基于云遗传退火的贝叶斯网络结构学习算法

曹如胜 倪世宏 张 鹏
(空军工程大学航空航天工程学院 西安 710038)

摘 要 针对贝叶斯网络结构学习对算法高效性的要求,提出将云遗传算法和模拟退火算法相结合的云遗传模拟退火算法,以云遗传算法的选择、云交叉和云变异来完成模拟退火算法中的更新解操作;同时,针对算法在特定条件下陷入早熟收敛的问题,提出了改进的云交叉算子和云变异算子。仿真实验结果表明,所提云遗传模拟退火算法能有效提高贝叶斯网络学习的效率和准确性。

关键词 云模型,遗传算法,模拟退火,结构学习
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.09.045

Bayesian Networks Structure Learning Algorithm Based on Cloud Genetic Annealing

CAO Ru-sheng NI Shi-hong ZHANG Peng
(College of Aeronautics and Astronautics Engineering, Air Force Engineering University, Xi'an 710038, China)

Abstract In view of the highly active requirement of Bayesian networks structure learning, a learning strategy was proposed based on cloud genetic annealing algorithm which combines cloud genetic algorithm and simulated annealing algorithm. Update solution operation are accomplished by selection, cloud cross and cloud variation. In view of the shortcomings of algorithm being involved into the local optimization untimely, this paper put forward an adaptive cloud crossover operation and cloud mutation operator. The simulation shows that the accuracy of learning and operational efficiency are increased.

Keywords Cloud model, Genetic algorithm, Simulated annealing, Structure learning

1 引言

如何确定贝叶斯网络(Bayesian Network,BN)结构是BN学习的难题之一。BN结构学习就是在许多变量中找出它们之间的相互影响关系,从已获得的一定数量的数据中学习出各变量之间的因果关系。BN结构学习属于机器学习的范畴,利用BN模型由计算机自动寻找各个变量之间的依赖与影响关系,通过适当的算法对所给数据进行分析,得到各变量之间相互影响的关系图[1]。因此提高BN结构学习算法的效率是一个重要的研究内容。

目前,BN结构学习的研究热点之一是如何通过学习自动确定和优化网络的拓扑结构。Cooper等[2]基于爬山搜索算法和贝叶斯评分准则提出了K2算法,该算法得到了广泛引用,但是在结构学习前需已知BN节点的顺序。Chickering等[3]提出基于模拟退火算法(Simulated Annealing,SA)的BN结构学习,但随着温度降低,评分值较低的解被选择的概率逐渐降低,该算法容易陷入局部最优,且耗时较长。文献[4]提出基于云遗传算法的BN结构学习,通过借助云理论的随机性和稳定倾向性避免了搜索陷入局部极值并能很快地定位全局最值。文献[5]提出了一种自适应遗传退火算法(Genetic Algorithm and Simulated Annealing,GASA)用于BN结构学习,该算法结合了两种智能算法的优点,有效提高了BN结构学习的准确性和运行效率。

本文将云遗传算法(Cloud Genetic Algorithm,CGA)和模拟退火算法相结合,提出了一种自适应的云遗传模拟退火算法(CGASA)用于贝叶斯网络结构学习。CGASA算法以模拟退火算法为骨架,在每一次退火时,采用CGA算法的选择、云交叉和云变异来自适应地产生新解;同时,提出一种改进的云交叉、变异算子来提高算法的收敛速度和避免陷入局部最优的能力。以ASIA网络标准测试集的仿真实验为例,验证了本文算法的有效性。

2 基于CGASA的BN结构学习

遗传算法是一种借鉴自然界中的生物进化原理而产生的高度并行、自适应、随机的全局搜索算法[6]。它把生物进化过程中的自然选择、优胜劣汰、适者生存和遗传变异的思想应用到求解空间搜索最优解的问题上,是一种多参数、多群体的并行优化算法。但在实际应用中,其往往出现早熟收敛和收敛性能差等缺点[7]。云遗传算法根据适应度函数值的大小,通过云模型的X条件发生器自适应产生交叉和变异概率,可以

