0 分别为0.6、0.7、0.8、0.9, t_{step} 分别为5、10、15,它们共计有12种组合形式.表4是CACOB使用不同参数

表 4 CACOB 算法不同 q_0 和 t_{step} 组合标准化的 f_{BDEu} 、 d_{SHD} 和 n_{NSC}

t_{step}	f_{BDEu} 均值(标准差)				d_{SHD} 均值(标准差)				n_{NSC} 均值(标准差)			
	$q_0=0.6$	$q_0=0.7$	$q_0=0.8$	$q_0=0.9$	$q_0=0.6$	$q_0=0.7$	$q_0=0.8$	$q_0=0.9$	$q_0=0.6$	$q_0=0.7$	$q_0=0.8$	$q_0=0.9$
5	1.000	1.000	1.000	1.001	1.000	1.101	0.995	1.288	1.000	0.889	0.857	0.816
	(1.02×10^{-3})	(7.99×10^{-4})	(7.43×10^{-4})	(2.03×10^{-3})	(0.403)	(0.794)	(0.303)	(0.679)	(0.414)	(0.204)	(0.147)	(0.349)
10	1.000	1.000	1.000	1.001	0.960	0.990	1.288	1.157	1.531	1.525	1.434	1.270
	(9.42×10^{-4})	(7.19×10^{-4})	(1.39×10^{-3})	(2.11×10^{-3})	(0.283)	(0.375)	(0.836)	(0.632)	(0.687)	(0.569)	(0.488)	(0.333)
15	1.000	1.000	1.000	1.001	0.944	0.949	0.894	1.348	1.957	1.914	1.941	1.627
	(7.63×10^{-4})	(2.15×10^{-4})	(1.82×10^{-5})	(1.9×10^{-3})	(0.208)	(0.159)	(0.068)	(0.759)	(0.611)	(0.484)	(0.455)	(0.546)



(a)CACOB 算法



(b)ACOB 算法

圆圈表示算法每迭代一次后记录其相应的 n_{NSC} 与 f_{BDEu}

图 6 两种算法的 f_{BDEu} 随 n_{NSC} 的变化

组合进行 20 次独立实验在 f_{BDEu} 、 d_{SHD} 和 n_{NSC} 这 3 个指标上的均值和标准差。3 个指标所有实验数据使用 $q_0=0.6$ 、 $t_{\text{step}}=5$ 的相应数据进行了标准化。

由表 4 可以得到如下结论。

(1)随着 t_{step} 减小以及 q_0 增大, n_{NSC} 会逐渐减小,收敛速度逐渐加快。这种现象是合理的,若 t_{step} 越小,则调用局部搜索算法 2iCHC 就越频繁,从而收敛越快;同时, q_0 越大,则 CACOB 的随机性越小,从而收敛越快。因此,在 12 种参数组合中当 $q_0=0.9$ 、 $t_{\text{step}}=5$ 时, CACOB 拥有最小的 n_{NSC} 。

(2)随 t_{step} 增加, d_{SHD} 减小,求解质量提高。这是因为当 t_{step} 较大时,调用局部搜索算法的时间间隔较大,从而 CACOB 有充足的运行次数找到新的更优解。但是,有两种参数组合违反了这个规律,分别是 $t_{\text{step}}=10$ 、 $q_0=0.8$ 和 $t_{\text{step}}=15$ 、 $q_0=0.9$ 。

综合以上结论,要在优先考虑收敛速度的前提下保证求解质量,建议参数设置为 $q_0=0.8$ 、 $t_{\text{step}}=5$ 。

4 结 论

本文针对 ACOB 收敛速度较慢的问题,提出了一种能够同时进行条件独立性判断和得分搜索的改进算法 CACOB。新算法的基本思想是根据 BDEu

的局部一致性原理,设计了增边规则,利用启发式信息作为条件独立性判断的依据,对候选边集合进行约简,从而减小了搜索空间。

仿真实验结果证明了以下结论。

- (1)当解决较大规模的 BNs 结构学习问题时, CACOB 比 ACOB 收敛速度快,且两者构建的 BNs 质量相当。
- (2)CACOB 在样本量较大时表现较好。
- (3)推荐 CACOB 的参数取 $q_0=0.8$ 、 $t_{\text{step}}=5$ 。

参考文献:

[1] HECKERMAN D. Bayesian networks for data mining [J]. Data Mining and Knowledge Discovery, 1997, 1(1): 79-119.

