第 38 卷第 5 期
2010 年 5 月

同济大学学报(自然科学版)
JOURNAL OF TONGJI UNIVERSITY(NATURAL SCIENCE)

Vol. 38 No. 5
May 2010

文章编号：0253-374X(2010)05-0744-05

DOI:10.3969/j.issn.0253-374x.2010.05.020

一种利用不可行解的贝叶斯网学习算

李小琳[1]，何湘东[2]，陈传明[1]

(1.南京大学 管理学院，江苏 南京 210093；2.南京大学 网络信息中心，江苏 南京 210093)

摘要：现有的基于打分搜索的贝叶斯网学习方法都是利用满足有向无环图的可行解进行学习.在搜索过程中遇到不可行解时，这类算法简单地去除不可行解或将不可行解转化为可行解.然而，有的不可行解中往往蕴含着有价值的信息.本文提出一种新的贝叶斯网学习方法 ISEC，同时利用可行解和不可行解学习贝叶斯网络，并提出针对不可行解的选择策略，在学习过程中可以有效地利用不可行解中的有用信息.实验结果表明，ISEC 能够比仅利用可行解的方法更快地学习到更优的贝叶斯网.

关键词：机器学习；贝叶斯网；结构学习；最小描述长度；进化计算

中图分类号：TP 301　　**文献标识码**：A

An Approach to Learning Bayesian Network by Using Infeasible Solutions

LI Xiaolin[1], *HE Xiangdong*[2], *CHEN Chuanming*[1]

(1. School of Management, Nanjing University, Nanjing 210093, China; 2. Network & Information Center, Nanjing University, Nanjing 210093, China)

Abstract: Existing Bayesian network learning approaches based on search and scoring usually work with feasible solutions which satisfy directed acyclic graph. This kind of approaches often removes infeasible solutions or converts infeasible solutions to feasible solutions when the solutions are infeasible. However, some infeasible solutions are much informative. This paper proposes the ISEC method for learning Bayesian network by using feasible and infeasible solutions synchronously based on an infeasible solution selection strategy. Then, the method can take advantage of the information in the infeasible solutions. Experiments show that the proposed approach can achieve better performance in less time than the approaches which use feasible solutions only.

Key words: machine learning; Bayesian network; structure learning; minimum description length; evolutionary computing

贝叶斯网[1]是联合概率分布的图形表示，它具有坚实的理论基础、形象直观的知识表示形式、灵活的推理能力和接近人类思维特征的决策机制，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之一[2].

基于数据建立贝叶斯网是近 10 年来贝叶斯网研究的主要内容之一，已相继产生了许多著名的算法[3-7]，有力推动了贝叶斯网理论和应用的研究进程.它在医疗诊断、软件智能化、金融风险分析、宏观经济决策、生物信息分析及 Internet 信息处理等方面得到广泛应用.目前贝叶斯网络结构学习算法有：基于评分搜索的方法[4]，基于依赖分析的方法[8]，以及把这两种方法结合起来的混合学习方法[9].基于评分搜索算法的基本思想是根据评分函数搜索得到对样本数据拟合得最好的贝叶斯网结构，此类算法一般来说效率较低，而且易于陷入局部最优解，当变量较多时较难实现.基于依赖分析的算法是通过分析样本中蕴含的依赖关系来构造网络结构，此类算法的时间复杂度相对较低，但对训练数据集依赖性大，如果训练数据集数据量不是充分大或者存在误差，结果通常是不可靠的.此外，很多基于依赖分析的算法需要知道结点的顺序，而在真实问题中这一条件往往很难满足.一般来说，基于评分搜索的贝叶斯网学习方法是在模型空间中进行搜索，利用某种评分函数对不同的模型进行评估，并利用评分结果指导下一轮的搜索.该过程反复进行，直到模型的评分收敛为止.通过学习产生的贝叶斯网一定要满足贝叶斯网的约束条件，也就是说必须是有向无环图.但在产生模型的过程中往往会出现一些不满足有向无环图约束条件的模型，即不可行解.对于这种情

