doi:10.3969/j.issn.1005-3026.2015.09.008

一种改进的多模块贝叶斯网络局部推理算法

赵建喆,李 凯
(东北大学 工商管理学院,辽宁 沈阳 110819)

摘 要:针对多模块贝叶斯网络的局部推理的时间和空间复杂度高的问题,提出了一种改进的多模块贝叶斯网络局部推理算法.该算法用面向对象语言重新定义了多模块贝叶斯网络模型,在联合树推理算法的基础上结合图论中"顶点度"的概念对局部推理算法进行了优化,针对三角化结果不唯一的问题,给出了一种一般性的解决方案,使三角化后的结果能够将消息传递得更快,有效地缩短推理时间.给出了算法的仿真实例并进行实验分析,结果表明改进后的推理算法有效减小时间、空间复杂度.

关 键 词:多模块贝叶斯网络(MSBN);局部推理;联合树算法;顶点度;三角化
中图分类号:TP 18 **文献标志码**:A **文章编号**:1005-3026(2015)09-1251-05

An Improved Local Inference Algorithm for Multiply Sectioned Bayesian Networks

ZHAO Jian-zhe, *LI Kai*
(School of Business Administration, Northeastern University, Shenyang 110819, China. Corresponding author: ZHAO Jian-zhe, E-mail: zhaojz@swc.neu.edu.cn)

Abstract: Due to the temporal and spatial complexity in the local inference of multiply sectioned Bayesian networks (MSBN), an improved algorithm for the local inference of MSBN was proposed. The algorithm redefined the model of MSBN with an object-oriented language. Combined with the concept of vertex degree in graph theory, the algorithm was optimized based on the joint tree algorithm. Considering that the outcome of triangulation was not single, the improved algorithm offered a general solution, which helped to convey message faster and greatly shorten inference time. Finally, an instance of the algorithm was given for experimental analysis, whose results showed that the improved inference algorithm significantly reduces both temporal and spatial complexity.

Key words: multiply sectioned Bayesian network; local inference; joint tree algorithm; vertex degree; triangulation

20世纪90年代以来,贝叶斯网络(Bayesian networks, BNs)成为人工智能、决策科学、机器学习、模式识别等众多分支学科中的理论研究热点.目前,随着应用领域的数据存储量增大,在大数据环境下对复杂的贝叶斯网络进行推理并提高时间效率对传统的贝叶斯网络提出了较大挑战.已证明传统的贝叶斯网络的推理是一个NP-hard问题[1].

Koller等科学家将模块化和面向对象的思想引入到贝叶斯网络中[2],提出多模块贝叶斯网络(multiply sectioned Bayesian networks, MSBNs).MSBNs是贝叶斯网络的一种拓展,将大型的复杂网络分解为几个子网分别进行建模,并且将子网抽象成类,使得子网具有良好的复用性与封装性,解决了复杂网络模型的构造问题.目前,MSBNs的研究将传统贝叶斯网络忽视的结构化信息(层次结构、数据类型结构)重新变成研究的重点,并利用结构化知识改进BNs,从而增强贝叶斯网络的解释性,降低贝叶斯网络推理的复杂度.MSBNs的推理可分为每个贝叶斯子网中的局部推理以及子网与子网之间的全局推理两个层次,

收稿日期:2014-08-25
基金项目:国家自然科学基金资助项目(61202085);教育部高等学校博士学科点专项科研基金资助项目(2012004,2120010).
作者简介:赵建喆(1982-),女,吉林白山人,东北大学博士研究生;李 凯(1957-),男,辽宁昌图人,东北大学教授,博士生导师.

整个推理可以通过有限次局部推理完成[3-4]. 局部推理中应用了许多经典贝叶斯网络的推理算法,精确推理算法与近似推理算法都可以应用到局部推理中.

