包含隐变量的贝叶斯网络增量学习方法

田凤占[1],黄 丽[2],于 剑[1],黄厚宽[1]
(1. 北京交通大学计算机与信息技术学院,北京 100044;2. 河北理工大学理学院,河北唐山 063000)

摘 要: 提出了一种贝叶斯网络增量学习方法——ILBN. ILBN 将 EM 算法和遗传算法引入到了贝叶斯网络的增量学习过程中,用 EM 算法从不完整数据计算充分统计量的期望,用遗传算法进化贝叶斯网络的结构,在一定程度上缓解了确定性搜索算法的局部极值问题. 通过定义新变异算子和扩展传统的交叉算子,ILBN 能够增量学习包含隐变量的贝叶斯网络结构. 最后,ILBN 改进了 Friedman 等人的增量学习过程. 实验结果表明,ILBN 和 Friedman 等人的增量学习方法存储开销相当,但在相同条件下,学到的网络更精确;实验结果也证实了存在不完整数据和隐变量时,ILBN 的增量学习能力.

关键词: 贝叶斯网络; 增量学习; 遗传算法; 隐变量
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2005) 11-1925-04

An Incremental Approach to Learning Bayesian Networks Containing Hidden Variables

TIAN Feng-zhan, HUANG Li, YU Jian, HUANG Hou-kuan
(1. *School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044, *China*;
2. *School of Science, Hebei Polytechnic University, Tangshan, Hebei* 063000, *China*)

Abstract: An incremental approach to learning Bayesian networks based on genetic algorithm, namely ILBN, is put forward in this paper. ILBN introduces the EM algorithm and genetic algorithm into the incremental process of Bayesian network learning, calculates the expectation of the sufficient statistics with incomplete data using EM algorithm and evolves network structures using genetic algorithm, that could avoid getting into local maxima to some extent. Furthermore, by defining a new mutation operator and extending the traditional crossover operator, ILBN could incrementally learn and evolve Bayesian networks containing hidden variables. Finally, ILBN improves the incremental process by Friedman et al. The experimental results show that, in terms of storage cost, ILBN is comparable with the method by Friedman et al, while under the same experimental conditions, ILBN could learn more accurate networks than that of Friedman et al. The experimental results also verify the validity of ILBN in presence of incomplete data and hidden variables.

Key words: Bayesian networks; incremental learning; genetic algorithm; hidden variables

1 引言

近年来,贝叶斯网络的学习技术成为机器学习、数据挖掘、模式识别等领域的一个热点研究问题,其中贝叶斯网络的增量学习方法受到了众多研究学者的广泛关注. 增量学习方法不仅可以利用以前获得的学习结果,缩短学习时间,而且可以解决因数据集太大、无法全部存储在内存时所造成的学习困难,在实时和嵌入式系统中具有广泛的应用.

最早关于贝叶斯网络增量学习的研究是 1991 年 Buntine 提出的算法[1],但 Buntine 没有给出其算法的相关实验结果. 1994 年,Lam 等人对他们的贝叶斯网络学习方法进行了扩展,使其在观测到新数据时能够实现增量学习[2]. 不过,他们的算法要求当前被更新的贝叶斯网络结构已经比较精确,否则难以得到理想结果.

在贝叶斯网络的增量学习方面,最重要的贡献是 1997 年 Friedman 等人提出的增量学习方法[3]. 他们的方法存储了多个候选网络,能够在存储空间开销和学习网络的精度之间进行折衷,而且采用集束(beam)搜索,使得学习过程有更大的机会达到较高的局部最大值.

2002 年 Alcobé 提出了一种增量学习树形贝叶斯网络的算法[4]. 2004 年,他又提出了两种将批量爬山算法转换成增量爬山算法的启发式方法[5].