到稿日期:2016-08-21 返修日期:2016-12-31

曹如胜(1993—),男,硕士生,主要研究方向为数据处理与挖掘,E-mail:crsloveyss@163.com;倪世宏(1963—),男,教授,博士生导师,主要研究方向为飞行状态监控与地面数据处理;张 鹏(1982—),男,博士,讲师,主要研究方向为飞机故障诊断、故障预测与健康管理。

弥补传统遗传算法存在的早熟收敛和易陷入局部最优的缺点[8]。模拟退火算法的过程为:在某一初温下,伴随温度参数的不断下降,结合概率突跳特性在解空间中随机寻找目标函数的全局最优解[9]。将两种算法融合能够更有效地避免陷入局部最优,提高算法收敛性能和收敛速度。本文将两种算法结合,对云交叉算子和云变异算子进行了改进并用于贝叶斯网络的结构学习。

2.1 CAGSA 算法的关键操作

(1)编码

BN 网络结构由网络节点和节点间的指向边组成,通常用 0-1 矩阵对 BN 结构进行编码,n 个节点的 BN 用 $n\times n$ 的 0-1矩阵来表示,0 表示对应行列节点没有连接关系,1 表示该行对应节点是该列对应节点的父节点[10]。

(2)适应度函数

适应度函数可以体现种群中个体的优劣性,BN 结构学习通常采用贝叶斯网络评分函数作为适应度函数,网络结构评分越高,其适应度值就越高,网络结构就越优秀。本文选择 BIC 评分函数[11]作为适应度函数。

(3)选择操作

选择操作指从旧种群中以一定概率选择出优良个体组成新的种群,以繁殖得到下一代个体。个体被选中的概率与适应度值相关。本文将种群中的个体按适应度值的大小排序,个体适应值越高,其被选中的概率越大,采用轮盘赌法选择出下一代。

(4)改进的交叉、变异操作

交叉操作是指两个相互配对的个体按照某种方式相互交换各自的部分基因,从而形成两个新的个体。变异操作是将个体的染色体编码串中的某些基因座上的基因值用该基因座上的其他等位基因来替换,从而形成一个新个体的过程[12]。引入云理论后,利用云模型的云滴随机性和稳定倾向性的优点,由云模型的 X 条件云发生器根据适应度函数值的大小自适应产生交叉概率 P_{cr} 和变异概率 P_{mt}。但在父代个体适应度接近最大适应度值 $F_{\max}$ 时,交叉率和变异率虽然并非绝对为零,但无法避免在某次迭代时接近于零,因此减弱了交叉、变异的效果,导致算法陷入局部最优。对此,本文提出在父代个体适应度值为 $0.95F_{\max}$ 到 $F_{\max}$ 时,交叉概率取 $P_{cr}=t_3$,变异概率取 $P_{mt}=s_3$。改进后的 P_{cr} 和 P_{mt} 如下。

1)改进的云交叉算子

$$P_{cr}=\begin{cases} t_1 e^{\frac{-(f-E_x)^2}{2En'^2}}, & \overline{F}\leqslant f\leqslant 0.95F_{\max} \\ t_2, & f<\overline{F} \\ t_3, & f>0.95F_{\max} \end{cases} \tag{1}$$

其中,t_1,t_2,t_3 均为常数,$\overline{F}$ 为父代种群适应度均值,$F_{\max}$ 为父代个体适应度最大值。$f=\max(f_a,f_b)$,$E_x=(f_a+f_b)/2$,$En=m_1(F_{\max}-F_{\min})$,$He=n_1E_n$,$m_1$ 和 n_1 为控制系数。En' 是以 En 为期望、以 He 为标准差的一个正态随机数。

2)改进的云变异算子

$$P_{mt}=\begin{cases} s_1 e^{\frac{-(f-E_x)^2}{2En'^2}}, & \overline{F}\leqslant f\leqslant 0.95F_{\max} \\ s_2, & f<\overline{F} \\ s_3, & f>0.95F_{\max} \end{cases} \tag{2}$$

