

用遗传算法实现信度网结构学习:一种新的编码方案^{*}

张 聪^{1,2} 沈一栋³ 张 勤⁴ 刘启元⁵

(重庆大学计算机学院 重庆400044)¹ (重庆交通学院 重庆40007)²

(中国科学院软件研究所计算机科学重点实验室 北京100080)³

(重庆大学自动化学院 重庆400044)⁴ (重庆电子商务有限公司 重庆400012)⁵

摘 要 近年来,信度网已经成为表达一组随机变量间的概率关系的常见方法。在大的变量域中信度网的结构生成是信度网应用的难点。为了解决信度网的结构学习问题,一些研究者研究如何从数据集自动学习结构。本文研究采用遗传算法进行信度网结构学习时存在的问题和编码方法,并给出了一种新的信度网编码方案。数值试验显示遗传算法能够给出理想的结果。

关键词 信度网,结构学习,遗传算法

Structure Learning of Belief Network by Genetic Algorithms: A New Network Encoding Method

ZHANG Cong^{1,2} SHEN Yi-Dong³ ZHANG Qin⁴ LIU Qi-Yuan⁵

(Computer School, Chongqing University, Chongqing 400044)¹

(Computer Department, Chongqing Jiaotong University, Chongqing 400074)²

(Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100080)³

(Automation School, Chongqing University, Chongqing 400044)⁴ (www.cqsunway.com)⁵

Abstract In the last few years Belief networks have become a popular way of modeling probabilistic relationships among a set of variables for a given domain. It's a very hard task that creates a good belief network for large domains. Therefore, some researchers have studied how this construction can be automated. This work introduces how to do structure learning by genetic algorithms and discuss the encoding method using in GA. A new encoding method has presented. A case study has shown that GA is good for structure learning.

Keywords Belief network, Structure learning, Genetic algorithm

1 引言

近年来,信度网已经成为表达一组随机变量间的概率关系的常见方法,它利用了随机变量之间的条件独立性以高效地编码随机变量的联合概率分布。条件独立性的引入,使得分布的表达更加精简,便于知识的理解和建模,同时,降低了推理的计算复杂性。由于信对应,它表达了在 X 内部各变量间的条件独立性:

$X_i \perp \{X_1, X_2, \dots, X_{i-1}\} \setminus Pa_i | Pa_i$ (采用 Dawid 的表达方式),也即是,当且仅当一个变量 i 是另一个变量 j 的直接(概率)原因时,信度网中存在一条从节点 i 指向节点 j 的有向边。由于这一因果语义,信度网不存在有向环。

一组条件概率表 $P(X_i | Pa_i), i=1, \dots, n$ 。令 $x_i^k, k=1, \dots, r_i$ 表示节点 X_i 的所有可能取值, $pa_i^j, j=1, \dots, q_i$ 表示其父节点集合的所有可能取值,则条件概率表中的一个参数 $\theta_{i,k}$ 代表概率 $P(X_i=x_i^k | Pa_i=pa_i^j)$ 。

这两部分一道表达了关于变量组 X 的联合概率分布:

$$P(X) = \prod_{i=1}^n P(X_i | Pa_i).$$

信度网不仅有效地表达了一个联合概率分布,同时表达了各节点之间的因果关系。这使得采用信度网为基础的专家

系统,能够有效地利用专家知识(即有效地使用先验知识)。这主要在于因果关系同条件独立性之间有良好的对应关系。通常导致一个事件的直接原因均已得知的情况下,该事件的发生与否同其它原因无关(Markov 因果定义)。可以将该事件对应于图形结构 S 上的一个节点 X_i ,该事件的直接原因对应于该节点的父节点 Pa_i 。Markov 的因果定义可表达为:已知 Pa_i, X_i 发生与否同其它原因无关,即在已知 Pa_i 下 X_i 同其它原因条件独立: $X_i \perp \{X_1, X_2, \dots, X_{i-1}\} \setminus Pa_i | Pa_i$ 。因此 Bayes 网络可以有另外一种解释:Bayes 网络中的节点表示事件,节点间的有向边表达了事件间的直接因果关系,节点内部的条件概率表表达了该节点父节点对该节点的直接因果强度。

信度网的结构学习,就是从观察到的数据集,得出一个能够解释该数据集的信度网模型,或者说,挖掘数据之间最可能的因果联系,从而能够预测或解释新的数据。用 S 表示一个信度网结构, D 表示数据集,根据贝叶斯公式:

$$P(S|D) = P(S)P(D|S)/P(D) \quad (1)$$

式中的分母部分 $P(D)$ 与信度网结构的选择无关,因此我们简单地将其处理为归一化常数。而在没有任何先验知识的情况下,我们将每种结构出现的概率 $P(D)$ 处理为均匀分布:

$$P(S|D) \propto P(D|S) \quad (2)$$

因此,信度网的结构学习就是寻找一个信度网结构 S 使

^{*} 本文受国家自然科学基金(69883009)、教育部跨世纪优秀人才培养基金资助。张 聪 博士研究生,讲师。沈一栋 研究员,博士生导师。张勤 教授,博士生导师。刘启元 博士。