然而,上述贝叶斯网络增量学习算法主要基于完整数据,在不完整数据条件下增量学习贝叶斯网络仍然是一个难题,尤其是当网络结构中包含隐变量的时候. 本文研究了在不完整数据和隐变量条件下贝叶斯网络的增量学习问题,提出了一种贝叶斯网络增量学习方法——ILBN (Incremental Learning of Bayesian Networks). ILBN 将 EM 算法和遗传算法引入到贝叶斯网络的增量学习过程中,定义了新变异算子,扩展了传统的交叉算子,最后,ILBN 改进了 Friedman 等人的增量学习过程.

收稿日期:2004-12-07;修回日期:2005-04-15
基金项目:国家自然科学基金(No. 60303014);北京交通大学科技基金(No. 2004RC063)

2 贝叶斯网络的增量学习

增量学习贝叶斯网络的最直接方法是 MAP(Maximum A-posteriori Probability)方法,MAP 用最大后验网络模型作为后续计算的先验模型,迭代过程中只需存储观测到的新数据和当前的最大后验网络模型. 但是,MAP 学习过程会严重偏向于当前的最大后验网络,经过若干次迭代,最终结果将锁定在某个特定的网络模型,失去对新数据的适应能力.

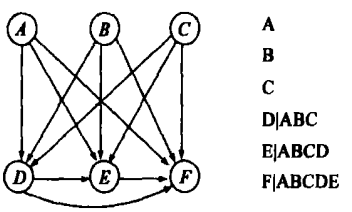
不用单个网络模型表示先验,Friedman 等人的增量学习算法将当前最有希望的网络结构组成一个候选网络集,并将该候选集定义为网络结构的搜索"边界",同时存储用于评价"边界"中的候选网络的充分统计量. 在观测到一条新数据后,算法更新充分统计量的取值,并且每观测到 k 条数据,检查"边界"中是否存在比当前网络模型评分更高的网络结构. 如果存在,则用该网络结构替换当前的网络模型. 然后,启动搜索过程确定下一个"边界",并对充分统计量集合进行相应的更新. 重复应用上述过程,直到算法结束.

当存在不完整数据时,由于无法精确计算网络结构所对应的充分统计量,Friedman 等人将 SEM(Structural EM)算法[6]扩展到其增量学习的框架中,用充分统计量的期望来评价候选网络. SEM 算法交替进行网络结构的贪婪搜索和网络参数的 EM 估计. 然而,当网络结构空间巨大且具有多峰值时,贪婪搜索方法会收敛到局部极值,因此,基于 SEM 算法的增量学习方法也容易陷入局部极值,不完整数据量越大,这种情况出现可能性也越大. 另外,Friedman 等人的增量学习算法没有考虑网络中存在隐变量时贝叶斯网络的增量学习.

3 贝叶斯网络增量学习方法——ILBN

3.1 学习包含隐变量的网络结构

为了从复杂的网络结构空间中搜索到好的网络模型,避免陷入局部极值,本文将遗传算法引入到贝叶斯网络结构的增量学习过程中. 一个贝叶斯网络可以分解成若干个局部结构,每个局部结构可以看作一个基因,父结点集可以看作是等位基因,整个网络结构可以表示成一条染色体,如图 1 所示. 可以采用 MDL 评价函数作为适应度函数来对网络结构进行评价[7],并且在完整数据集条件下,每个基因可以单独评分,并且所有基因的评分值之和就是整个网络结构的适应度. 当存在不完整数据时,本文借鉴 Friedman 等人的思想,用 EM 算法计算充分统计量的期望,并用充分统计量的期望来评价候选网络.



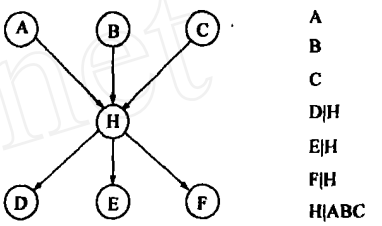
(a) 贝叶斯网络示例 (b) 基因表示

图 1 贝叶斯网络结构示例及其基因表示

对于网络结构染色体,本文使用均匀参数化交叉算子. 而基本的变异算子包括三种:其中两种是在等位基因中增加一个父亲结点或删除一个父亲结点,相当于在网络结构中增加或者删除一条指向变量的弧;第三种变异算子是逆转一条弧,这相当于删除一条从父亲到孩子的弧,添加一条从孩子到父亲的弧. 为了能够增量学习包含隐变量的网络结构,本文定义了新变异算子.

