

基于子网融合的贝叶斯网络结构学习算法

钟坤华, 陈芋文, 秦小林

引用本文

钟坤华, 陈芋文, 秦小林. [基于子网融合的贝叶斯网络结构学习算法](#)[J]. 计算机科学, 2022, 49(11A): 210800172-7.

ZHONG Kun-hua, CHEN Yu-wen, QIN Xiao-lin. [Sub-BN-Merge Based Bayesian Network Structure Learning Algorithm](#) [J]. Computer Science, 2022, 49(11A): 210800172-7.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[面向图像分类的小样本学习算法综述](#)

Survey on Few-shot Learning Algorithms for Image Classification

计算机科学, 2022, 49(5): 1-9. <https://doi.org/10.11896/jsjcx.210500128>

[基于贝叶斯攻击图的动态网络安全分析](#)

Dynamic Network Security Analysis Based on Bayesian Attack Graphs

计算机科学, 2022, 49(3): 62-69. <https://doi.org/10.11896/jsjcx.210800107>

[基于树增益朴素贝叶斯网络的服务定价策略](#)

TAN-based Service Pricing Strategy

计算机科学, 2021, 48(6A): 203. <https://doi.org/10.11896/jsjcx.200900024>

[高效计算因果网中的最大可能解释](#)

Efficient Computation of MPE in Causal Bayesian Networks

计算机科学, 2021, 48(4): 14-19. <https://doi.org/10.11896/jsjcx.200500155>

[基于数据挖掘的指定航班计划延误预测方法](#)

Prediction Method of Flight Delay in Designated Flight Plan Based on Data Mining

计算机科学, 2020, 47(11A): 464-470. <https://doi.org/10.11896/jsjcx.200600001>

基于子网融合的贝叶斯网络结构学习算法

钟坤华^{1,2,3} 陈芋文^{1,2,3} 秦小林^{1,3}

1 中国科学院成都计算机应用研究所 成都 610041
2 中国科学院重庆绿色智能技术研究院 重庆 400714
3 中国科学院大学 北京 100049
(zhongkunhua@cigit.ac.cn)

摘要 针对贝叶斯网络结构学习 K2 算法要求提供实际难以获得的准确先验节点顺序信息以及爬山算法对初始网络结构依赖性较强且容易陷入局部最优的问题,提出了一种基于子网融合的贝叶斯网络结构学习算法 Sub-BN-Merge。该算法首先为每个节点构造一个子网,并以 Voting 的方式融合生成每个节点的候选父节点集,然后基于评分函数在候选集中为每个节点搜索最优父节点集,最后消除所得网络结构中的环路,并以此为初值进一步采用启发式搜索方法对其进行优化。在小型网络 Asia、中型网络 Alarm 和大型网络 Win95pts 上进行了实验验证,同时分析了算法在数据存在缺失值情况下的性能。实验结果证明了算法的有效性,Sub-BN-Merge 算法在结构汉明距和算法正确率方面优于对比算法。

关键词: 贝叶斯网络;结构学习;子网融合;K2 算法;爬山算法

中图分类号 TP181

Sub-BN-Merge Based Bayesian Network Structure Learning Algorithm

ZHONG Kun-hua^{1,2,3}, CHEN Yu-wen^{1,2,3} and QIN Xiao-lin^{1,3}
1 Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, China
2 Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China
3 University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Aiming at the problem that Bayesian network structure learning method K2 requires to provide accurate prior node order information which is difficult to obtain in practice, and hill-climb algorithm is highly dependent on initial network structure and easy to fall into local optimal, a sub-BN merge based Bayesian network structure learning algorithm, (Sub-BN-Merge) is proposed in this paper. Firstly, for each node, the Sub-BN-Merge algorithm constructs a sub-net and merges them via voting to get a candidate parent set. Next, based on a scoring criterion, the algorithm searches the temporary optimal parent set of every nodes. Then, we breaks the cycle in the directed graph of the search result to get an directed acyclic graph. At last, further optimization is executed via a heuristic search method with directed acyclic graph as the initial value to get the final Bayesian network structure. Experiment is carried out on small network Asia, medium network Alarm and large network Win95pts. we also analyze the algorithm performance for the missing value situation. Experimental results show the effectiveness of the proposed algorithm. Our Sub-BN-Merge algorithm outperforms the comparison algorithms in term of structural Hamming distance and algorithm correct rate.