其中,s_1,s_2,s_3 均为常数,$E_x=f_a$,f_a 为单个父代个体的适应度值。

2.2 CGASA 算法的实现过程

(1)初始化:取初始温度 T_0 足够大,令 $T=T_0$,随机生成初始解,确定每个 T 时的迭代次数,即 Metropolis 链长 L;

(2)对当前温度 T 和 $k=1,2,\cdots,L$,重复步骤(3)—步骤(7);

(3)对当前解通过选择、云交叉和云变异生成新解;

(4)计算产生新解后的适应度函数的差值;

(5)若适应度函数差值小于零,则接受新解作为当前解,否则按 Metropolis 准则接受新解;

(6)如果达到迭代次数,则转到步骤(7),否则转到步骤(3);

(7)如果满足终止条件,则输出当前解为最优解,结束程序,否则按衰减函数衰减 T 后返回步骤(2)。

以上步骤结合了云遗传算法和模拟退火算法的思想,模拟退火中每次降低温度后,由云遗传算法的选择、云交叉和云变异来生成新解,改进的云交叉算子和云变异算子能更好地防止算法陷入早熟收敛,图 1 所示为算法的流程图。

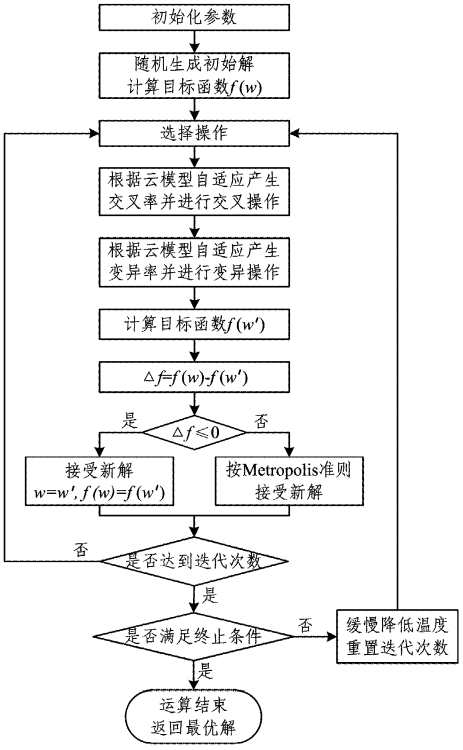


图 1 CGASA 算法结构学习流程图

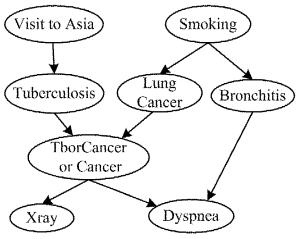
3 仿真与分析

选择贝叶斯网络标准测试集中的 ASIA 网络和 Car Trouble-Shooter 网络作为已知的标准网络。对于 ASIA 网络,分别选择 300,500,1000,5000 规模的样本用于学习;对于 Car Trouble-Shooter 网络,分别选择 100,500,800,1000 规模的样本用于学习。参数取 $t_1=t_2=1$,$s_1=s_2=0.6$,$t_3=s_3=0.2$。

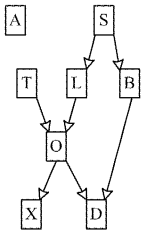
3.1 算法准确性和时间分析

以丢失边、反向边和增加边的个数为评价指标,对比分析在不同样本规模下模拟退火算法、遗传模拟退火算法以及本

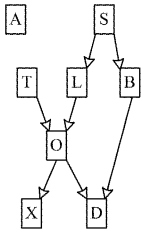
文提出的云遗传模拟退火算法学习得到的网络结构。ASIA 网络以样本规模 5000 为例，Car Trouble-Shooter 网络以样本规模 1000 为例，图 2 和图 3 示出了不同算法某次学习得到的网络结构。