许多学者对实际应用领域中的 MSBNs 建模研究,将一些相同或类似的子系统抽象成类. 郭文强等应用 MSBN 对农业车辆处于复杂农田环境的识别信度定量分析问题进行研究[5]. 此类研究需要对类进行具有实际意义的子网划分,这对问题本身有一定的要求. 2003 年清华大学的田凤占等[6]通过实验证明多模块贝叶斯网络推理算法的时空复杂度主要取决于各个子网的推理复杂度,并指出在模型建立初期就应该尽量简化子系统的内部结构,将整个系统划分为若干个松散耦合的稀疏网络的子系统.

联合树算法是通过步骤的共享来加快推理的一种算法[7]. 许多科学家为了使消息传播的速度更快,在联合树算法的基础上提出了基于传统推理算法的拓展算法,例如, Shafer - Shenoy and lazy propagation 算法[8],这些算法能够加快信息的收集和传播速度. 但是这些算法普遍存在对树形结构进行三角化的过程中三角化方案不唯一的问题[9]. 三角化的不唯一直接导致联合树树形结构的不同,最终导致消息传递的路径不同,会对推理的时间和空间复杂度造成较大影响.

针对这一问题,本文提出一种基于联合树算法改进的多模块贝叶斯网络局部推理算法,该算法结合图论中“顶点度”的概念,提出一种一般性的三角化方案. “顶点度”表示的是节点与其邻居节点之间关联边的个数,节点顶点度数越大表示其与周围节点关联越多,消息传递的范围越广. 因此,在三角化的时候将“顶点度”考虑在内,从而保证消息最大范围的传播,加快推理的进行.

1 面向对象的 MSBN 模型及改进的局部推理算法

1.1 面向对象的 MSBN 模型

MSBNs 提出了引入面向对象思想的贝叶斯网络,为搭建面向对象思想的贝叶斯网络模型提供了研究方向与框架,但是并没有通过面向对象编程语言对模型进行具体的定义. 本文提出一个面向对象的 MSBN 模型,用面向对象的编程语言对多模块贝叶斯网络的元素进行定义,并且确定了模型中的存储结构. 面向对象的 MSBN 模型,将贝叶斯网络划分成子网并将相同或类似的子网结构抽象成类,使其具有继承、封装、多态的特性,为以后的推理计算以及参数、结构学习提供了前提与基本框架.

定义 1 节点类(Node). 节点类是对贝叶斯网络节点的一个抽象,一个节点类对应一个节点,即随机变量将其属性进行封装. 节点类中包含属性:名字(Name),编号(Number),状态(State),顶点度(Degree)和条件概率参数(CPD). 节点类是一个 MSBN 子网中最小的结构单元,其数据结构如下:

```
Class Node{
attributes:
 - Name;
 - Number;//从 0 开始计数
 - State;
 - CPD;
 - Degree;//先默认为 0,在联合树算法中构建 Moral 图后再赋值
 ⋮
}
```

定义 2 边(Directededge[i][j]). 定义矩阵 $\boldsymbol{D}_1 = \begin{cases} 1, x_i, x_j \text{ 之间有关联边}, \\ 0, x_i, x_j \text{ 之间无关联边} \end{cases}$ 来表示节点之间的关系,即贝叶斯网络中连接节点的边,其中,x 为节点名字,i,j 为节点编号. 矩阵中每行代表节点的出度,每列代表节点的入度,出度为零的节点为叶子节点,没有子节点,入度为零的节点为根节点,其没有父节点. 但是,根节点不代表是信息输入节点,可能有许多根节点但是只有部分节点集合是信息输入节点;叶子节点也同理,并不是所有叶子节点都是信息输出节点. 除此之外,因为贝叶斯网络是一个有向无环图,所以对角线上的数值全部为零.

定义 3 子网类(SubNetwork). 子网类是模型中拥有相同或相似网络结构的贝叶斯网络的抽象. 子网中包含贝叶斯网络的基本元素:节点 Nodes 与边 Directededge[ ][ ],同时它还应该包含多模块贝叶斯网络特有的属性:消息传入节点 InputNodes、消息传出节点 OutputNodes 以及一般节点 NormalNodes. 根据不同的推理方法可以在子网类中定义不同的推理函数 inference( )进行不同的概率推理算法,其数据结构如下:

```
Class SubNetwork{
attributes:
 - Nodes;
 - Directededge[ ][ ];
```