本文定义的变异算子通过在网络中添加新结点以及与之相连的弧,并删除掉那些在变量之间相互密集连接、关系错综复杂的弧,从而将若干个结点之间的相互依赖关系通过一个隐变量来表达,大大简化了网络结构,更易于用较少的数据学习到较精确的网络模型. 变异算子运算过程如下:当发现网络结构染色体的等位基因中多次出现公共结点集时,则在染色体中添加一个对应隐变量的新基因,新基因的等位基因是上述的公共结点集,而原来包含公共结点集的等位基因替换为隐变量. 实验表明,当公共结点集出现三次时,就可以应用该变异算子进行进化操作.

例如,图 2(a)中的网络结构可以通过应用本文的变异算子由图 1(a)中的网络结构演化得到. 具体运算过程如下:图 1(a)中的网络结构所对应的染色体如图 1(b)所示. 通过分析图 1(b)中的染色体可以发现,结点集{A, B, C}在等位基因中出现了三次,而等位基因中包含{A, B, C}的基因所对应的结点为 D、E、F. 因此,在染色体中添加一个对应隐变量 H 的新基因,它的等位基因是结点集{A, B, C},并且,将 D、E 和 F 的等位基因替换为隐变量 H. 形成的染色体如图 2(b)所示,该染色体所对应的网络结构如图 2(a)所示.



(a) 图 1 中网络引入一个隐变量后的贝叶斯网络结构 (b) 基因表示

图 2 包含隐变量的网络结构及其基因表示

虽然新变异算子可以发现包含隐变量的网络结构,但是它的引入也带来了新的问题. 因为基因的增加,使染色体的长度在应用该算子的过程中发生了变化,这给交叉算子的应用带来了困难. 因此,本文对传统交叉算子进行了扩展. 具体做法是:当进行交叉运算的两条染色体长度不相等时,给较短的染色体增加相应的虚位基因,使二者的长度相等. 所谓虚位基因,就是指该基因的等位基因为空,其对应的变量也不出现在其它基因的等位基因中. 增加虚位基因相当于在网络结构中添加孤立结点. 增加虚位基因后,两条染色体就可以进行传统的交叉运算了.

例如,图 3 中两条长度不同的染色体进行交叉运算,首先给较短的一条染色体增加虚位基因 F,结果如图 3 中的"父代"所示,进行交叉运算以后,结果如图 3 中的"子代"所示. 图中的双向箭头表示交叉运算发生的位置.

3.2 改进的增量学习流程

Friedman 等人的增量学习算法每读入 k 条数据,才向前搜索一步,当初始网络和初始"边界"与当前数据集蕴含的最优网络距离较远时,Friedman 等人的方法可能会在达到最优网络之前就停止了,无法保证学到当前数据集所蕴含的最优网络结构. 这种推测在 Bromberg 等人的课程实验[8]中得到了证实. 另外,新的充分统计量集合是在原来的充分统计量集合