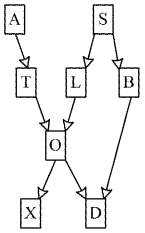
(a)标准 ASIA 网络结构



(b)SA 学习到的网络结构



(c)GASA 学习到的网络结构

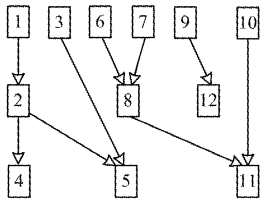


(d)CGASA 学习到的网络结构

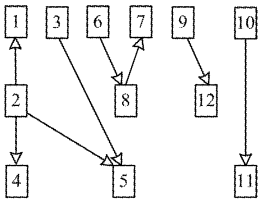
图 2　样本规模为 5000 时各算法学习到的 ASIA 网络结构

对于 Car Trouble-Shooter 网络，为方便网络节点的观察，将网络中的 12 个节点 BatteryAge，Battery，Starter，Lights，TurnsOver，FuelPump，FuelLine，FuelSubsys，Fuel，Spark，Starts，Gauge 分别用数字 1～12 来表示。

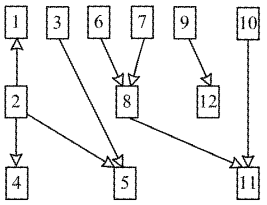
从图 2 和图 3 可以看出，在样本规模相同的情况下，本文提出的 CGASA 算法学习得到的网络结构与 ASIA 网络标准结构及 Car Trouble-Shooter 网络标准结构相同，而 SA 算法和 GASA 算法学习得到的 ASIA 网络结构与 ASIA 网络标准结构相比均存在某些边缺失。SA 算法学习得到的 Car Trouble-Shooter 网络结构相比于标准结构存在反向边和缺失边，GASA 算法学习得到的 Car Trouble-Shooter 网络结构存在反向边。因而本文所提出的 CGASA 算法在样本规模相同的情况下学习准确性更高。汉明距离是多余边、缺失边和反向边之和，可以直观地反映算法结构学习的准确性，表 1 列出了 3 种算法在不同数据规模下多次学习的汉明距离和时间的均值统计结果。



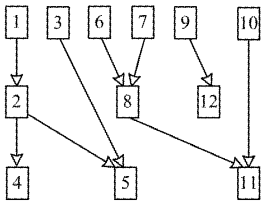
(a)标准 Car 网络结构



(b)SA 学习到的网络结构



(c)GASA 学习到的网络结构



(d)CGASA 学习到的网络结构

图 3　样本规模为 1000 时各算法学习到的 Car 网络结构

表 1　3 种算法学习结果的正确性对比

算法	搜索内容	不同样本规模的搜索结果(ASIA 网络)				平均值	不同样本规模的搜索结果(Car Trouble-Shooter 网络)				平均值
		300	500	1000	5000		100	500	800	1000	
SA	多余边数	1	1	2	0	1	0	0	0	0	0
	缺失边数	1	1	1	1	1	0	1	3	1	1.25
	反向边数	1	1	1	1	1	3	1	0	2	1.5
	汉明距离	3	3	4	2	3	3	2	3	3	2.75
	时间	39.61	42.41	38.52	78.14	49.67	87.97	44.52	41.34	41.33	53.79
GASA	多余边数	0	1	1	0	0.5	0	0	0	0	0
	缺失边数	1	1	1	0	0.75	0	0	1	1	0.5
	反向边数	1	1	1	1	1	2	1	1	0	1
	汉明距离	2	3	3	1	2.25	2	1	1	1	1
	时间	23.70	34.06	29.16	45.77	33.17	74.66	17.34	21.24	22.16	33.85
CGASA	多余边数	0	0	1	0	0.25	0	0	0	0	0
	缺失边数	1	1	1	0	0.75	0	0	0	0	0
	反向边数	0	0	0	0	0	1	1	1	0	0.75
	汉明距离	2	2	2	0	1.5	1	1	1	0	0.75
	时间	6.33	9.48	12.69	19.83	12.08	19.72	10.28	11.33	9.28	12.65