收稿日期：2009-03-13

基金项目：国家自然科学基金资助项目(60803055)；教育部人文社会科学一般资助项目(08JC630041)；中国博士后科学基金资助项目(20080441031)；江苏省博士后科研资助计划(0801038C)；南京大学人才引进培养基金资助项目

作者简介：李小琳(1978—)，女，副教授，博士，主要研究方向为机器学习，数据挖掘，商务智能和决策分析. E-mail:lixl@nju.edu.cn

况,以往算法通常将不可行解去除或将其转化为可行解.然而,不可行解并非完全没有用,有的不可行解很可能包含了非常有用的信息,能够提供关于搜索最优解方面的更有用的信息.

本文提出一种可以有效利用搜索过程中遇到的不可行解的贝叶斯网学习算法 ISEC(infeasible solutions-based evolutionary computation).ISEC 利用竞争选择保留每代中部分优秀的不可行解,使搜索从可行解空间和不可行解空间两个方向搜索最优解,从而学习到最终的贝叶斯网结构.

1　贝叶斯网和相关工作

贝叶斯网[10]是一种有向无环图,图中每个结点代表一个向量.每个向量与一个条件概率表相关.关于一组变量 $X=\{x_1,x_2,\cdots,x_n\}$ 的贝叶斯网由两部分组成:一个表示 X 中变量条件独立关系的网络结构 S;与每一个变量相联系的局部条件概率表 P.

贝叶斯网是一种描述变量间依赖关系和条件独立性关系的图形和数量的表示方法.贝叶斯网 S 中的结点一一对应于 X 中的变量,结点之间没有弧线连接的表示结点之间是条件独立的.在给定结点的父结点集的情况下,图中的结点由变量及它们的条件概率表表示.根据条件独立性,联合概率分布可表示为

$$P(x_1,\cdots,x_n)=\prod_{i=1,\cdots,n}P(x_i\mid\pi(x_i)) \quad (1)$$

其中 $\pi(X_i)$ 是结点 X_i 的父结点集.

值得注意的是对同一个联合概率分布来说,贝叶斯网并不是唯一的.一个给定的联合概率分布能够表示成不同的网络拓扑结构,这主要依赖于给定的结点次序.相同的联合概率分布大约能有 $n!$ 种网络结构的表示方式.显然,父结点集就取决于结点次序和变量间的内在联系.

从数据中学习贝叶斯网有两个方面的任务:参数估计和结构学习.参数学习的目的是对一个已知结构的模型,从数据中学习出相关的参数.也就是说,算法的输入必须包括依赖结构和训练数据集;而结构学习不需要额外的输入,学习的目的则是要从训练数据集中学习出一个概率模型结构.显然,在贝叶斯网的学习中,结构学习是核心内容.

结构学习是一个富有挑战性的问题,其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.大多数基于评分搜索的结构学习算法主要关注三个方面:假设空间、评分函数和搜索算法.已经证明,找到具有最高评分的贝叶斯网结构是一个 NP 问题[11].因此,在实际的结构学习中,通常引入启发式搜索技术来寻找具有最高评分的结构.这类算法成功的关键在于潜在父亲结点集的选择.显然,错误的初值将会使最终的结构较差.

基于评分搜索的结构学习算法主要有贝叶斯评分方法[4,6]、基于熵的方法[12]、基于最小描述长度的方法[5,13]、基于最大互信息的方法[14]等.这些方法在进行模型选择时通常要求模型满足贝叶斯网学习的约束条件,也就是说待评估的模型必须是有向无环图.已有方法均未考虑在搜索过程中不可行解可能蕴含的有用信息,并将其利用到搜索过程中.