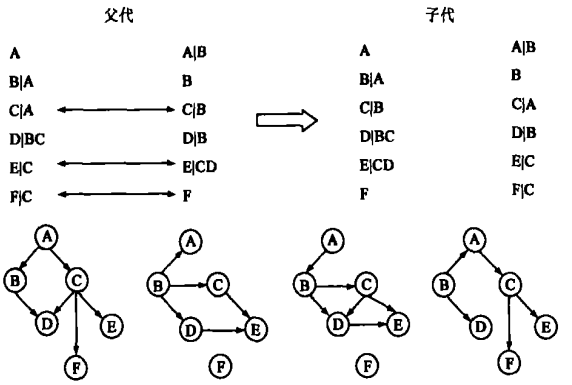


图 3　包含隐变量的贝叶斯网络的交叉运算

中增加或者删除一些充分统计量而得到的. Friedman 等人的增量学习算法利用读入的 k 条数据计算新增的充分统计量,这 k 条数据主导了新"边界"中所有网络的评价结果,从而引导了网络搜索进行的方向. 这样,当 k 值较大时,由于每读入 k 条数据,算法才向前搜索一步,因此算法的效率较低. 当 k 值较小时,虽然能够加快搜索进程,但是由于小数据集难以反映统计规律,因此算法的稳定性较差. 这与 Friedman 的实验结果相符,因此其算法在 k 值的选择上存在矛盾.

本文对 Friedman 等人的增量学习流程进行了改进,改进的流程如下所示:

令 F 为初始搜索边界;
令 G 为边界中的任意网络;
WHILE D 非空
{
读入 K 条数据;
REPEAT
令 $S = \bigcup_{G' \in F} Suff(G')$;
用 k 条数据更新 S 中的充分统计量信息;
令 $G = \arg\max_{G' \in Nets(S)} MDL(G' \mid S)$;
更新边界 F;
UNTIL 终止条件
}; /*WHILE 循环结束*/
利用 S 计算网络 G 的最优参数 Θ;
输出 $(G; \Theta)$.

在上面的过程中,$Suff(G) = \{N_{Xi, Pa(X_i)} \mid 1 \leq i \leq n\}$ 代表评价网络结构 G 所需的充分统计量集合;给定一个充分统计量集合 S,$Nets(S) = \{G \mid Suff(G) \subseteq S\}$ 表示可以用 S 中的充分统计量评价的网络结构集合;$MDL(G' \mid S)$ 是 MDL 评价函数[3];终止条件可以人为设定,一般的原则是搜索边界不再变化即可终止 REPEAT 循环.

3.3　ILBN 算法的时空复杂度分析

ILBN 算法的存储开销取决于为评价边界中的网络所需存储的充分统计量的数目. 在进化过程中父代群体与子代群体的网络模型之间大部分局部结构是相同的,可以用相同的充分统计量评价这些局部结构,只需补充很少的充分统计量来评价不相同的局部结构. 再考虑到评价不相同的局部结构时还会删除某些充分统计量,因此,当保存的充分统计量的规模达到一定程度时,所占的存储开销将趋于稳定,而不再进一步增长.

此外,ILBN 算法每读入 k 条数据进行一趟搜索,每趟搜索一直进行到发现当前数据集所蕴含的最优网络模型为止. 循环中的 k 值在可能的情况下越大越好,因为 k 值越大,数据越具有统计规律,使得搜索更能向着有利于找到最优网络结构的方向进行,搜索过程也更稳定. 另外,k 值越大,意味着 WHILE 循环的执行次数越少,从而算法的效率越高. 在理想情况下,当新数据集能够一次被读入内存时,上面流程中的 WHILE 循环只需要执行一次,算法的时间代价只由 REPEAT 循环决定,与对新数据集进行一次批量学习的时间代价相同. 这一特性可以作为何时启动增量学习算法的参考原则,即当观测到的数据量达到能够一次读入内存的临界点时启动增量学习算法效率最高.

4　实验分析

一方面,在完整数据条件下,本文就所需的存储开销对 ILBN、批学习算法以及 Friedman 等人的增量学习算法进行了实验比较. 从 ALARM 网[9]抽取出 10 个数据集,每个数据集包含 10,000 条数据. 实验结果是 ILBN 在 10 个数据集上运行结果的平均.

