doi:10.3969/j.issn.1673-4785.2011.01.009

利用互信息学习贝叶斯网络结构

李冰寒[1],高晓利[1],刘三阳[1],李战国[2]
(1.西安电子科技大学 数学系,陕西 西安 710071; 2.西安交通大学 机械工程学院,陕西 西安 710049)

摘 要:由数据构造贝叶斯网络结构是NP-难问题,因此提出了一种基于互信息的改进算法.该算法根据互信息构造初始框架,其次利用最大支撑树算法精简初始框架,并通过条件独立测试添加方向,最后利用贪婪算法得到最优网络结构.数值实验表明,改进算法无论是在BIC的得分值,还是在结构的误差上都有一定的改善,并且在迭代次数、运行时间上均有明显降低,能较快地确定出与数据匹配程度最高的网络结构.

关键词:贝叶斯网络;结构学习;互信息;独立测试;最大支撑树
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-4785(2011)01-0068-05

Learning Bayesian network structures based on mutual information

LI Binghan[1], GAO Xiaoli[1], LIU Sanyang[1], LI Zhanguo[2]
(1. Department of Mathematics, Xidian University, Xi'an 710071, China; 2. Department of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Constructing Bayesian network structures from data is an NP-hard problem, and an improved algorithm was proposed based on mutual information. This algorithm built the initial skeleton using mutual information, refined the initial skeleton by employing the maximum spanning tree algorithm, and then oriented edges according to conditional independence tests. Finally, the optimal network structure was obtained using a greedy search. Numerical experiments show that both the BIC score and structural error made some improvements from previous results, and the number of iterations and running time was greatly reduced. Therefore the structure with highest degree of data matching was shown to be relatively faster as determined by the improved algorithm.

Keywords: Bayesian network; structure learning; mutual information; independence test; maximum spanning tree

贝叶斯网络是表示随机变量间依赖和独立关系的网络模型,也称为贝叶斯网、因果概率网络或信念网络.它由节点集、有向边集和条件概率分布集组成,其中节点表示随机变量,是对过程、事件、状态等实体的某些特征的描述,有向边表示变量间的概率依赖关系.贝叶斯网络可对现实世界的各种状态或变量进行分析,生活中的许多实例如医疗诊断[1]、设备故障诊断[2]、故障预测[3]等都可以用贝叶斯网络进行建模.并且由于图形表示的直观性和易理解性,贝叶斯网络已逐渐成为在不确定情况下进行推理和决策的一种很受欢迎的问题表示结构[4-6].

学习贝叶斯网络可分为结构学习和参数学习,其中结构学习主要有2种方法:一是基于依赖性测试的方法[7-9],它是在给定数据下评估变量之间的条件独立性关系,构建网络结构;二是基于得分搜索的方法[9-14],其原理是在所有节点的结构空间内按照一定的搜索策略及得分准则构建贝叶斯网络结构.

利用互信息、最大支撑树算法和条件独立测试,本文提出了一种构建最优网络结构的改进算法.通过数值实验表明,与文献[15]中提到的贪婪算法相比,改进算法能较快地确定出与数据匹配最好的网络结构.

1 贝叶斯网络

定义1 贝叶斯网络[16] $N_B=(G,\theta)$,其中 $G=(V,E)$ 是一个非循环有向图,简称为DAG(directed acyclic graph),θ 是一个条件概率分布集,$\theta_i\in\theta$ 表

收稿日期:2010-07-13.
基金项目:国家自然科学基金资助项目(60974082).
通信作者:李冰寒. E-mail: binghan2020@126.com.

示在给定节点 X_i 的父节点时 X_i 的条件概率,节点集 V 的联合概率分布可由式(1)表示.

$$P(X_1 = a_1, X_2 = a_2, \cdots, X_n = a_n) = \prod_{i=1}^{n} P(X_i = a_i \mid \pi_i = F_{a_i}). \quad (1)$$

式中:π_i 表示节点 X_i 在 G 中的父节点集.