2　ISEC 方法

能够有效地评价不同的网络结构以找到和数据样本匹配程度最高的概率图模型,是学习贝叶斯网结构的关键问题之一.可以利用打分函数来选择网络结构,例如:MDL(minimum description length)标准[5,13,15].MDL 标准源于信息论中的交叉熵.用于贝叶斯网学习的 MDL 标准包括两个部分,即贝叶斯网结构的描述长度与数据的描述长度.它综合考虑网络结构的描述精度和网络结构的复杂性两个方面,试图找到一个既精确又简洁的网络结构.使用 MDL 标准,较好的网络结构应具有更小的分值.同其它评分函数一样,对于完备数据来说,MDL 准则是可以分解的.一个贝叶斯网模型的 MDL 评分是模型中每个属性 X_i 的父亲结点集 $\|\prod(X_i)\|$ MDL 评分的总和.由 MDL 标准为贝叶斯网 S 评分,可以表示为

$$M_{\mathrm{DL}}(S)=\sum_{X_i}MDL\left(X_i,\prod(X_i)\right) \quad (2)$$

根据 MDL 标准的可分解性,式(2)可以写成

$$M_{\mathrm{DL}}(S)=N\sum_{i=1}^{N}\sum_{X_i,\prod(X_i)}P\left(X_i,\prod(X_i)\right)\log P(X_i,\prod(X_i))-\sum_{i=1}^{N}\frac{\log N}{2}\left\|\prod(X_i)\right\|(\|X_i\|-1) \quad (3)$$

其中:N 是数据样本的大小;$\|X_i\|$ 表示 X_i 所有可能取值的个数;$\|\prod(X_i)\|$ 是结点 X_i 的所有可能父亲结点集取值的个数.

然后利用 MDL 标准对不同的模型进行评估,并利用评分结果指导下一轮的搜索.该过程反复进行,直到连续几轮搜索中模型的评分不再有明显提高为止.在此搜索过程中,往往会产生一些不满足约束条

件的不可行解,ISEC 方法将部分优秀的不可行解引入到搜索过程中,而不是将所有不可行解作为无价值的个体加以去除或通过删除产生环的边将不可行解转化为可行解. 如果能够有效地利用这些优秀不可行解中蕴含的信息,那么学习将被有效地促进.

直观地考虑,在搜索过程中,如果不可行解比可行解更靠近最优解,也就是说该不可行解能够提供有关最优解的更有用的信息,则说明不可行解比可行解更优秀,因此应该将此不可行解包括在群体中参与学习. 如图 1 所示,其中曲线范围内为可行解空间,小菱形表示不可行解,小圆形表示可行解.

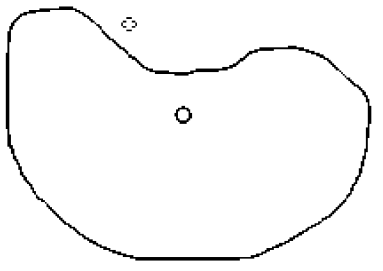

图 1 可行解与不可行解

Fig.1 Feasible solution and infeasible solution

从图 1 中可以看出全局最优解位于可行解空间的左侧顶点,因此小菱形表示的不可行解比小圆形表示的可行解更靠近最优解. 如果在学习中同时产生图中所示的可行解与不可行解,那么显然此不可行解比可行解更靠近最优解. 若能够将此不可行解包含在学习过程中,使搜索从可行解空间和不可行解空间两个方向搜索最优解,从而学习到最终的贝叶斯网结构,则可以有效地提高学习效率.

ISEC 利用进化计算作为搜索方法,并针对不可行解设计相应的选择策略,使不可行解参与学习,以此利用部分优秀不可行解中的信息帮助学习. 然而在进化过程中,直到进化过程终止前最优解是未知的. 因此,在进化过程中如何根据最优解判断保留哪些不可行解是算法的核心问题. ISEC 方法利用近似最优解来代替最优解对不可行解进行选取. 具体方法是在进化过程的各代中,将当前最优解作为近似最优解代替最优解进行比较,选择更接近近似最优解的不可行解作为群体中的个体参与进化.