从表 1 可以看出，相同样本规模下，CGASA 算法学习的汉明距离小于 GASA 算法和 GA 算法。通过对比不同样本规模下学习的汉明距离和所需时间的平均值可以看出，本文所提的 CGASA 算法学习得到的网络较 GASA 算法和 SA 算法学习所得网络更接近真实标准网络，学习得到的网络准确性更高，所用时间也有较为明显的缩减。

3.2　算法效率分析

图 4—图 9 给出了相同样本规模下 3 种算法运行时的 BD 得分与标准网络 BD 得分的比值随着个体数的变化情况，选择样本规模为 5000，通过对比 3 种算法的 BD 比值收敛情况可以看出，本文提出的 CGASA 算法相比于其他两种算法收敛速度更快，学习效率更好。

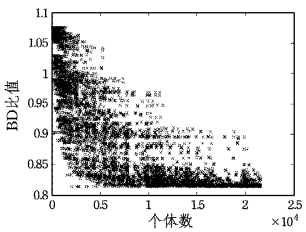


图 4　SA 算法收敛情况(ASIA 网络)

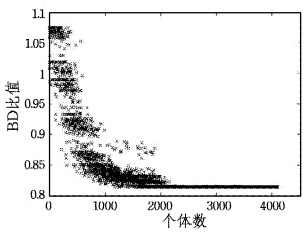


图 5　GASA 算法收敛情况(ASIA 网络)

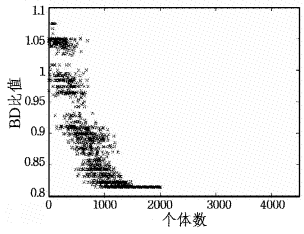


图 6 CGASA 算法收敛情况
(ASIA 网络)

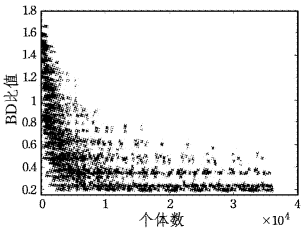


图 7 SA 算法收敛情况
(Car 网络)

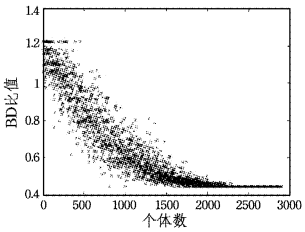


图 8 GASA 算法收敛情况
(Car 网络)

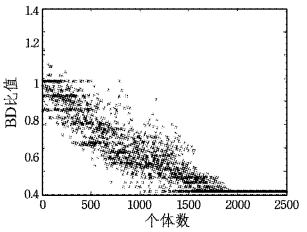


图 9 CGASA 算法收敛情况
(Car 网络)

3.3 改进的云交叉和变异算子分析

本次实验中,当没有对 CGASA 算法的云交叉算子和变异算子做出改进(即没有考虑父代个体适应度接近最大适应度的情况)时,交叉率和变异率取值存在接近于零的情况,CGASA 算法学习得到的结果中存在的最优结构为全零矩阵,即没有学习到网络节点间存在的因果关系,此时算法陷入早熟收敛。通过分析可以发现,交叉和变异操作的频度受交叉概率和变异概率的控制,父代中某些超常的个体在交叉变异中存在交叉率和变异率接近于零的情况,使得交叉变异操作效果减弱。为此,提出在 $\overline{F} \leqslant f \leqslant 0.95F_{\max}$ 时取 $P_{cr}=0.2$, $P_{mt}=0.2$,为了验证本文改进的云交叉和变异算子对算法学习效果的影响,以 ASIA 网络 4 种样本规模学习得到的网络结构的汉明距离均值为指标,P_{cr} 和 P_{mt} 取值范围为 0.1～0.4,当 P_{cr} 和 P_{mt} 取值过大时会减弱云交叉、变异算子的自适应性,因此取值范围只考虑 0.1～0.4,实验结果如表 2 所列。

表 2 改进的云交叉、变异算子的学习效果

汉明距离均值		P_{cr}			
		0.1	0.2	0.3	0.4
P_{mt}	0.1	2.0	1.9	2.1	2.1
	0.2	3.6	1.5	1.7	2.6
	0.3	2.4	3.1	3.2	3.5
	0.4	3.5	3.1	3.3	3.0

