文章编号:1000－2367(2015)04－0016－05 DOI:10.16366/j.cnki.1000－2367.2015.04.003

基于蜂群算法和遗传算法的贝叶斯网络结构混合学习方法

汪春峰[1],将 妍[2,3]

(1.河南师范大学 数学与信息科学学院,河南 新乡 453007;2.河南大学 环境与规划学院,河南 开封 475001;
3.郑州旅游职业学院 基础部,郑州 450009)

摘 要:变量之间的关系对解释数据具有重要作用,而贝叶斯网络恰恰是表示变量之间关系的重要工具.针对贝叶斯网络结构学习问题,基于蜂群算法(ABC)和遗传算法(GA),提出一个新的混合型算法(ABC-GA).由于 ABC-GA 融合了 ABC 算法和 GA 算法的长处,所以可以弥补单独使用任一算法的缺陷.数值试验结果表明:ABC-GA 算法具有较高的计算效率和计算精度.

关键词:贝叶斯网络结构学习;ABC 算法;GA 算法;无约束优化

中图分类号:TP301 **文献标志码**:A

贝叶斯网络(BNs)是联合概率分布的图形表示,它以其坚实的理论基础、形象直观的知识表示、灵活的推理能力和接近人类思维特征的决策机制等特点,已成为机器学习和数据挖掘等领域中处理不确定性的主要方法.贝叶斯网络研究的一个重要方向是贝叶斯网络结构学习.最早,贝叶斯网络的构建是由人工完成的,但是人工构建 BNs 是一个非常复杂且耗时的过程,因此如何利用样本自动学习 BNs 结构成为该领域最重要的研究课题之一.经过近 10 年的发展,在基于数据建立贝叶斯网络结构研究方面,已相继产生了许多著名算法[2－5].尽管这些算法的求解策略不同,但是从求解机制的角度来看,贝叶斯网络结构学习算法大致可以分为:基于评分搜索的方法[3],基于依赖分析的方法[6]以及这两种方法结合起来的混合方法[7].由于贝叶斯网络结构学习是 NP-难问题[8],所以,对于节点比较多的情况,现有大多数算法都采用了随机搜索机制.在过去十几年里,人们已经提出了许多随机搜索方法,如遗传算法[9],蚁群优化算法[10]和粒子群算法[11].

基于蜂群算法(ABC)和遗传算法(GA),本文提出一个学习贝叶斯网络结构的混合算法(ABC-GA).该算法与其他学习算法性能的比较显示,本文方法具有更高的执行效率和学习精度.

1 贝叶斯网络

贝叶斯网络是一个有向无环图,图中每个节点代表一个向量,每个向量与一个条件概率表有关.贝叶斯网络由两部分组成:一个表示 X 中变量条件独立关系的网络结构 G;一个表示变量相联系的局部条件概率表 P.

通常贝叶斯网络 B 可表示为一个二元组 $B=<G,\theta>$,其中 $G=<X,E>$ 是网络结构图,其节点对应于随机变量 X 内的相应变量,$E=\{e_{ij}\mid 1\leq i,j\leq n,\text{且 } i\neq j\}$ 是有向边集合,e_{ij} 是从节点 x_i 到节点 x_j 的有向边,表示随机变量 x_i 对 x_j 的条件依赖关系;$\theta=<\theta_1,\theta_2,\cdots,\theta_n>$ 为网络参数,是一组条件概率表,刻画了局部变量间的随机依赖关系,θ_i 是对应于随机变量 x_i 的条件概率表.

从数据中学习贝叶斯网络包括结构学习和网络参数学习,其核心内容是结构学习.本文主要解决网络

收稿日期:2014-11-11;**修回日期**:2015-03-23.

基金项目:国家自然科学基金(U1404105;11171094);河南省科技攻关研究计划项目(142102210058);河南师范大学国家级科研项目培育基金(01016400105);河南师范大学博士科研启动课题项目(qd12103);河南师范大学校级骨干教师培养项目;河南师范大学青年科学基金项目(2013qk02).