得概率 $P(D|S)$ 最大。在已知信度网 S (也就是已知一个联合概率分布) 的前提下, 得到 n 个数据的数据集 D 的概率 $P(D|$

$S) = \prod_{i=1}^n P(d_i|S)$, 其中 d_i 表示一个单独的观察数据。任意给定随机变量的一个顺序 $(X_1, X_2, \dots, X_n)$, 在参数独立和服从指数族先验分布这样的合理假设下^[5], 这一概率可以计算为:

$$P(D|S) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(\theta_{ij})}{\Gamma(\theta_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\theta_{ik} + N_{ik})}{\Gamma(\theta_{ik})} \quad (3)$$

其中 θ_{ij} 表示 X_i 的父节点的先验分布, N_{ij} 表示 X_i 的父节点观察值的计数, 同样, θ_{ik} 和 N_{ik} 分别表示 X_i 的先验条件概率和观察值的计数。在没有多余信息的条件下, 我们假设先验分布参数是均匀的。这样, 我们就得到了评价一个结构是否符合一个观察到的数据集的标准, 因此可以从备选的结构中选出最优的结构。但是, 信度网的结构空间是节点数的指数规模, 检验每一个可能的结构具有 NP 复杂度^[6], 因此需要采用某种搜索方法来寻找。“贪心搜索法”^[7]可以得到不错的结果。这一算法从一个没有边的结构开始, 每一步都增加一条使得式(3)的增量最大的边, 直到添加任何一条边都不再对式(3)有任何贡献为止。本文接下来将讨论如何采用遗传算法来进行信度网结构学习。

3 用标准遗传算法进行信度网结构学习

遗传算法(GA)^[8]出现以来, 作为一种先进的全局搜索算法, 在很多领域得到广泛的应用。遗传算法是一种启发式的随机搜索算法, 在多数问题上, 算法以概率1收敛于最优解^[9]。

3.1 信度网的二进制编码

采用遗传算法进行信度网结构学习, 首先要解决信度网的编码问题。信度网的有向无环图(DAG)结构, 一般的编码方式是邻接矩阵。关联矩阵由于涉及的边的数目可变, 不能在遗传算法中很好地采用。但是, 邻接矩阵要求对角线上的元素为0, 这就使得在变异操作时, 需要增加一个限制性检验, 从而增加了计算的时间。更严重的是, 在交叉操作时, 会产生大量的有向回路。我们在这里采用数值试验给出了节点数从2到60的任意 DAG 随机进行单点交叉操作后代中不出现有向回路的比例, 结果见图1。

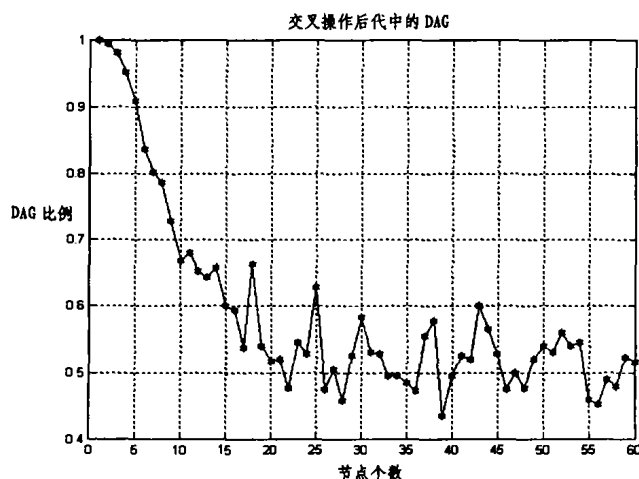


图1

图1的每一个数据点由1万对随机生成的 DAG 交叉后统计产生。从图中可以看出, 随着 DAG 节点数目的增长, 后代中 DAG 的比例很快趋近于40%~60%。对此结果进行数学上严格分析比较复杂, 本文暂不涉及, 但是上述试验表明采用

邻接矩阵编码对于信度网结构的遗传算法学习的效率确实有严重的影响。

文[10,11]采用了父节点描述这一自然的编码方案, 即用一个字符串 $c_{11}c_{21}\dots c_{n1}c_{12}c_{22}\dots c_{n2}\dots c_{1n}c_{2n}\dots c_{nn}$ 来编码一个信度网, 其中 $c_{ij}=1$ 表示节点 j 具有父节点 i 。这一编码方案是邻接矩阵的线性展开, 不能避免出现有向环后代。为了解决后代的有向环问题, 文[10]引入了一个“修复算子(repair operator)”, 对非法的信度网结构进行修复。文[11]则采用随机删除有向环的边来避免有向环的出现。