－Name；
－InputNodes；
－OutputNodes；
－NormalNodes；
－JointTree
methods：
＋addNodes( )；
＋addEdges( )；
＋IsEmpty( )；
＋getFirstNeighbor(int *v*)；
＋getNextNeighbor(int *v*1 , int *v*2)；
＋findNode( )；//找到节点
＋findCircle( )；//找到回路
＋createMoralGraph( )；//构建 Moral 图
＋createTriangulatedGraph( )；//图形的三角化算法
＋createJointTree( )；//通过对三角化后的图形进行操作形成联合树结构
＋inference( )；//推理函数
⋮
}

定义 4　团结点类(Clique). 团结点类是联合树结构上团节点进行封装的类,包含多个节点 Nodes 以及条件概率在联合树上的映射函数 EnergyFunction,操作方法中最重要的有消息更新方法 update( ),该方法用于消息传递,使联合树达到一致状态,其数据结构如下：

Class Clique{
attributes：
－Nodes；
－Name；
－Number；
－EnergyFunction；
methods：
＋update( )；
⋮
}

定义 5　分割集类(Separator.) 分割集类是联合树算法中生成的一种数据类型,分割集类是对分割集的抽象. 分割集类包含多个节点 Nodes,其条件概率是联合树上的映射函数 Energy Function,操作方法中最重要的有消息更新方法 update( ),该方法用于消息传递,使联合树达到一致状态,其数据结构如下：

Class Separator{
attributes：
－Nodes；
－Name；
－Number；
－EnergyFunction；
methods：
＋update( )；
⋮
}

定义 6　联合树类(JointTree). 对分割集和团结点进行封装的类称为联合树类. 联合树类是联合树推理算法进行当中构建的一个新的数据类型,用以存储团结点 Cliques 与分割节点 Separators,联合树类可以对整个联合树进行消息更新等操作,其数据结构如下：

Class JointTree{
attributes：
－SubNetwork；
－Cliques；
－Separators；
－Name；
－Number；
－List < Object > list;//Clique 与 Separator 不是一个数据类型,用 List 存储边在二维数组中的位置
－TreeEdge[ ][ ];//用矩阵对边进行存储
methods：
＋update( )；
⋮
}

定义 7　多模块贝叶斯网络类(MSBN). 一个 MSBN 是对多个子网类的封装,一个贝叶斯网络与一个 MSBN 一一对应.

Class MSBN{
attributes：
－SubNetworks；
－Nodes；
－RootNode；
methods：
＋inference( )；
⋮
}

1.2　改进的局部推理算法

经典联合树算法是通过消息的共享来加快推理的进行,其难点在于使团结点之间消息能够更好地进行传递与分享. 当加入新的证据节点后,联合树结构破坏了其全局一致性,通过消息的传递

重新使联合树达到全局一致性,消息传递的过程就是贝叶斯网络的推理过程. 在构建 Moral 图之前将一个子节点的父节点两两相连,子节点与两个父节点构成汇连节点,当知道子节点的概率时,三个变量间是相互关联的,所以要将它们三个变量放在同一个团结点当中,以便于消息共享. 条件独立关系确定后,团结点内部的节点都相互关联能够起到很好的消息共享作用,使整个联合树的结构更加清晰. 在确立了条件独立关系后就要在图形转换阶段对构建的 Moral 图进行三角化操作. 将无向图中的一个封闭回路进行三角化的目的是为了使消息传递的更加稳定,不出现漏传或者重传的现象. 三角化的操作保证了消息传递的正确性与稳定性,但并不是所有的三角化结果都是唯一的[10],三角化不唯一的结果直接导致联合树树形结构的不同,最终导致消息传递的路径不同,所以选择一种合适的三角化方法能够更有效地提高算法的计算速度. "顶点度"表示的是节点与其邻居节点关联边的个数,节点顶点度数越大表示其与周围节点关联越多,消息传递的范围越广,所以在三角化的时候将"顶点度"考虑在内,保证消息最大范围的传播,从而加快推理的进行.