定义 2　DAG G 的 V-结构是指由 3 个节点构成的有序节点组(X,Y,Z),满足下面 2 个条件:

1)G 包含有向边 $X\rightarrow Y$ 和 $Z\rightarrow Y$;

2)在 G 中 X 和 Z 不相邻.

定义 3　随机变量 X 和 Y 的互信息定义如下[17]:

$$I(X;Y) = H(X) - H(X \mid Y).$$

式中:$H(X)$ 和 $H(Y)$ 分别是 X 和 Y 的信息熵,$H(X\mid Y)$ 是在给定 Y 下 X 的条件信息熵. $H(X)$、$H(X\mid Y)$ 定义如下:

$$H(X) = -\sum_{i=1}^{n} P(X_i)\log(P(X_i)),$$

$$H(X \mid Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(X = x_i, Y = y_i) \cdot \log(P(X = x_i \mid Y = y_i)).$$

式中:n 和 m 分别表示 X 和 Y 的状态个数. X 和 Y 之间的互信息是对称的,即 $I(X;Y) = I(Y;X)$.

2　构造贝叶斯网络的改进算法

算法的思想是从完全无向图开始,第 1 阶段删除互信息小于给定阈值 ε 节点间的边,构造初始无向图 G_1;第 2 阶段利用图论中的 Prim 算法寻找 G_1 的最大支撑树 G_2,精简初始无向图 G_1;第 3 阶段根据条件独立测试对 G_2 中的无向边添加方向,得到部分有向图 G_3;第 4 阶段对 G_3 中剩余的无向边任意添加方向,构造可能的非循环有向图 G_4;第 5 阶段从 G_4 开始,利用贪婪算法[15]在等价类空间上进行搜索,得到和数据匹配最好的贝叶斯网络结构的本质图 G^*. 具体算法可分为以下 5 个阶段.

2.1　构造初始无向图 G_1

由于随机变量 X 和 Y 之间的互信息度量了在已知 Y 的新观察值时对 X 取值的不确定性的减少,反之亦然,从而随机变量间的依赖程度可通过评价它们之间的互信息值来衡量. 如果两节点 X_i、X_j 之间的互信息小于某个临界值,即 $I(X_i;X_j) < \varepsilon$,则在真实网络中这两节点极大可能不相邻,因为反映它们之间依赖程度的互信息值小于阈值 ε,从而它们之间不存在边. 具体算法如下:

1)令 $V = \{X_1, X_2, \cdots, X_n\}$,$E_1 = \{X_i - X_j \mid i \neq j, i, j = 1,2,\cdots,n\}$;

2)计算 $I(X_i;X_j)$,$i \neq j, i, j = 1,2,\cdots,n$;

3)对 $\forall X_i, X_j \in V, i \neq j$,若 $I(X_i;X_j) < \varepsilon$,其中 ε 为阈值$(0 < \varepsilon < 1)$,则令 $E_1 = E_1 \setminus \{X_i - X_j\}$.

应注意的是,阈值 ε 的取值与 G_1 中无向边的个数有关. 如果 ε 过大(接近于 1),则很大可能会删除真实网络中实际存在的边;如果 ε 过小,则 G_1 中可能包含较多不存在于真实网络的冗余边. 在贝叶斯网络的结构学习中,通常取 ε 的值为 0.01,因为这样能保证真实网络中的大部分边存在于 G_1 中,而且 G_1 中含有较少的冗余边.

2.2　确定 G_1 的最大支撑树 G_2

本阶段将图中每条边上的权定义为节点间的互信息,然后利用 Prim 算法构造无向图 $G_1 = (V, E_1)$ 的最大支撑树. 由于第 1 阶段构造的初始无向图 $G_1 = (V, E_1)$ 可能包含多于三节点的环,而这些环的存在将会导致当前无向图不可一致延拓;但是在等价类空间上进行贪婪搜索的输入图应该是一个可一致延拓的本质图,故第 2 阶段通过构造最大支撑树来精简初始无向图,使得精简后的无向图 $G_2 = (V, E_2)$ 可一致延拓.