图 4 展示了在完整数据集上 ILBN 与批量学习方法和 Friedman 等人方法在空间代价方面的实验对比. 其中,Friedman 等人方法的结果来自于文献[3]中的实验,批量学习方法和 ILBN 的结果来自本文的实验. 实验中,只计算必需的数据存储,不包括任何附加存储,并就 k 取 1000 和 2000 分别运行了 ILBN. 由图 4 可知,批量学习方法需要越来越多的内存来存储所有的观测数据. 相反,ILBN 和 Friedman 等人方法在观测数据量较小时,随着新数据的不断加入内存开销逐渐增加,而当网络结构的估计接近稳定状态时,这两种方法的存储开销也趋于稳定. 在开始阶段,ILBN 比 Friedman 等人的方法需要占用更多的存储,到大约 6500 条观测数据时,两种方法达到了同样的存储开销.

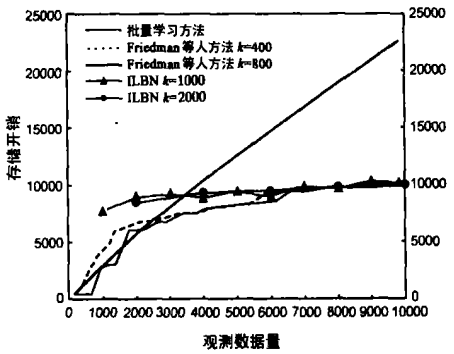


图 4　从完整数据集学习网络结构的存储开销

另一方面,本文也对存在不完整数据和隐变量时 ILBN 的学习能力进行了实验分析. 因为 Friedman 等人没有就这两方面进行实验,因此,本文将 ILBN 与批量学习方法进行了实验比较,实验中用标准化误差度量学到的模型与真实模型的接

近程度.对前面实验中使用的 10 个数据集,分别随机去掉 10 %和 20 %的数据值,从而得到两组不完整数据集:10 个含 10 %缺值的数据集和 10 个含 20 %缺值的数据集.利用这两组不完整数据集,对批量学习方法、ILBN k = 1000 和 ILBN k = 2000 进行实验对比,实验结果如图 5 所示.

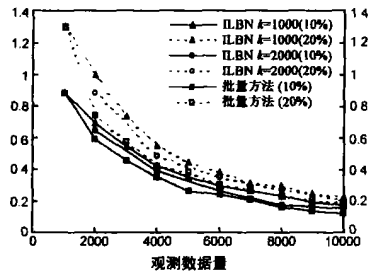


图 5 不完整数据条件下两种方法的标准化误差

从图 5 可以看出,从含 20 %缺值的数据集学到的网络的标准化误差明显高于从含 10 %缺值的数据集学到的网络的标准化误差,不过,随着观测数据量的增加,两种情况下的标准化误差都逐渐降低,并且二者的差距逐渐缩小,到 7000 条数据时二者的差距达到最小.此外,即便对于 20 %的缺值数据,到 10000 条数据时,ILBN 与批量学习方法学到的网络的标准化误差已非常接近.这表明 ILBN 在从不完整数据集学习时仍然具有较高的精度,而且也说明大数据集对于缺值带来的误差有很好的补偿作用.

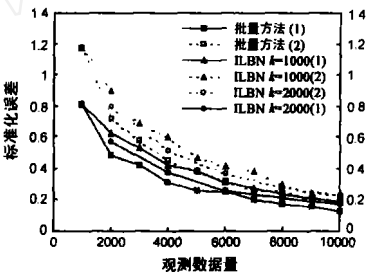


图 6 网络中包含隐变量时两种方法的标准化误差

对于最初的 10 个数据集,分别随机去掉其中的 1 个和 2 个变量的取值,则分别得到 10 个含 1 个和含 2 个隐变量的数据集.用这两组数据集对批量学习方法、ILBN k = 1000 和 ILBN k = 2000 进行实验分析,实验结果如图 6 所示.从图 6 可以发现,学习包含 2 个隐变量的网络的标准化误差明显大于学习包含 1 个隐变量的网络的标准化误差,不过,随着数据量的增加,两种情况下的标准化误差都逐渐降低,并且二者的差距逐渐缩小,到 9000 条数据时二者的差距达到最小,此时学到的包含 2 个隐变量的网络也相当精确.这说明 ILBN 也能够学习包含隐变量的贝叶斯网络,而且具有较高的精度.另外,从图 5 和图 6 都可以发现,设定较大的 k 值有利于学习到更精确的网络.