3.2 新的编码方案

为了避免后代中出现大量的非法结构, 我们提出了如下的编码方案:

设需要编码的 DAG 具有 n 个节点, 编码分为两个部分, 第一部分为整数 $k \in \{0, \dots, n!-1\}$ 的二进制编码共 $\lceil \log_2 n! \rceil$ 个 bits, 每一个数 k 表示图中 n 个节点的一种排列顺序。第二部分为父节点描述编码, 但是根据第一部分所决定的顺序, 排在后面的节点不能是排在前面的节点的父节点。第二部分的编码长度为 $\sum_{i=1}^{n-1} (n-i)$ 个 bits。

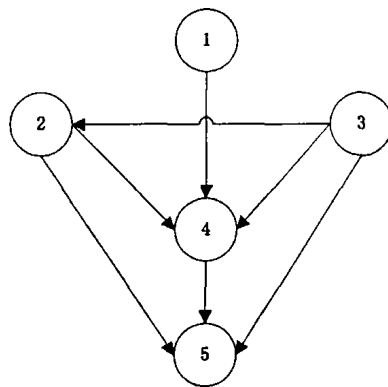


图2 信度网编码示意图

例如, 图2中的信度网编码为: 0000110 0010 111 11 1。编码的第一段是图中节点的顺序“1 3 2 4 5”的二进制编码。这一编码根据组合数学中关于依序枚举排列的序数法^[13]得到。第二段是节点1与后续节点的关系的描述, 第三段是节点3与后续节点的关系描述, 以此类推。这一编码方案很好地避免了邻接矩阵编码和父节点编码的缺点, 在个体上的任意的交叉和变异操作都是合法的, 不再需要定义修复算子。而且, 由于

$$\lceil \log_2 n! \rceil \leq \log_2 n! = n \log_2 n < \frac{n^2}{2}, \text{ 当 } n > 5,$$

我们的编码方案的编码长度也小于用邻接矩阵编码, 这对于提高计算效率具有十分明显的好处。

3.2 演化处理

余下的工作, 我们采用 Holland^[8]介绍的标准遗传算法进行, 选择算法采用“轮盘赌”算法, 交叉算法采用单点交叉, 以及简单的均匀概率变异。

在采用遗传算法进行信度网结构学习时, 可以很容易地加入对于结构的先验知识。例如, 我们已知某一变量可能不会受到别的变量的影响, 则可以在适应度计算的时候, 对于具有指向该变量的边的结构进行惩罚, 从而将搜索导向别的结构。

另外一种附加的指导性信息是采用近似判据 Bayesian information criterion(BIC)^[12]来计算适应度,

$$\log p(D|S) \approx \log p(D|\hat{\theta}, S) - \frac{d}{2} \log N \quad (4)$$

式(4)称为贝叶斯信息判据,其中, d 是结构参数 θ 的维度的线性函数。对于大的数据集,参数的估计值 $\hat{\theta}$ 可以采用参数对于数据的最大似然估计。BIC对于复杂的结构,给予一个较大的惩罚项,因此会指导遗传算法演化到结构比较简单的结构。在数据较少时,信度网和其他有指导的学习均存在过度拟合(overfitting)的现象^[14]。而BIC导出的结构,能够较好地避免过度拟合,从而较好地预测和解释新的数据。采用BIC判据,种群会较快地演化到简单结构上,从而能够提高计算的速度。在其他一些依赖计算较多的应用中^[14],信度网的结构学习也采用BIC判据。

4 数值试验

为了便于比较结果,我们采用了 Sewell 和 Shah^[15]在

表1 Sewell and Shah (1968)年的统计数据

4	349	13	64	9	207	33	72	12	126	38	54	10	67	49	43
2	232	27	84	7	201	64	95	12	115	93	92	17	79	119	59
8	166	47	91	6	120	74	110	17	92	148	100	6	42	198	73
4	48	39	57	5	47	123	90	9	41	224	65	8	17	414	54
5	454	9	44	5	312	14	47	8	216	20	35	13	96	28	24
11	285	29	61	19	236	47	88	12	164	62	85	15	113	72	50
7	163	36	72	13	193	75	90	12	174	91	100	20	81	142	77
6	50	36	58	5	70	110	76	12	48	230	81	13	49	360	98