Keywords Bayesian network, Structure learning, Sub-net merge, K2 algorithm, Hill climb algorithm

1 引言

贝叶斯网络(Bayesian Networks, BN)起源于 20 世纪 80 年代中期对人工智能中不确定性问题的研究,已经成为人工智能的一个重要领域。贝叶斯网络是贝叶斯方法的扩展,是

人工智能、概率理论、图论和决策分析相结合的产物,是图形化概率模型的一种,是目前不确定知识和概率推理领域中最有效的理论模型之一,在临床决策^[1]、基因分析^[2]、金融分析^[3]及工业^[4]等领域应用广泛。

贝叶斯网络结构学习是贝叶斯网络的核心问题之一,是

基金项目:国家重点研发计划项目(2018YFC0116704);四川省科技计划(2019ZDZX0006, 2020YFG0010, 2020YFQ0056);中科院 STS 区域重点 A 类(KFJ-STIS-QYZD-2021-21-001);全国科学院联盟合作项目(中国科学院成都分院-重庆科学技术研究院)

This work was supported by the National Key Research & Development program of China(2018YFC0116704), Sichuan Science and Technology Program(2019ZDZX0006, 2020YFG0010, 2020YFQ0056), Science and Technology Service Network Initiative(KFJ-STIS-QYZD-2021-21-001) and National Academy of Science Alliance Collaborative Program(Chengdu Branch of Chinese Academy of Sciences-Chongqing Academy of Science and Technology).

通信作者:秦小林(qinxl2001@126.com)

贝叶斯网络参数学习和贝叶斯网络推理的基础。贝叶斯网络结构学习方法主要分为基于依赖分析的方法、基于评分搜索的方法以及混合方法。基于依赖分析的方法对条件独立性检验失败非常敏感;基于评分搜索的方法是一个 NP-Hard 问题,通常采用启发式搜索。

K2 算法^[5]和爬山算法^[6]均是经典的基于评分搜索的贝叶斯网络结构学习方法,具有很好的性能。然而,K2 算法需要事先提供实际中难以获得的全局节点顺序作为输入。爬山算法采用贪婪策略进行搜索,对初始网络结构依赖严重,容易陷入局部最优解。为了避免 K2 算法中的节点顺序要求,且为爬山算法提供一个较好的初始网络结构,本文提出一种基于子网融合的贝叶斯网络结构算法 Sub-BN-Merge,该算法结合机器学习中的 Voting 思想,以子网融合的方式构建节点的局部序,基于贪婪策略,借鉴 K2 算法为每个节点搜索最优父节点集的思路,构建初步的有向图网络结构,消除其中可能存在的环并以此作为初始网络结构,进一步采用启发式搜索方法对其进行优化,最终得到较为理想的贝叶斯网络结构。本质上,Sub-BN-Merge 是一种基于评分搜索的方法。最后,通过模拟数据集在小型网络 Asia、中型网络 Alarm 和大型网络 Win95pts 上验证了算法的性能,并从结构汉明距、算法正确率和计算效率 3 个方面做了分析。

2 贝叶斯网络及其结构学习

2.1 贝叶斯网络

贝叶斯网络也被称为信念网络(Belief Networks),是描述数据变量之间依赖关系的一种图形模型,也是一种用来进行推理的模型。BN 由两部分构成:网络结构和网络参数。BN 的结构是一个有向无环图(Directed Acyclic Graph, DAG),DAG 中的每个节点表示特定域中的一个变量,节点之间的有向弧表示相互之间的依赖关系,是对节点之间关系的定性刻画。每个节点都有一个与之相联系的条件概率分布函数,即网络参数,其定量地描述了变量节点与其父节点之间的依赖关系。若变量是离散的,则网络参数表现为给定父节点状态情况下该节点的条件概率表。BN 节点之间的依赖关系满足一阶马尔可夫性质。若网络中任意变量均为离散值,则称网络为离散 BN,本文研究所涉及的都是离散 BN。