首先,对任意两个个体 x_i,x_j,定义它们的规范化欧氏距离 D 为

$$D(x_i,x_j)=\sqrt{\frac{1}{n}\sum_{k=1}^{n}\left|\frac{x_i(k)-x_j(k)}{b_k-a_k}\right|^2} \quad (4)$$

其中:假设每个个体有 n 个分量;$x_i(k)$和 $x_j(k)$为其在个体 x_i,x_j 的第 k 个分量上的取值;a 和 b 为搜索空间的边界.

针对不可行解的竞争选择方法是:对于一个选定的不可行解,在与其他 q 个个体进行竞争时,如果它与当前代最优个体的规范化欧氏距离 D 小于与其竞争的个体,则对这个不可行解而言,其获得一次可能的获胜机会.

由于本文目的是验证进化过程中引入某些蕴含有价值信息的不可行解来促进优化的有效性,而并非是针对进化算法本身的探讨,因此,ISEC 方法采用相对比较简单的进化规划方法作为搜索算法. 算法中采用三种变异算子(增加边、删除边和转向边)产生后代,每次执行变异操作时,三种变异操作以相同的概率被选择. 如果变异操作产生的后代是可行解,则按照进化规划的竞争选择方法进行选择. 如果变异操作产生的后代是不可行解,则按照上面的竞争选择方法进行选择.

ISEC 方法为

(1) 产生种群规模 PS 个贝叶斯网络模型作为学习的初始群体,记作 $Pop(0)$;

(2) 使用 MDL 标准为 $Pop(0)$中的贝叶斯网络打分;

(3) 当 $t\leqslant G$,其中 G 为最大进化代数

①$Pop(t)$中的每个个体通过执行变异操作产生一个子代;②所有 $Pop(t)$中的个体及产生的所有子代个体组成一个临时群体 $\mathrm{Pop}'(t)$,个体数量为 $2\times PS$,$Pop'(t)$由可行解群体 $Pop'(t_F)$和不可行解群体 $Pop'(t_I)$构成;③对于 $Pop'(t_F)$中的每一个贝叶斯网络(可行解),设 S_i 为每个贝叶斯网的结构,从其它贝叶斯网中随机等概率地选取 q 个贝叶斯网与它进行比较. 设 S_{ij},$1\leqslant j\leqslant q$,为随机选出的贝叶斯网结构. 如果 $M_{\mathrm{DL}\,i}(S_i)\leqslant M_{\mathrm{DL}\,i}(S_{ij})$,$1\leqslant j\leqslant q$,则称该个体获得一次胜利. ④对于 $Pop'(t_I)$中的每一个不可行解,设 S_a 为每个不可行解的结构,从其他不可行解中随机等概率地选取 q 个不可行解与它进行比较. 设 S_{ab},$1\leqslant b\leqslant q$,为随机选出的不可行解结构. 如果$D_a(S_a)\leqslant D_a(S_{ab})$,$1\leqslant b\leqslant q$,则称该个体获得一次胜利. 显然,一个个体最多获得 q 次胜利. ⑤从 $Pop'(t)$中选择获胜次数最多的 PS 个个体作为下一代群体$Pop(t+1)$;⑥$t=t+1$.

(4) 当进化达到最大进化代数时,若最终群体中存在不可行解,即有环图,则删除产生环的边将其转化为可行解,并计算每个个体的 MDL 值. 如果存

在多种组合形式,随机选取一种形式删除;

(5) 返回各代中搜索到的具有最高打分值的贝叶斯网结构作为最终结果.

3　实验测试

实验选择两个问题域对 ISEC 进行测试. 首先,选定的贝叶斯网是含有 37 个结点的 ALARM 网. ALARM 网[16]是依据专家知识建立的医疗诊断贝叶斯网.