对于无向图 $G_1 = (V, E_1)$,设它的权矩阵为 $\boldsymbol{W} = (w_{i,j})$,$S$ 为 V 的一个非空子集,$\overline{S} = V \setminus S$,定义:

$$w_{i,j} = \begin{cases} I(X_i;X_j), & \text{若 } X_i - X_j \in E_1; \\ 0, & \text{其他.} \end{cases}$$

$$[S,\overline{S}] = \{X_i - X_j \in E_1 \mid X_i \in S, X_j \in \overline{S}\}.$$

构造 G_1 最大支撑树的具体步骤如下:

1)任取 $X_i \in V$,令 $S_0 = \{X_i\}$,$E_0 = \varnothing$,$k = 0$.

2)若 $S_k = V$,结束,以 S_k 为顶点集、E_k 为边集的图即是 G_1 的最大支撑树;否则转 3).

3)构造$[S_k,\overline{S}_k]$,若$[S_k,\overline{S}_k] = \varnothing$,则 G_1 不连通,算法停止;否则,设 $w(e_k) = \max\limits_{e \in [S_k,\overline{S}_k]} w(e)$,$e_k = X_k - X'_k$,$X_k \in S_k$,令 $S_{k+1} = S_k \cup \{X'_k\}$,$E_{k+1} = E_k \cup \{e_k\}$,$k = k + 1$,转 1).

2.3　确定最大支撑树 G_2 中边的方向得到 G_3

根据条件独立测试可以确定出形如 $X - Z - Y$ 和 $X \rightarrow Y - Z$ 的子结构中无向边的方向. 具体算法如下:

1)令 $G_3 = G_2$.

2)确定 G_3 中所有的子结构 $X - Z - Y$,其中 X 和 Y 不相邻.

3)对于每个子结构 $X - Z - Y$,若 $P(X \mid Z) \neq P(X \mid Y, Z)$ 且对于添加边 $X - Y$ 后的三角环 $X - Z - Y - X$,令 $S = \varnothing$. 对每个节点 $W \in V \setminus \{X, Y, Z\}$,计算 W 与其他各节点(不包含 Z)的互信息并降序排列,

设前3个与 W 的互信息最大的节点集为 T_W,若 $X\in T_W$ 且 $Y\in T_W$,则 $S=S\cup\{W\}$. 若 $S=\varnothing$,则令 $E_3=E_3\backslash\{X-Z,Y-Z\}\cup\{X\to Z,Y\to Z\}$.

4)确定 G_3 中所有的子结构 $X\to Y-Z$.

5)对于每个子结构 $X\to Y-Z$,若等式 $P(Z|Y)=P(Z|Y,X)$ 成立,则令 $E_4=E_4\backslash\{Y-Z\}\cup\{Y\to Z\}$;否则,对添加无向边 $X-Z$ 后的子框架 $X-Y-Z-X$,令 $S=\varnothing$. 对每个节点 $W\in V\backslash\{X,Y,Z\}$,计算 W 与其他各节点(不包含 Y)的互信息并降序排列,设前3个与 W 的互信息最大的节点集为 T_W,若 $X\in T_W$ 且 $Z\in T_W$,则令 $S=S\cup\{W\}$. 若 $S=\varnothing$,则令 $E_3=E_3\cup\{Z\to X\}$.

2.4 确定非循环有向图 G_4

由于第3阶段产生的图是非循环部分有向图,因此需要对尚未确定方向的无向边添加方向. 具体算法为:对 G_3 中尚未确定方向的无向边添加任意方向使之变为非循环有向图 G_4.

2.5 优化 DAG G_4 确定最优网络结构的本质图 G^*

由于第4阶段得到的 G_4 可能含有或缺少真实网络中的边,因此需要优化非循环有向图 G_4. 具体算法如下:

1)利用 Find-Compelled 算法[18]确定 G_4 的本质图 G_4^*;

2)从 G_4^* 开始,在等价类空间上进行贪婪搜索,得到最优结构的本质图 G^*.