5 结语

本文提出了一种贝叶斯网络增量学习方法——ILBN.ILBN 将 EM 算法和遗传算法引入到了贝叶斯网络结构的增量学习过程中,用 EM 算法从不完整数据计算充分统计量的期望,用遗传算法进化贝叶斯网络的结构,实现了从不完整数据的增量学习.通过定义新变异算子和对传统交叉算子进行扩展,ILBN 也能学习和进化包含隐变量的贝叶斯网络结构.实验结果表明,ILBN 和 Friedman 等人的增量学习方法存储开销相当,但在相同条件下,ILBN 学到的网络更精确.而且在包含不完整数据和隐变量的情况下,实验结果也验证了 ILBN 的增量学习能力.

未来的研究工作之一是通过实验测试 ILBN 的时间性能;另一个工作是将 ILBN 应用到涉及大量数据的实际领域中,如数据挖掘、监测系统等.

参考文献:

[1] Buntine W. Theory refinement on Bayesian networks[A]. Ambrosio B D, Smets P Proceedings of the Seventh Annual Conference on Uncertainty in Artificial Intelligence[C]. Los Angeles: Morgan Kaufmann, 1991. 52 - 60.

[2] Lam W, Bacchus F. Using new data to refine a Bayesian network[A]. Ramon López de Mántaras, David Poole Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence[C]. San Mateo: Morgan Kaufmann, 1994. 383 - 390.

[3] Friedman N, Goldszmidt M. Sequential update of Bayesian network structure[A]. Dan Geiger, Prakash P. Shenoy. Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence[C]. Morgan Kaufmann, 1997. 165 - 174.

[4] J R Alcobé An incremental algorithm for tree-shaped bayesian network learning[A]. Frank van Harmelen Proceedings of the 15th European Conference on Artificial Intelligence[C]. Lyon: IOS Press, 2002. 350 - 354.

[5] J R Alcobé Incremental hill-climbing search applied to bayesian network structure learning[A]. Proceedings of the 15th European Conference on Machine Learning[C], Pisa, Italy, 2004.

[6] Friedman N. Learning belief networks in the presence of missing values and hidden variables[A]. Proceedings of the 14th International Conference on Machine Learning[C]. Madison, 1997. 452 - 459.

[7] Myers J W, Laskey K B, DeJong K A. Learning bayesian networks from incomplete data using evolutionary algorithms[A]. Banzhaf W, Daida J, Eiben A E et al. Proceedings of the Genetic and Evolutionary Computation Conference[C]. San Francisco: Morgan Kaufmann, 1999. 458 - 465.

[8] F Bromberg, B Patterson, S Yaramakala. Mining Bayesian Networks from Streamed Data[Z]. CS 561 Final Report, Spring 2003.

[9] Beinlich I, Suermondt G, Chavez R, et al. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks[A]. Proceedings of the Second European Conference on Artificial Intelligence in Medicine[C]. Berlin: Springer-Verlag, 1989. 247 - 256.

作者简介:

田凤占 男,1972 年 9 月生于河北武强,2002 年获清华大学计算机应用技术专业博士学位,2004 年从清华大学模式识别与智能系统博士后流动站出站,目前在北京交通大学计算机科学与工程研究所工作,主要研究领域为机器学习、数据挖掘、数据仓库、智能信息处理等.
E-mail: fztian@center.njtu.edu.cn.

黄 丽 女,1971 年生于新疆,1994 年毕业于华北电力大学计算应用专业,2001 年获得北京地质大学计算机系硕士学位,目前在河北理工大学理学院工作,主要研究领域为知识工程、数据挖掘等.