作者简介(通信作者):汪春峰(1978－),男,河南开封人,河南师范大学副教授,博士,主要从事贝叶斯网络结构学习、最优化理论方法及应用的研究,E-mail: wangchunfeng09@126.com

结构学习问题. 结构学习主要困难在于如何从众多可能的结构中找到最合适的依赖结构. 对于 n 个节点所构成的网络中包含贝叶斯网络的数目计算公式如下：

$$g(n) = \sum_{i=1}^{n} (-1)^{i+1} C_i^n 2^{i(n-1)} g(n-i). \tag{1}$$

由式(1) 知，当 $n = 10$ 时，需要搜索的模型个数将达 4.17×10^{18}，可见搜索空间巨大.

在学习贝叶斯网络结构的算法中，目前研究比较多的是基于评分搜索的算法. 此类算法所采用的打分函数都是在统计数据集合的基础上设计的，大体上可以分为依据信息论原理设计的打分函数和依据贝叶斯方法设计的打分函数. 这些打分函数通常具有可分解性，即：

$$f(G : D) = \sum_{i=1}^{n} f(V_i, Pa(V_i) : N_{V_i}, Pa(V_i)), \tag{2}$$

其中 $N_{V_i, Pa(V_i)}$ 是 V_i 和 $Pa(V_i)$ 在给定数据 D 中的统计量，更详细内容可参看文献[12].

2　贝叶斯网络结构学习的混合算法(ABC-GA)

在这一部分，基于 ABC 算法和 GA 算法，给出本文所提的混合算法 ABC-GA. 由于该算法结合了两类算法的长处，新算法有望具备更好的性能.

2.1　经典 ABC 算法

ABC 算法[13] 是由 Karaboga 在 2005 年提出的，该算法是一种群智能算法，它是模拟了蜂群采蜜行为的随机优化算法. 人工蜂群有 3 个基本组成部分：雇佣蜂、观察蜂和侦察蜂. 蜜源的搜索过程可概括为：雇佣蜂确定食物源 X_i，计算该食物源处的适应度 $f(x_i)$，与观察蜂共同分享相关信息；观察蜂以一定的概率 P_i 在临近食物源中选择开发对象，其中 P_i 定义如下：

$$P_i = \frac{f(X_i)}{\sum_{k=1}^{N} f(X_k)}, \tag{3}$$

这里 N 为食物源的个数；被放弃的食物源处的蜜蜂转换为侦查蜂，并开始随机搜索新的食物源.

ABC 算法步骤

(1) 初始化

(2) 重复以下步骤

(a) 将雇佣蜂与食物源一一对应，同时确定食物源的适应度.

(b) 观察蜂根据雇佣蜂提供的信息，以(3)所确定的概率选择食物源，同时确定食物源的适应度.

(c) 确定侦查蜂，寻找新的食物源.

(d) 记忆迄今为止最好的食物源.

(3) 判断终止条件是否成立

在 ABC 算法中，每个食物源代表优化问题的一个可能解，食物源的适应度对应于解的质量.

2.2　贝叶斯网络学习的遗传算法(GA)

遗传算法是基于达尔文进化论，模拟生物界自然进化和遗传过程的随机搜索算法. 遗传算法通过选择、交叉和变异三种基本操作寻找最优个体，是处理复杂优化问题的一类算法，具有较高的鲁棒性和广泛的适用性，已被应用于许多领域，包括贝叶斯网络结构学习. 下面对遗传算法进行贝叶斯网络结构学习的基本环节介绍.

2.2.1　编码及基本操作算子

由于算法的搜索空间是由贝叶斯网络结构组成的，且贝叶斯网络结构可表示成如下邻接矩阵 $(C_{ij})_{n\times n}$ 的形式：$c_{ij} = \begin{cases} 1 & \text{如果 } i \text{ 是 } j \text{ 的父节点}, \\ 0 & \text{否则}. \end{cases}$ 所以，在学习贝叶斯网络的遗传算法中，其染色体可由贝叶斯网络结构对应邻接矩阵的压缩行向量表示：

$$c_{11}c_{12}\cdots c_{1n}c_{21}c_{22}\cdots c_{2n}\cdots c_{n1}c_{n2}\cdots c_{m}.$$

在介绍了编码方式后,下面通过简单例子介绍遗传算法学习贝叶斯网络结构的基本操作算子.

例 考虑如图 1 所示的两个 3 节点的贝叶斯网络

其邻接矩阵分别为

$$\begin{pmatrix}0&0&1\\0&0&1\\0&0&0\end{pmatrix}\text{和}\begin{pmatrix}0&0&0\\0&0&0\\1&1&0\end{pmatrix}.$$

采用上面介绍的编码方式,这两个贝叶斯网络的编码分别为:001001000 和 000000110.