3 复杂度分析

设 n 和 m 分别为节点个数和样本数,在第1阶段中需要计算 $n(n-1)/2$ 次互信息,而计算每个互信息的复杂度为 $O(m)$,故第1阶段的复杂度为 $O(mn^2)$;Prim 算法的复杂度为 $O(n^2)$,从而第2阶段的复杂度为 $O(n^2)$;在第3阶段中确定形如 $X-Z-Y$和 $X\to Y-Z$ 的子结构的复杂度为 $O(n^3)$,而最多有 n 个这样的子结构,对每个子结构进行条件独立测试的复杂度为 $O(m)$,从而第3阶段的复杂度为 $O(mn)+O(n^3)$;第4阶段的复杂度为 $O(|E|)$;算法 Find-Compelled 的复杂度为 $O(n^2)$,故第5阶段的复杂度为 $O(n^2)$. 由此可知,整个算法的复杂度为 $O(mn^2)+O(n^3)$,也就是说,改进算法的复杂度在最坏情况下是节点个数的多项式函数和样本个数的线性函数. 由于在实际生活中算法搜索时遇到的网络结构都是相当稀疏的,即使网络节点数很大,由于网络本身很稀疏,即具有较少的有向边,从而参与计算的节点很少,那些没有边相连的节点间的互信息很小,在第1阶段就已经被忽略掉了,从而后续阶段的计算复杂度相对来说比较小.

4 数值实验

下面以 Alarm 网络为例测试改进算法的性能,其中 Alarm 网络如图1所示,其本质图如图2所示,并将测试结果与原始贪婪算法[15]进行比较.

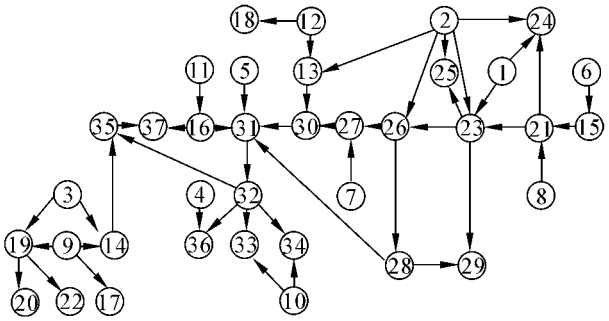


图1 Alarm 网络

Fig. 1 Alarm network

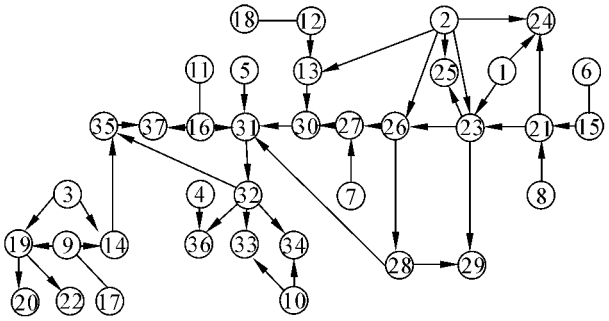


图2 Alarm 网络的本质图

Fig. 2 Essential graph of Alarm network

利用 Matlab 随机抽取4 000、6 000、8 000和10 000个样本作为实验数据,其中统计参数如下所示:

1)S_C:最优结构的 BIC 得分;

2)C_{ES}:与真实网络结构的本质图相比,算法确定出的真实边的个数;

3)M_{ES}:与真实网络结构的本质图相比,算法未确定出的边的个数;

4)W_{OS}:与真实网络结构的本质图相比,算法确定出的不具有方向或具有相反方向的边的个数;

5)W_{CS}:与真实网络结构的本质图相比,算法确定出的冗余边的个数;

6)G_{ES}:$M_{ES}+W_{OS}+W_{CS}$;

7)N_I:算法迭代次数.

表1给出了原始贪婪算法和本文提出的改进算法在不同样本容量数据集合上的实验结果. 由实验结果可以看出:与原始贪婪算法相比,改进算法确定的最优结构具有相对较少的缺失边和冗余边,迭代

次数明显减少,从而收敛速度更快,而且最优结构具有较高的 BIC 得分. 由此可知,改进算法能够获得更好的求解质量,并且算法的计算复杂度得到了明显的改进.