根据 ALARM 网的概率描述,抽样生成含有 500, 1 000,2 000,3 000,5 000 个事例的数据集,用 MDL 标准在这些数据集上为 ALARM 网打分的结果分别为 10, 598. 54, 18 602. 27, 34 265. 69, 49 732. 88, 81 219.74.算法分别对这 5 个数据集进行测试,每个数据集测试 10 次.进化计算中,由于适应度函数的计算复杂度,进化计算中种群规模不宜过大,而为了预防早熟收敛现象的发生种群规模又不宜过小.本文将种群规模 PS 设为 30,q 值设为 5.算法仅以数据集作为输入.表 2 所示为在这些测试数据集上分别学习 10 次得到的网络结构打分结果.

表 1　不同数据集上分别学习 ALARM 网 10 次的打分结果

Tab.1　Results of ten runs for ALARM network on different datasets

学习次数	评分结果				
	500	1 000	2 000	3 000	5 000
1	10 620.34	18 678.77	34 299.29	49 744.58	81 246.24
2	10 634.78	18 632.29	34 278.98	49 753.29	81 229.53
3	10 625.62	19 853.75	34 314.46	49 759.15	81 233.46
4	11 968.24	18 629.34	34 359.23	49 787.65	81 257.78
5	10 691.65	18 619.87	34 953.21	49 765.29	81 242.62
6	11 895.45	18 795.47	35 285.14	50 793.44	82 451.97
7	10 654.98	18 622.89	34 282.37	49 750.37	81 248.36
8	10 639.37	18 635.91	34 296.62	49 762.98	81 255.41
9	11 624.94	19 640.22	34 301.82	49 775.23	81 250.08
10	10 658.85	18 627.39	34 290.46	49 998.02	81 238.57

从表 1 可以看出,多数情况下 ISEC 能够学习到较好的网络结果,少数情况下 ISEC 陷入局部最优值,从而使学习到的结果较差.可以采取重开始策略或小生境的方法预防早熟收敛现象的发生.

然后在含有 5 000 个事例的数据集上对 ISEC 与使用遗传算法的学习方法进行比较.两种方法各学习 10 次,最后选择 10 次学习中的最优网络结构作为学习的最终结果.基于遗传算法的学习方法使用单点交叉操作和变异操作.在进化过程中若产生不可行解,则为不可行解赋予一个较差的适应度值.同样,群体规模为 30,最大进化代数 5 000.交叉概率 p_c 取值 0.9,变异概率 p_m 取值 0.01.如图 2 所示为应用 ISEC 和基于遗传算法的学习方法从具有5 000 个事例的数据集上学习到的网络结构的 MDL 打分结果比较.

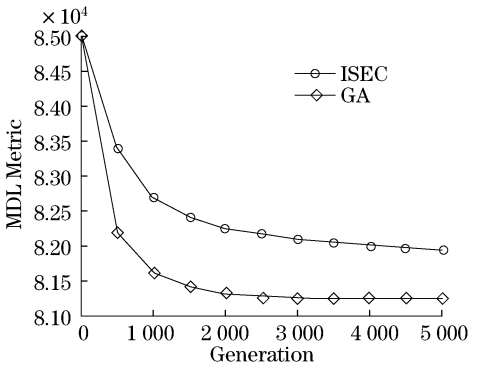

图 2　5 000 个事例学习到的 ALARM 网 MDL 打分结果比较

Fig.2　Comparison of the MDL scores on ALARM structures learned from 5 000 examples

从图 2 可以看出,ISEC 学习到的结果明显优于基于遗传算法的学习方法.另外,ISEC 方法得到最优网络的平均进化代数为 2 805.5,而用遗传算法的方法得到最优网络的平均进化代数为4 495.4.这并不难以理解,将部分优秀的不可行解引入学习进程,不仅能够有效地推动学习朝着有益的方向发展,还能够有效地提高学习的效率.因此,我们可以得出这样的结论:与应用遗传算法的方法比较,ISEC 能够在更短的时间里学习到更优的网络结构.

第二个问题域是 PRINTD 网络[17].这是一个含有 26 个结点的打印问题诊断贝叶斯网络.根据 PRINTD 网络结构及其概率描述生成含有 5 000 个事例的数据集,用 MDL 标准为此数据集打分的结果为 106 568.12.使用同 ALARM 网同样的方法进行测试.如图 3 所示为应用 ISEC 和基于遗传算法的学习方法从具有 5 000 个事例的数据集上学习到的 PRINTD 网络结构的 MDL 打分结果比较.