交叉算子:假定以上两个个体要进行交叉操作,且交叉位置为随机确定的第 6 个位置,则交叉后产生的后代为分别为:00100110 和 000000000,其对应的网络结构如图 2 所示.

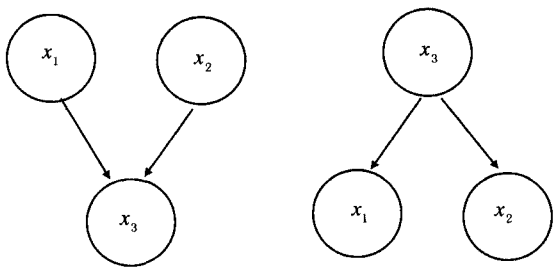

图 1 3节点贝叶斯网络

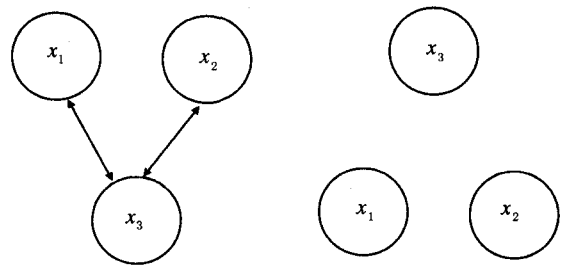

图 2 交叉后的网络结构

显然,交叉后的网络不一定满足有向无环这一要求,因此,上述交叉算子不是闭算子.

变异算子:考虑如图 3(a)所示的贝叶斯网络,其个体编码为:010001000.假定在第 7 个位置处要进行变异操作,变异后产生的个体编码为:010001100,其对应的网络结构图如图 3(b).

显然变异算子也不是闭算子.

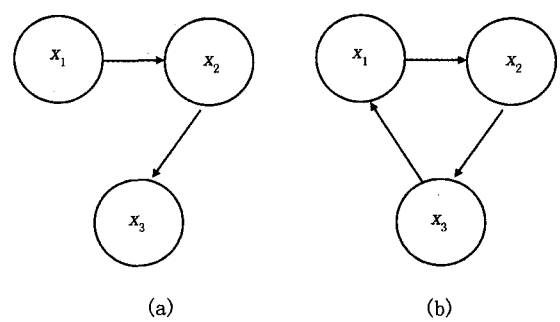

图3 变异前 (a) 、后 (b) 的网络结构

2.2.2 混合算法(ABC-GA)

在前文基础上,下面给出本文学习贝叶斯网络结构的混合算法(ABC-GA).

算法描述

输入:数据 D

输出:贝叶斯网络结构

1. 初始化

2. 设置雇佣蜂和观察蜂的数量 N,算法最大迭代次数 It,后代数量 DeN,交叉概率 p_c,变异概率 p_m.

3. 采用文献[14]中的方法,求解一无约束优化问题,以减少算法的搜索空间.在此基础上随机产生初始

食物源 $X_i(i=1,\cdots,N)$,计算每个食物源的适应度 $f(X_i)$.

4. 主循环

5. 当 $t<=It$

6. 通过使用选择、交叉和变异等算子产生 DeN 个后代.

7. 从 DeN 个后代中选出 N 个适应度高的食物源.

8. 对于 $i=1:N$ % 观察蜂阶段

9. 依据(3)选择食物源 X_i.

10. 使用变异算子产生新的食物源 X_i'.

11. 如果 $f(X_i')>f(X_i)$,则令 $X_i=X_i'$.

12. 对于 $i=1:N$ %侦查蜂阶段

13. 如果 $X_i=X_i-1$,则令 $c_i=c_i+1$;否则,令 $c_i=0$.

14. 如果 $c_i=$limit,则由步骤 3 得到的无向图基础上随机产生一个新的食物源 X_i.

15. 确定当前最优食物源 X_{best}.

16. 返回具有最高 BIC 分值的贝叶斯网络.