从表 2 可以看出,针对 ASIA 网络结构学习实验,云交叉、变异概率在取 $P_{cr}=0.2$, $P_{mt}=0.2$ 时,学习得到的网络的汉明距离均值最小,学习效果最佳;同理,针对 Car Trouble-Shooter 网络结构学习实验,云交叉、变异概率取 $P_{cr}=0.2$, $P_{mt}=0.3$。

结束语 本文将云遗传算法和模拟退火算法的思想相融合,并结合其各自优势提出了一种用于贝叶斯网络结构学习的云遗传模拟退火算法。在模拟退火算法的基础上加入云遗传算法的选择、云交叉和云变异作为更新解的操作,同时针对算法在应用中存在的早熟收敛情况,提出了改进的云交叉、变异算子。仿真结果表明,本文所提 CGASA 算法学习得到的网络结构的准确性更高,收敛更快,改进的云交叉、变异算子能有效避免算法的早熟收敛。由于算法中各参数的设置对算法的运行结果有较大影响,下一步将对算法中各参数的设置做进一步研究。

参 考 文 献

[1] 肖秦琨,高嵩. 贝叶斯网络在智能信息处理中的应用[M]. 北京:国防工业出版社,2012.

[2] COOPER G F, HERSKOVITS E. A Bayesian Method for the Induction of Probability Networks from Data[J]. Machine Learning, 1992, 9(4): 309-347.

[3] CHICKERING D, GEIGER D, HECKERMAN D. Learning Bayesian Networks: Search Methods and Experimental Results [C] // Proceedings of the 5th Conference on Artificial Intelligence and Statistics. 1995: 112-128

[4] QIN S, FENG L, WEI S, et al. Learning Bayesian Network Structure using a Cloud-based Adaptive Immune Genetic Algorithm[J]. Proc Spie, 2011, 8050(4): 80500s-80500s-10.

[5] LING A, XIAO B, ZHU Y, et al. Bayesian Network Structure Learning Based on an self-adaption Genetic and Simulated Annealing Algorithm[J]. Journal of Air Force Early Warning Academy, 2014, 28(2): 119-122. (in Chinese)
林傲,肖兵,朱艺,等. 一种用于 BN 结构学习的自适应遗传模拟退火算法[J]. 空军预警学院学报,2014,28(2):119-122.

[6] XIONG J, GAO D T, DU S D, et al. Genetic algorithm with mutation probability and population size adaptation[J]. Journal of Southeast University (Natural Science Edition), 2004, 34(4): 553-556. (in Chinese)
熊军,高敦堂,都思丹,等. 变异率和种群数目自适应遗传算法[J]. 东南大学学报(自然科学版),2004,34(4):553-556.

[7] 史峰,王辉,郁磊,等. MATLAB 智能算法 30 个案例分析[M]. 北京:北京航空航天大学出版社,2011.

[8] KANG L S, XIE Y, YOU S Y, et al. Non-numerical parallel algorithm simulated annealing algorithm [M]. Beijing: Science Press, 1998.

[9] DAI C H, ZHU Y F, CHEN W R. Adaptive genetic algorithm based on cloud theory[J]. Control Theory and Applications, 2007, 24(4): 1419-1423. (in Chinese)
戴朝华,朱云芳,陈维荣. 云自适应遗传算法[J]. 控制理论与应用,2007,24(4):1419-1423.

[10] WANG C F, ZHANG Y H. Bayesian network structure learning based on unconstrained optimization and genetic algorithm[J]. Control and Decision, 2013, 28(4): 618-222. (in Chinese)
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):618-222.

[11] LI S H, ZHANG J, SUN B L, et al. An incremental structure learning approach for bayesian network[C] // Proc of the 26th Chinese Control and Decision Conference. 2014: 4817-4822.

[12] 梁旭,黄明,宁涛,等. 现代智能优化混合算法及其应用[M]. 北京:电子工业出版社,2014.