记 BN 结构的有向无环图为 $G = \langle V, E \rangle$,其中,节点集合 $V = \{X_1, X_2, \dots, X_n\}$ 表示变量, E 表示有向边集合。每个节点 X_i 和它的父节点集合 $Pa(X_i)$ 都对应一个条件概率表,且满足:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | Pa(X_i))$$

2.2 贝叶斯网络结构学习

贝叶斯网络结构学习是指利用训练数据样本集,尽可能结合先验知识,获得最优的贝叶斯网络拓扑结构。从数据中学习获得贝叶斯网络的过程是机器学习的一个特例。贝叶斯网络结构近似学习方法主要分为基于依赖分析的方法、基于评分搜索的方法以及混合方法。

基于依赖分析的方法将贝叶斯网络看作表示独立变量关系的网络模型,通常利用统计或信息论的方法,通过计算节点

之间的互信息和条件独立性来分析变量之间的依赖关系,最终寻求一种符合这种关系的网络结构。典型的基于依赖分析的方法有 SGS 算法^[7]和 PC 算法^[6]。这类方法较为直观有效,但对条件独立性检验中的失败非常敏感。

基于评分搜索的方法将网络结构学习视为最优化问题,学习目标是搜索出评分最高的网络结构。随着节点的增多,待搜索的网络结构空间呈指数级增长,这是一个 NP-Hard 问题^[8],通常采用启发式搜索方法获得最终的网络结构。常用的评分函数包括 BDeu 评分、K2 评分^[5]、贝叶斯信息准则 BIC 评分^[6]和最短信息准则 MDL 评分。搜索空间包括:DAGs 空间、节点序空间和等价类空间。

基于 DAGs 空间搜索的方法直接对贝叶斯网络的拓扑结构进行搜索,通常采用局部搜索策略,容易陷入局部最优。典型算法包括 K2 算法和爬山算法。K2 算法具有非常突出的学习性能,但是需要提供节点之间的先后顺序和最大父节点数作为输入参数,而这在实际中难以满足。为了解决 K2 算法中的节点序问题,2018 年 Tabar 等^[9]基于 L1 正则化马尔可夫毯和改进的评分函数为每个节点寻找其父节点候选集;2020 年 Behjati 等^[10]采用递归方法寻找强连通分量中的节点序,从而构建出完整的节点序。针对局部搜索算法容易陷入局部最优的问题,诸多优化算法被引入 BN 结构学习。2019 年 Yan 等^[11]引入无需调参的 Jaya 算法并结合遗传算法的交叉变异思想设计了更新策略,Liu 等^[12]提出一种基于改进鲸鱼优化策略的具有全局收敛性的 BN 结构学习算法,Liu 等^[13]结合樽海鞘算法和差分进化算法对局部搜索算法进行改进,在收敛精度和寻优效率上得到了一定的提升。2020 年 Li 等^[14]采用了布谷鸟搜索(Cuckoo Search,CS)和粒子群优化(Particle Swarm Optimization,PSO)算法,在算法收敛速度和精度上均取得了很好的效果。

基于等价类空间搜索的方法,其搜索空间中的点不再是单个 BN 结构,而是一类具有等价关系的结构,其经典算法为贪婪等价搜索算法(Greedy Equivalent Search,GES)^[15]。然而,GES 算法的效率并不高,同时也存在局部最优的问题。2013 年,Alonso-Barba 等^[16]通过限制等价类空间,对 GES 算法进行改进以提升计算效率。2018 年,Alonso 等^[17]基于局部搜索的元启发式方法对 GES 方法做了改进,提升了所得网络的评分。

基于节点序空间搜索的方法,对节点之间的顺序进行搜索,然后在得到的最优节点顺序基础上构建贝叶斯网络结构。2003 年,Friedman 等^[18]提出了一种基于 MCMC 的节点序搜索方法,利用 Metropolis-Hastings 采样算法保证采样得到的节点序服从后验概率,从而使该算法具有良好的学习精度。2018 年,Liu 等^[19]提出通过支撑树权重矩阵构建对节点序评价的适应度函数,利用动态学习因子和倒置变异策略,提升遗传算法节点序寻优的性能,最后在所得节点序的基础上采用 K2 算法得到最终的网络结构,在所得网络结构的评分值上有所提升。