由于交叉和变异算子不是闭算子,所以步骤 6 和步骤 10 得到的不一定是有向无环图.如果不是,算法将调用两个子程序 repairbyMutualInfo(·)和 repairMaxparents(·),其作用分别为:当产生的新网络结构中含有环时,程序 repairbyMutualInfo(·)将计算环节点间的互信息,并删除环中具有最小互信息节点之间的

边，实现去环操作；当产生的新的有向无环图中某节点的父节点个数超过了最大限 u 时，repairMaxparents(·)程序会从父节点中选择出不超过最大上限的最好子集作为父节点集.

3　数值实验

为测试本文方法 ABC-GA 学习贝叶斯网络结构的性能，选取了 Asia 网络(图 4) 和 Car Trouble-Shooter 网络(图 5) 作为学习对象，从运行时间和汉明距离两个方面进行了比较，其中汉明距离＝多余边＋丢失边＋反向边，该指标可以反映算法的学习精度. 比较结果分别见表 1～表 4. 程序实现采用 Matlab 7.0，数值实验平台为 Pentium 4，3.06 GHz CPU，512 M 的微机，显著性水平 $\alpha=0.5\%$，交叉概率 $p_m=0.99$，变异概率 $P_c=0.001$，蜂群规模 $N=100$，后代 $DeN=200$，迭代次数 $It=15$，limit＝20.

3.1　Asia 网络

对于 Asia 网络，将本文的 ABC-GA 方法与文献[14]中的 CGA 方法做了比较，比较结果见表 1.

由表 1 可以看出，在同等学习精度下，本文提出的 ABC-GA 方法学习效率比文献[14]中的 CGA 方法要好. 在样本数量为 22 000，蜂群数量为 100 时，算法迭代 11 次，运行 178.773 203 s，ABC-GA 寻找到了真正最优的贝叶斯网络.

3.2　Car Trouble-Shooter 网络

对于 Car Trouble-Shooter 网络，将本文方法 ABC-GA 与文献[15]中的 U-ACO-B 方法做了比较，比较结果见表 3 和表 4.

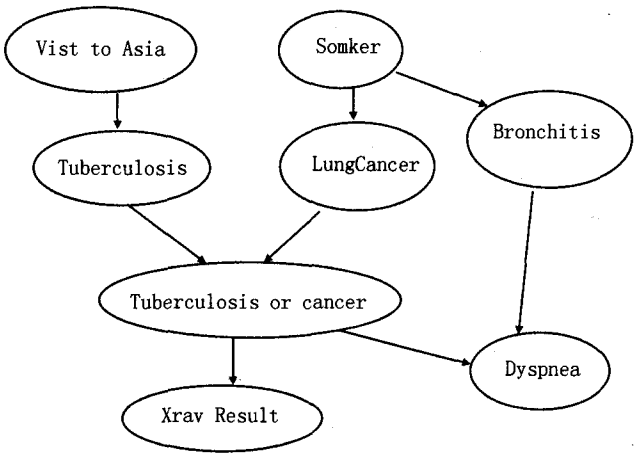

图 4　Asia网络

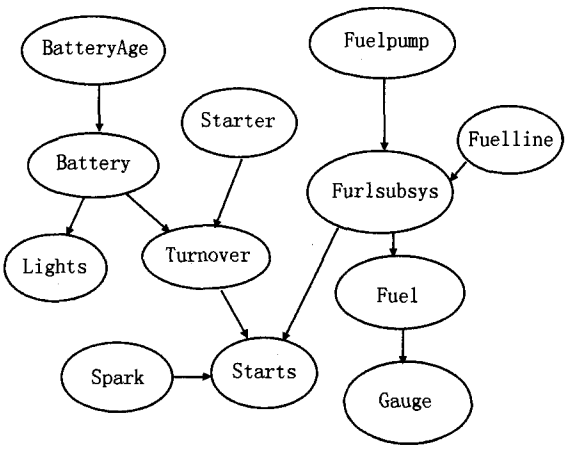

图 5 Car Trouble-Shooter网络

表 1　ABC-GA 和 CGA[14] 算法学习 Asia 网络的结果比较

样本数量	统计结果	算法	
		ABC-GA	CGA
200	运行时间/s	43.593	50.093
	汉明距离/边	2	2
500	运行时间/s	47	81.469
	汉明距离/边	2	2
800	运行时间/s	50.735	153.845
	汉明距离/边	1	1