顶点度数大的节点消息传播的范围广,消息到达同一地点所花费的时间短. 虽然图形转换过程中会出现三角化结果不唯一的情况,但并不是所有的图形三角化的情况都不唯一,所以当最大回路小于 3 时就不存在三角化不唯一的问题.

因此,本文提出的这种结合"顶点度"的三角化方法是对 Moral 图一种改进的三角化操作方法,能够加快消息传播. 其具体操作过程是对 Moral 图顶点的一一消去过程,其具体步骤是:

1) 去掉不构成回路大于 3 的顶点,将构成回路大于 3 的环保存在一个集合中,如果有两个或两个以上的回路构成连通图那么将这个图进行分割,分别将这些回路存储在集合 H 中,然后再对回路集合 H 中的回路进行添加边的操作. 其算法伪代码如下:

输入:$G \leftarrow G_M = \{V, E\}$
输出:回路集合 H
BEGIN
while(G. hasNext( )){
$V_i \leftarrow G$. getNode( );
$C_i \leftarrow G$. findCircle(V_i);//找到所有回路
if(C_i. circle. road( ) > 3){//如果回路长度大于 3
circle($C_1, \cdots, C_n$);//将所得回路按照长度从小到大排序
C = null; H = null;//C 集合用来存储回路,用于后来的判断
while(circle( ). hasNext( )){
If($C \cup C_i \neq C$){//C 中回路的节点与 C_i 中的节点不同
C. add(C_i);
$H_i \leftarrow C_i$;
H. add(H_i);//将得到的回路保存在集合 H 中
}
}
if(G. leftNode( ) = C){
G. delete(C);
break;
}
}else{
G. delete(V_i);
}
}
END

2) 集合 H 是回路的集合,在 H 中选择一个回路,在回路中选择顶点度数最大的顶点,在该顶点的左右邻居间添加边,然后删除该顶点,最终直到该回路中没有顶点可删除则完成操作,得到的添加边集合 E 即为所求. 其算法伪代码如下:

输入:算法一中输出的回路集合 H
输出:三角化过程中添加的边集合 E
BEGIN
while (H. hasNext( )){
$H_i \leftarrow H$. getCircle( );
while(H_i. lefNode( ) > 3){
$V_i \leftarrow H_i$. getNode( );//找到顶点度数最大的节点
If(V_i. haveTwoNeighbours( )){
$M \leftarrow V_i$. getFirstNeighbour( );
$N \leftarrow V_i$. getNextNeighbour( );
H_i. addedge(M, N);//在 M 与 N 之间添加边
$ed_j \leftarrow$ edge(M, N);
E. add(ed_j);//获得添加边的集合
H_i. delete(V_i);
}
}
H. delete(H_i);
}
END

2　仿真实验与结果分析

本文通过一个实例对多模块贝叶斯网络推理与改进的多模块贝叶斯网络局部推理的复杂度进行比对与分析. 通过团结点个数 $f_{\text{cloud\_size}}$、最大团结点宽度 $f_{\text{max\_size}}$、最小团结点宽度 $f_{\text{min\_size}}$、割点集个数 $f_{\text{s\_size}}$、消息传播花费的时间 f_{cost} 这些判断复杂度的标准来对其空间、时间复杂度进行分析.

图 1 为一个多模块贝叶斯网络模型,首先将给出的 MSBN 模型进行联合树算法的图形转换操作,其次再将给出的 MSBN 模型进行改进的局部推理算法的图形转换操作,最后再对这两个算法构建的联合树树形结构进行分析.

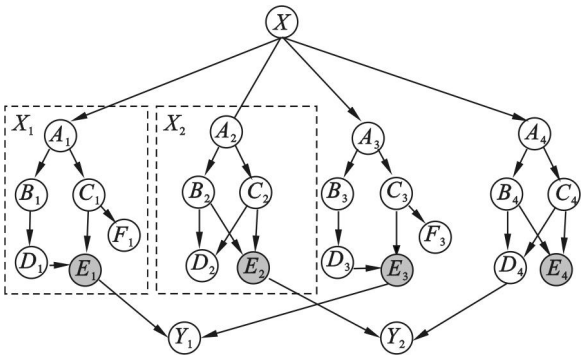

图 1　复杂网络实例

Fig. 1　Instance of complex network

MSBN 推理算法改进前后比对结果如表 1 所示,从 MSBN 中抽象出的两个子网类分别是 X_1 和 X_2,对子网类进行算法改进前后比对的结果如表 1.