混合搜索的 BN 结构学习方法结合两类方法的优点,首先利用条件独立性检测缩减网络结构空间大小,再进一步采用评分搜索得到最优网络结构。

3 Sub-BN-Merge

3.1 Sub-BN-Merge 算法原理

K2 算法通过为每个节点搜索父节点集合的方式来构建贝叶斯网络结构,且要求提供所有节点之间的整体顺序。然而,为每个节点搜索父节点集合并不以所有节点之间的整体顺序为前提,只需要在确保所得网络结构中无环的情况下为每个节点提供候选父节点集合即可,或者以一定的方式消除网络结构中的环。

以每个节点为中心,若能够挑选出少量节点并通过现有的结构学习方法获得一个 Sub-BN,则该 Sub-BN 能够以一定精度为其中的节点对提供先后顺序信息,即有向边的信息。为每个节点构建一个这样的 Sub-BN,则能够获得诸多节点对之间的顺序关系。当然,一对节点之间的先后顺序可能在某个 Sub-BN 和另一个 Sub-BN 中截然相反。但是,对所有的 Sub-BN 的有向边信息进行计数,并以 Voting 的方式进行融合,则可以为每个节点提供一个较为准确的候选父节点集合,这可以视为节点之间的一种局部顺序。构建 Sub-BN 时所需的少量节点可以通过节点之间的相关性来挑选,而互信息恰是一种较好的相关性度量。在整体的 BN 中,虽然相距较远的节点之间的互信息不一定很小,但是相邻节点之间的互信息必然不会过小。为此,本文以互信息为标准选择子网的节点集。K2 算法中,节点顺序表中某个节点之前的所有节点均为其候选父节点。Sub-BN-Merge 中采取类似处理,以节点的所有前驱节点构成其候选父节点集合,而非仅采用其直接前驱节点。

通过上述子网融合的方式可以获得每个节点的候选父节点集合,然后借鉴 K2 的流程,以评分搜索的方式为每个节点搜索最优的父节点集合,即可获得一个整体的有向图网络结构。此时,所得有向图网络中可能存在环。以一定的方式消除网络中的环,所得 DAG 将是一个较为接近真实网络结构的 DAG。

另一方面,爬山算法对初始网络结构的依赖性很强,易陷入局部最优。初始网络结构越接近于真实网络结构,则爬山算法最后所得结果就越接近于真实网络。因此,为爬山算法选择一个良好的初始网络结构至关重要。以上述消除环之后所得 DAG 为初始网络结构,进一步采用爬山算法等启发式搜索方法进行优化,最后能够得到较优的 BN 网络结构。

综合上述思路,本文所提出的基于子网融合的贝叶斯网络结构学习算法包括如下 4 个主要步骤。

(1)子网构建及节点顺序信息融合。基于互信息选取子网节点,并构建 Sub-BN,对所有 Sub-BN 的有向边进行计数统计,Voting 融合得到每一个节点的候选父节点集合。