表 2　ABC-GA 和 CGA[15] 算法学习 Car Trouble-Shooter 网络的结果比较

样本数量	统计结果	算法	
		ABC-GA	CGA
1000	运行时间/s	42.137 176	43.9220
	汉明距离/边	3	3
1500	运行时间/s	53.295 336	58.3590
	汉明距离/边	1	3
2000	运行时间/s	57.632 593	74.6300
	汉明距离/边	1	2

由表 2 的计算结果可以看出，本文方法 ABC-GA 无论从学习效率还是从学习精度方面都比文献[15]中的 U-ACO-B 方法要好. 在样本数量为 20 000，蜂群数量为 100 时，算法迭代 16 次，运行 407.333 940 s，ABC-GA 寻找到了真正最优的贝叶斯网络.

4 结 论

针对贝叶斯网络结构学习问题，本文基于蜂群算法和遗传算法提出了一种混合型算法(ABC-GA). ABC-GA 方法将蜂群和遗传两种算法有效结合，克服了每种算法所存在的缺陷. 数值实验表明，本文方法可以在小样本数据下，以较短时间获得较好的结果. 下一步工作将对算法中的参数设置，以及不同交叉算子的使用对算法的影响等问题展开研究.

参 考 文 献

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

[2] Spirtes P, Glymour C, Scheines R. An algorithm for fast recovery of sparse causal graphs[J]. Social Science Computer Review,1991,9:62-72.

[3] Cooper G F, Herskovits E. A Bayesian method for the induction of probabilistic networks from data[J]. Machine Learning,1992,9(4):309-347.

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

[5] Cheng J,Bell D, Liu W R. Learning Bayesian networks from data: an efficient approach based on information theory[J]. Artificial Intelligence,2002,137(1/2):43-90.

[6] Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search[M]. 2nd. Cambridge:The MIT Press,2000.

[7] Iannis T, Laura E B, Constantin F A. The max-min hill-climbing Bayesian network structure learning algorithm[J]. Machine Learning,2006,65(10):31-78.

[8] Chickering D M. Learning Bayesian networks is NP-complete[C]. Learning from Data: Artificial Intelligence and Statistics, Berlin: Spronger,1996.

[9] Liu D Y, Wang F, Liu Y N, et al. Research on learning Bayesian network structure based on genetic algorithms[J]. Journal of Computer Research and Development,2001,38(8):916-922.

[10] Pinto P C, Nagele A, Dejori M, RunKler T A, Sousa J M. Using a local discovery ant algorithm for Bayesian network structure learning [J]. IEEE Transactions on Evolutionary,2009,13(4):767-778.

[11] Wang T, Yang J. A heuristic method for learning Bayesian networks using discrete particle swarm optimization[J]. Knowledge and Information Systems,2010,24:269-281.

[12] Jensen F. BNs and Decision Graphs[M]. New York:Springer-Verlag,2001.

[13] Karaboga D. An idea based on honey bee swarm for numerical optimization[C]. Technical Report TR06,Erciyes University,2005.

[14] 汪春峰,张永红. 基于无约束优化和遗传算法的贝叶斯网络结构学习方法[J]. 控制与决策,2013,28(4):618-622.

[15] Wang C F, Liu S Y, Zhu M M. Bayesian network learning algorithm based on unconstrainedoptimization and ant colony optimization [J]. Journal of Systems Engineering and Electronics,2012,23(5):784-790.

A Hybrid Algorithm for Learning Bayesian Network Structure Based on Artificial Bee Colony and Genetic Algorithm

WANG Chunfeng[1], JIANG Yan[2,3]

(1. College of Mathematics and Information Science, Henan Normal University, Xinxiang 453007, China;
2. College of Environment and Planning, Henan University, Kaifeng 475001, China;
3. Foundation Department, Zhengzhou Tourism College, Zhengzhou 450009, China)

Abstract: The relationships between variables plays an important role in the interpretation of the data, and Bayesian network is an important tool to express the relationship between variables. For Bayesian network structure learning problem, a new hybrid algorithm (ABC-GA) is proposed based on ant colony algorithm (ABC) and genetic algorithm (GA). Because ABC-GA combines the strengths of ABC algorithm and GA algorithm, so it can make up the defects using either method alone. Numerical results show that: the calculation efficiency and accuracy of ABC-GA algorithm is high.

Keywords: Bayesian network structure learning; ABC algorithm; GA algorithm; unconstraint optimization