表 1　MSBN 推理算法比对

Table 1　Comparison of MSBN inference algorithms

算 法	$f_{\text{cloud\_size}}$/个	$f_{\text{max\_size}}$/个	$f_{\text{min\_size}}$/个	$f_{\text{s\_size}}$/个	f_{cost}/个
原有算法	14	3	2	8	12
改进算法	14	3	2	8	8

从表 1 可知,经过图形转换后的联合树树形结构有 14 个团结点,43 个节点元素,而原来的贝叶斯网络结构只有 25 个节点元素,所以联合树算法的复杂度来源于团结点中节点元素的重复使用. 改进后的算法在没有增添空间复杂度的情况下将消息传播的花费从 12 个单位降低到 8 个单位,加快了信息传递的速度.

MSBN 中对类 X_1 实例化,因为在子网中存在三角化不唯一的问题,改进后的算法将消息传递到叶子团结点的花费从 8 个时间单位降低到 4 个时间单位. 而对类 X_2 实例化,因为子网中不存在三角化不唯一的问题,所以算法改进前后没有差别.

3　结　　论

本文提出一种改进的多模块贝叶斯网络局部推理算法是对精确推理算法联合树算法的优化,首先提出一个面向对象的 MSBN 模型,用面向对象的编程语言对多模块贝叶斯网络的元素进行定义并且确定了模型中的存储结构. 其次,引入顶点度概念对经典联合树算法的三角化问题给出了一个一般性的解决方案. 最后,通过举例对多模块贝叶斯网络推理算法改进前后进行了比较,得出的结论是,新的算法在空间复杂度不变的情况下加快了消息的传递.

参考文献:

[1] Haenni R, Romeijn J W, Wheeler G, et al. Probabilistic logics and probabilistic networks[M]. Berlin: Springer, 2010.

[2] Koller D, Pfeffer A. Object-oriented Bayesian networks[M]. London: Morgan Kaufmann Publishers Inc., 1997.

[3] Xiang Y, Jensen F V. Inference in multiply sectioned Bayesian networks with extended Shafer-Shenoy and lazy propagation [M]. London: Morgan Kaufmann Publishers Inc., 1999.

[4] Xiang Y. Belief updating in multiply sectioned Bayesian networks without repeated local propagations [J]. *International Journal of Approximate Reasoning*, 2000, 23(1): 1-21.

[5] 郭文强,高晓光,侯勇严,等. 采用 MSBN 多智能体协同推理的智能农业车辆环境识别[J]. 智能系统学报,2013(5): 453-458.
(Guo Wen-qiang, Gao Xiao-guang, Hou Yong-yan, et al. Environment recognition of intelligent agricultural vehicles based on MSBN and multi-agent coordinative inference [J]. *Journal of Intelligent Systems*, 2013(5): 453-458.)

[6] 田凤占,张宏伟,陆玉昌,等. 多模块贝叶斯网络中推理的简化[J]. 计算机研究与发展,2003,40(8): 1230-1237.
(Tian Feng-zhan, Zhang Hong-wei, Lu Yu-chang, et al. Multiple modules in Bayesian network inference of simplified [J]. *Journal of Computer Research and Development*, 2003, 40(8): 1230-1237.)

[7] Jensen F V, Lauritzen S L, Olesen K G. Bayesian updating in causal probabilistic networks by local computations [J]. *Computational Statistics Quarterly*, 1990, 4(1): 269-282.

[8] Madsen A L, Nilsson D. Solving influence diagrams using HUGIN, Shafer-Shenoy and lazy propagation [M]. London: Morgan Kaufmann Publishers Inc., 2001.

[9] Jensen F V, Jensen F. Optimal junction trees [M]. London: Morgan Kaufmann Publishers Inc., 1994.