利用表1的数据,用 matlab 进行编程,种群规模200,交叉概率0.3,变异概率0.01。由问题的性质,在适应度函数里,我们加入了如下两条先验知识:(1)SEX和SES没有父节点;(2)CP没有子节点。我们得到了和文[3]一致的结果,最佳的结构在图3中给出,在假设结构的先验概率为所有合法结构的均匀分布条件下,通过式(1)和(3)计算图2的后验概率 $P(S|D) \geq 1$ 。

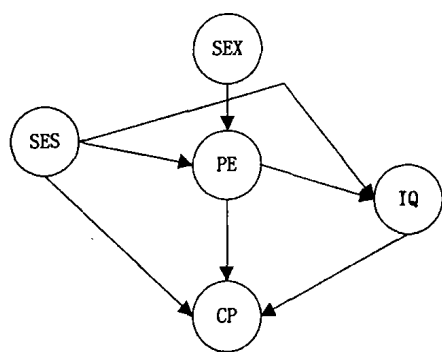


图3 最佳的结构

结论 遗传算法用于信度网结构学习,其主要的困难在于后代中如何避免有向环的问题。本文提出的编码方案较好地解决了这一问题,并且编码长度也比邻接矩阵编码更短,因此能够得到理想的结果。对于复杂的大型信度网学习,遗传算法的全局搜索性能具有比较明显的优势。但是,遗传算法在对式(3)或者其他的近似判据^[3]进行评估时,计算量较大,这多少影响了遗传算法在信度网结构学习中的使用。特别是,采用穷举搜索或者“贪心搜索法”时,能够利用信度网结构的局部变化,对结构评估进行增量计算^[3],使得小规模的信度网的启发式搜索或穷举搜索具有更大的可行性。

1968年收集的美国 Wisconsin 州10318名高中生是否计划上大学的数据来进行分析。这一数据集在文[3,4]中,以及其他一些统计学家的研究中曾经使用。文[3]采用了式(3)作为判据,对所有合法的结构进行了穷举搜索。数据集具有五个变量,其名称和取值分别为:性别(SEX):male, female;社会经济状态(SES):low, lower middle, upper middle, high;智商(IQ):low, lower middle, upper middle, high;父母鼓励(PE):low, high;以及上大学计划(CP):yes, no。表1中的数据的一项代表上述5个变量的一个特定的取值组合所观察到的个体数。第一项的组合是 SEX=male, SES=low, IQ=low, PE=low, CP=yes。其余的项分别是这5个变量轮流发生变化的各种组合,其中后面的变量(CP)先变化。例如,表的下半部分描述了 SEX=female 的数据。

参考文献

- Pearl J. From Bayesian Network To Causal Network. In: A. Gammerman, ed. *Bayesian Networks and Probabilistic Reasoning*, Alfred Walter Ltd., London, 1994
- Galles D, Pearl J. Axioms of Causal Relevance. In: *Proc. of the Fourth Intl. Conf. on Mathematics and AI*, Fort Lauderdale, FL, 1996. 64~67
- Heckerman D. Bayesian Networks for Data Mining. *Data Mining and Knowledge Discovery*, 1997(1): 79~119
- 慕春棣,戴剑彬,等. 用于数据挖掘的贝叶斯网络. *软件学报*, 2000, 11(5): 660~666
- Copper G, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, 1992, 9: 309~347
- Buntine W. Theory refinement in Bayesian networks. In: *Proc. 7th Conf. on Uncertainty in Artificial Intelligence*, Los Angeles, CA, 1991. 52~60
- Chickering D M, Geiger D, Heckerman D. Learning Bayesian networks is NP-hard. [Technical Report MSR-TR-94-17]. Microsoft Research, Advanced Technology Division, Microsoft Corporation, Redmond, WA, 1994
- Holland J H. *Adaptation in nature and artificial systems*. Arbor: University of Michigan, 1975
- 张文修,梁怡. *遗传算法的数学基础*. 西安交通大学出版社, 2000
- Larrañaga P, Poza M, Kuijpers C. Structure Learning of Bayesian Networks by Genetic Algorithms: A Performance Analysis of Control Parameters. *IEEE Trans. Pattern Anal. Machine Intell.*, 1996, 18(9): 912~926
- Etzeberria R, Larrañaga P, Picaza J M. Analysis of the behaviour of genetic algorithms when learning Bayesian network structure from data. *Pattern Recognition Letters*, 1997, 18: 1269~1273
- 杨振生. *组合学及其算法*. 中国科学技术大学出版社, 1997
- Schwarz G. Estimating the dimension of a model. *Annals of Statistics*, 1978, 6: 461~464
- Inza I, Larrañaga P, Etzeberria R, Sierra B. Feature Subset Selection by Bayesian network-based optimization. *Artificial Intelligence*, 2000, 123: 157~184
- Sewell W, Shah V. Social class, parental encouragement, and educational aspirations. *American Journal of Sociology*, 1968, 73: 559~572