[2] 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.

[3] DE CAMPOS L M, FERN N J M, MEZ G J A, et al. Ant colony optimization for learning Bayesian networks [J]. International Journal of Approximate Reasoning, 2002, 31(3): 291-311.

[4] TSAMARDINOS I, BROWN L E, ALIFERIS C F.

- The max-min hill-climbing Bayesian network structure learning algorithm[J]. Machine Learning, 2006, 65(1): 31-78.
- [5] JI J, ZHANG H, HU R, et al. Bayesian network learning algorithm based on independence test and ant colony optimization [J]. Acta Automatica Sinica, 2009, 35: 281-288.
- [6] CHICKERING D M. Optimal structure identification with greedy search[J]. The Journal of Machine Learning Research, 2003, 3: 554.
- [7] DORIGO M, MANIEZZO V, COLORNI A. Ant system: optimization by a colony of cooperating agents [J]. Systems, Man, and Cybernetics: Part B Cybernetics, 2002, 26(1): 29-41.
- [8] 张云,冯博琴,麻首强,等. 蚁群-遗传融合文本聚类算法[J]. 西安交通大学学报, 2007, 41(10): 1146-1150.
- ZHANG Yun, Feng Boqin, MA Shouqiang, et al. Text clustering based on fusion of ant colony and genetic algorithms[J]. Journal of Xi'an Jiaotong University, 2007, 41(10): 1146-1150.
- [9] 任志刚,冯祖仁,柯良军. 蚁群优化属性约简算法[J]. 西安交通大学学报, 2008, 42(4): 440-444.
- REN Zhigang, Feng Zuren, KE Liangjun. Ant colony optimization approach to attribute reduction problem [J]. Journal of Xi'an Jiaotong University, 2008, 42(4): 440-444.
- [10] 郑巍,刘三阳,寇晓丽. 一种面向传感器网络的蚁群优化路径恢复算法[J]. 西安交通大学学报, 2010, 44(1): 83-86.
- ZHENG Wei, LIU Sanyang, KOU Xiaoli. A route restoration algorithm for sensor network via ant colony optimization[J]. Journal of Xi'an Jiaotong University, 2010, 44(1): 83-86.
- [11] 张兆军,冯祖仁,任志刚. 采用序优化的改进蚁群算法[J]. 西安交通大学学报, 2010, 44(2): 15-19.
- ZHANG Zhaojun, FENG Zuren, REN Zhigang. Novel ant colony optimization algorithm based on order optimization[J]. Journal of Xi'an Jiaotong University, 2010, 44(2): 15-19.
- [12] DE CAMPOS L M, PUERTA J M. Stochastic local and distributed search algorithms for learning belief networks[C] // Proceedings of the III International Symposium on Adaptive Systems: Evolutionary Computation and Probabilistic Graphical Model. Havana, Cuba: ICIMAF, 2001: 109-115.
- [13] 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.
- [14] GAMES J A, MATEO J L, PUERTA J M. Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood [J]. Data Mining and Knowledge Discovery, 2011, 22(1): 106-148.

[本刊相关文献链接]

动态团队定向问题的模型及其优化算法. 西安交通大学学报, 2011, 45(6): 54-58.

一种高效的对等网络流媒体数据调度算法. 西安交通大学学报, 2011, 45(6): 20-26.

一种修复网络拓扑的 Steiner 树移动控制算法. 西安交通大学学报, 2011, 45(2): 39-43.

可演化网络中移动代理的性能分析及优化方法. 西安交通大学学报, 2010, 44(12): 5-9.

利用蚁群优化的非均匀分簇无线传感器网络路由算法. 西安交通大学学报, 2010, 44(6): 33-38.

采用序优化的改进蚁群算法. 西安交通大学学报, 2010, 44(2): 15-19.

一种面向传感器网络的蚁群优化路径恢复算法. 西安交通大学学报, 2010, 44(1): 83-86.

一种可演化网络的研究与实践. 西安交通大学学报, 2008, 42(10): 1193-1203.

蚁群优化属性约简算法. 西安交通大学学报, 2008, 42(4): 440-444.

用于无线传感器网络的比例公平队列调度算法. 西安交通大学学报, 2008, 42(2): 129-132.

蚁群-遗传融合文本聚类算法. 西安交通大学学报, 2007, 41(10): 1146-1150.

(编辑 荆树蓉 武红江)