从图 3 可以看出,ISEC 学习到的结果仍然优于基于遗传算法的学习方法,并且能够在较少的代数里搜索到最优结构.

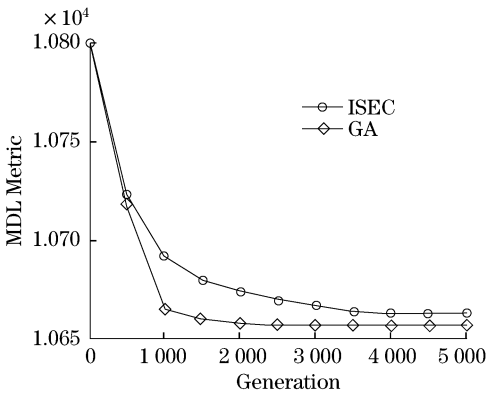

图3 5 000个事例学习到的PRINTD网MDL打分结果比较

Fig.3 Comparison of the MDL scores on PRINTD structures learned from 5 000 examples

4 结语

现有的基于打分搜索的贝叶斯网学习算法都是利用满足有向无环图的可行解进行学习,然而不可行解中可能蕴含着更有价值的信息.文中提出了一种同时利用可行解和不可行解学习贝叶斯网络的方法ISEC,使搜索从可行解空间和不可行解空间两个方向搜索最优解.实验结果显示与仅利用可行解的方法比较该算法能够在更短的时间内学习到更优的贝叶斯网络.

ISEC可能会陷入局部最优值,因此今后的一个工作是通过嵌入某些策略来预防早熟收敛现象的发生以提高算法的有效性.

参考文献:

[1] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference[M]. San Mateo: Morgan Kaufmann, 1988.

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

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

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

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

[6] Heckerman D, Geiger D, Chickering D M. Learning Bayesian networks: the combination of knowledge and statistical data[J]. Machine Learning, 1995, 20(3): 197.

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

[8] Spirtes P, Glymour C, Scheines R. Causation, prediction, and search [M]. 2nd ed. Cambridge: The MIT Press, 2000.

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

[10] 慕春棣,戴剑彬,叶俊.用于数据挖掘的贝叶斯网络[J].软件学报,2000,11(5):600.
MU Chundi, DAI Jianbin, YE Jun. Bayesian network for data mining[J]. Journal of Software, 2000, 11(5): 600.

[11] Chickering D M. Learning Bayesian networks is NP-complete [C]// Learning from Data: Artificial Intelligence and Statistics V. Berlin: Springer, 1996: 121-130.

[12] Herskovits E. Computer-based probabilistic network construction [D]. Stanford: Stanford University. Medical Information Sciences, 1991.

[13] SuzukiJ. Learning Bayesian belief networks based on the MDL principle: an efficient algorithm using the branch and bound technique[C] // Proceedings of the Thirteenth International Conference on Machine Learning. Bari: Morgan Kaufmann, 1996: 462-470.

[14] Wallace C, Korb K B, Dai H. Causal discovery via MML[C]// Proceedings of the Thirteenth International Conference on Machine Learning. Bari: Morgan Kaufmann, 1996: 516-524.

[15] Gao Q, Li M. The minimum description length principle and its application to online learning of handprinted characters[C]// Proceedings of the 11th International Joint Conference on Artificial Intelligence. Detroit: Morgan Kaufmann, 1989: 843-848.

[16] Beinlich IA, Suermondt H J, Chavez R M, Cooper G F. The ALARM monitoring system: a case study with two probabilistic inference techniques for belief networks[C] // Proceedings of the Second European Conference Artificial Intelligence in Medicine. Berlin: Springer, 1989: 247-256.

[17] Heckerman D, Wellman M. Bayesian networks [J]. Communication ACM, 1995, 38(8): 27.