表 1　不同容量下 2 种算法的实验结果

Table 1　Experimental results of the two algorithms on different sample capacities

样本容量	算　法	$S_C/10^4$	C_{ES}	M_{ES}	W_{OS}	W_{CS}	G_{ES}	N_I
4 000	改进算法	-4.471 5	31	12	3	4	19	1
	贪婪算法	-5.905 3	31	12	3	4	19	12
6 000	改进算法	-6.489 5	32	10	3	3	16	1
	贪婪算法	-8.831 1	31	12	3	4	19	12
8 000	改进算法	-8.582 5	32	10	3	3	16	1
	贪婪算法	-11.737 0	31	12	3	4	19	12
10 000	改进算法	-10.729 0	32	10	3	3	16	1
	贪婪算法	-14.705 0	31	12	3	4	19	12

5　结束语

树是极为简单又极为重要的一类图,而最大支撑树是图论中的一类非常简单的网络最优化问题. 结合支撑树的性质,本文提出了一个新的构想:将网络结构中边上的权定义为节点间的互信息,从而有效地将图论和互信息知识结合起来. 改进算法与原始贪婪算法相比具有更好的求解质量,并且收敛速度有明显改进,对于提高大多数智能算法如蚁群算法、免疫算法的收敛速度有很大帮助,因为改进算法可用于对这些智能算法进行局部优化,并且对在等价类空间上搜索最优贝叶斯网络结构具有重要的研究意义.

参考文献:

[1] HANSEN J F. The clinical diagnosis of ischaemic heart disease due to coronary artery disease[J]. Danish Medical Bulletin, 1980, 27: 280-286.

[2] WOLBRECHT E, AMBROSIO B D, PASSCH B, et al. Monitoring and diagnosis of a multi-stage manufacturing process using Bayesian networks[J]. Artificial Intelligence for Engineering Design, Analysis & Manufacturing, 2000, 14(1): 53-67.

[3] BEISER J A, RIGDON S E. Bayes prediction for the number of failures of a repairable system[J]. IEEE Transactions on Reliability, 1997, 46(2): 320-326.

[4] PEARL J. Graphical models for probabilistic and causal reasoning[M]//TUCKER A B. The computer science and engineering handbook. Boca Raton, USA: CRC Press, 1997: 697-714.

[5] SCHACHTER R D. Probabilistic inference and influence diagrams[J]. Operations Research, 1988, 36(4): 589-605.

[6] MEEK C. Causal inference and causal explanation with background knowledge[C]//Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann, 1995: 403-410.

[7] BROMBERG F, MARGARITIS D, HONAVAR V. Efficient Markov network structure discovery using independence tests [J]. Journal of Artificial Intelligence Research, 2009, 35 (1): 449-485.

[8] MILAN S, JIRI V. A reconstruction algorithm for the essential graph[J]. International Journal of Approximate Reasoning, 2009, 50: 385-413.

[9] 冀俊忠,张鸿勋,胡仁兵,等. 一种基于独立性测试和蚁群优化的贝叶斯网学习算法[J]. 自动化学报, 2009, 35 (3): 281-288.
JI Junzhong, ZHANG Hongxun, HU Renbing, 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] PEDRO C P, ANDEAS N, MATHAUS D, et al. Using a local discovery ant algorithm for Bayesian network structure learning[J]. Transactions on Evolutionary Computation, 2009, 13(4): 767-779.

[11] CHEN Xuewen, GOPALAKRISHNA A, LIN Xiaotong. Improving Bayesian network structure learning with mutual information-based node ordering in the k2 algorithm[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20: 1-13.

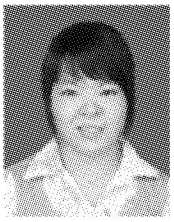
[12] LEVINE J, DUCATELLE F. Ant colony optimization and local search for bin packing and cutting stock problems [J]. Journal of the Operational Research Society, 2004, 55(7): 705-716.

[13] LOBONA B, AFIF M, FAIEZ G, et al. Imporving algorithms for structure learning in Bayesian networks using a new implicit score[J]. Expert System with Applixation, 2010, 37: 5470-5475.

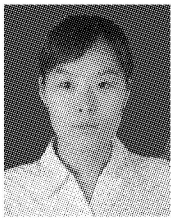
[14] TSAMARDINOS I, BROWN L E, ALIFERIS C F. The

max-min hill-climbing Bayesian network structure learning algorithm[J]. Machine Learming,2006,65(1):31-78.
[15] CHICKERING D M. Learning equivalence classes of Bayesian network structures [J]. Journal of Machine Learning Research,2002(2):445-498.
[16] RICHARD E N. Learning Bayesian networks[M]. Chicago,USA: Prentie Hall,2002:40-43.
[17] PROAKIS J. Digital communications [M]. Columbus, USA: McGraw-Hill,2000:6190.
[18] CHICKERING D M. A transformational characterization of equivalent Bayesian network structures[C]//Proceedings of the Eleventh Conference on Uncertainty in Arificial Intelligence. San Francisco,USA: Morgan Kaufmann,1995:87-98.

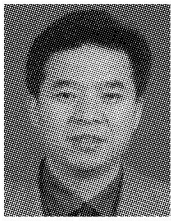
作者简介:



李冰寒,女,1986年生,硕士研究生,主要研究方向为数据挖掘、贝叶斯网络、最优化理论,发表学术论文5篇.



高晓利,女,1983年生,硕士研究生,主要研究方向为数据挖掘、贝叶斯网络、最优化理论,发表学术论文3篇.



刘三阳,男,1959年生,教授、博士生导师、博士,国家级教学名师,曾在法国作博士后研究,西安电子科技大学理学院院长兼数学系主任、工业与应用数学研究所所长.主要研究方向为应用数学、最优化和运筹学.先后主持10余项科研项目,获得多次省部级科技进步奖,国家级优秀教学成果二等奖2次,陕西省优秀教学成果特等奖、一等奖和二等奖多次.发表学术论文300余篇,其中被SCI检索50余篇,EI检索80余篇,出版专著6部.

第9届中国智能机器人学术研讨会征文通知

由中国人工智能学会智能机器人专业委员会主办、北京大学深圳研究生院承办的第9届中国智能机器人学术研讨会将于2011年11月11-13日在深圳举行,会议将组织赴香港和澳门的学术交流与参观活动,大会组委会热忱欢迎从事智能控制与智能机器人研究与应用的教师、学者、科技工作者和学生参加本次大会.现将大会征文范围和有关日期通知如下.

1.征文范围(但不限于)

·机器人理论与控制技术
·人工智能与智能控制技术
·智能机器人体系结构及实现方法
机器学习、算法与知识工程
基于网络的机器人结构与控制
·机器人传感技术、智能传感器
·多智能体系统理论与应用
多传感器集成与信息融合
·移动机器人及自主导航技术
·机器视觉、图像处理与模式识别技术
·机器人同时定位与地图创建(SLAM)
机器人结构设计及计算机仿真技术
·嵌入式计算机技术
·无线通信网络技术
服务机器人、特种机器人
足球机器人、仿生(人)机器人
·信息获取与数据挖掘
·智能机器人的应用

2.重要日期

投稿截止日期:2011年4月30日;录用通知日期:2011年5月31日;修改定稿日期:2011年6月15日.

3.稿件要求

论文必须未公开发表过,每篇论文的篇幅(含图、表)一般不超过6000字.录用论文在《华中科技大学学报(自然科学版)》(增刊)刊登.论文包括:题目、作者姓名、作者单位信息、中英文摘要、关键词、正文、参考文献、作者简介和E-mail地址.稿件要求附后或访问《华中科技大学学报》主页:http://xb.hust.edu.cn上的"投稿须知"栏目.

4.投稿方式

投稿方式:稿件用word电子文档发送到:liwei0828@mail.hust.edu.cn李炜副教授;
联系电话:027-87556242(办)、18971142368(移动);
投稿时请在稿件末尾的作者简介中注明作者的电话,以便联系.