(2)构建整体有向图。采用 K2 算法的贪婪策略,基于评分函数,为每一个节点从候选父节点集合中搜索最终的父节点集,得到一个可能存在环的有向图。

(3)消环。消除有向图中可能存在的环,得到一个有向无环图。

(4)启发式优化。以所得有向无环图为初始网络结构,采用启发式搜索方法进一步优化,得到最终的整体 BN 结构。

图 1 为 Sub-BN-Merge 算法流程示意图。由于算法需要

为每个节点先构建一个子网,因此其对于节点数较少的网络并不具有优势,更适用于中大型网络。

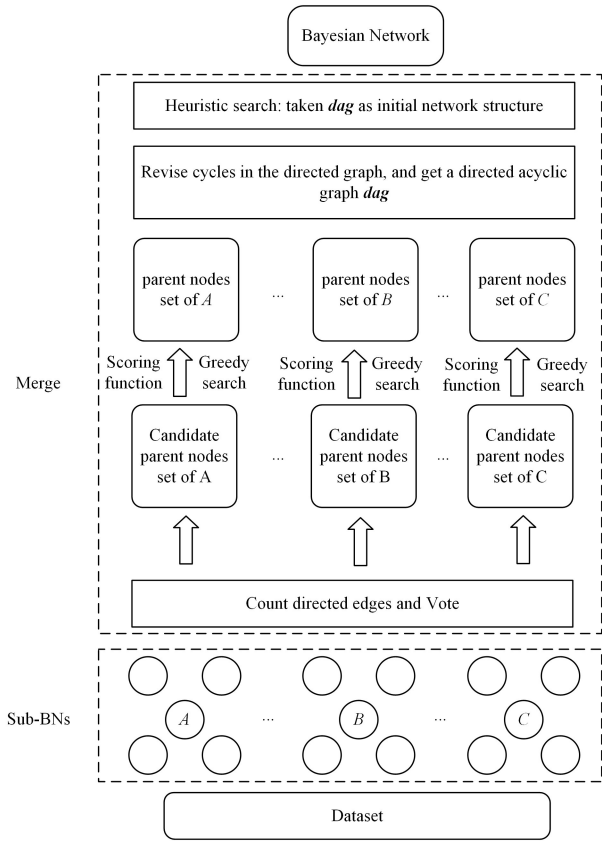


图 1 Sub-BN-Merge 算法流程示意图
Fig. 1 Flow chart of Sub-BN-Merge

3.2 Sub-BN-Merge 算法流程

3.1 节对 Sub-BN-Merge 算法给出了一个框架性的描述,这一节对其实现流程进行详细说明。

子网构建过程由于没有先验的节点顺序信息,故不能采用 K2 算法。本文采用爬山算法进行子网构建。

节点顺序信息Voting 融合过程中,将节点的候选父节点集分为直接候选集和扩展候选集,以节点 A 和节点 B 之间的先后顺序为例进行说明。首先,对所有 Sub-BN 中的有向边 $A \rightarrow B$ 和 $B \leftarrow A$ 计数,分别记为 a 和 b 。若 $a > 0$ 且 $b = 0$,则将节点 A 添加到节点 B 的直接候选集中,反之亦然。在 $a > 0$ 且 $b > 0$ 的情况下,若 $a - b \geq 2$,则将节点 A 添加到节点 B 的直接候选集中;若 $b - a \geq 2$,则将节点 B 添加到节点 A 的直接候选集中;若 $|b - a| < 2$,则节点 A 和节点 B 之间的先后顺序难以确定,故将节点 A 添加到节点 B 的扩展候选集中,同时将节点 B 添加到节点 A 的扩展候选集中。然后,对直接候选集中的节点进行递归搜索,即若 $A \rightarrow B \rightarrow C$,则 A 和 B 均为 C 的候选父节点。最后,将扩展候选集中的节点添加到候选父节点集合中。

消环过程中,采用基于评分的方式进行。对环中的每一条有向边遍历,从环中删除该条边,然后计算环中剩余部分的评分。选择最大评分所对应的删除操作来消除该环。删除有向边不会增加有向图中的环,只会减少环,甚至删除一条边可能消除多个环。以这种方式逐个消除有向图中的环,直至消除所有的环为止。

启发式优化过程中,本文采用爬山搜索算法。

Sub-BN-Merge 算法伪代码见算法 1。

算法 1 Sub-BN-Merge 算法

输入:包含 n 个节点的数据集 D,D 中包含 m 个数据样本;每个 Sub-BN 的节点个数 K;最大父节点数(最大入度) μ

输出:每个节点的父节点集

```
1. 计算所有节点的互信息矩阵 /* 对称矩阵 */
2. for i←1 to n do /* 子网构建 */
3.   选择与  $X_i$  互信息最大的 K-1 个节点;
4.   采用爬山算法学习子网 sub_bni
5. end /* for */
6. 对所得子网中的所有有向边计数,得到 edge_set;
7. 构建每个节点的直接候选父节点集与扩展的候选父节点集
8.   for i←1 to n do
9.     从直接候选父节点集中递归获取  $X_i$  的所有前驱节点构成集合 DPS;
10.    从扩展候选父节点集中获取  $X_i$  的扩展的候选父节点构成集合 EPS;
11.    合并 DPS 和 EPS,得到  $X_i$  的候选父节点集 CPS;
12.    采用贪婪策略,以  $\mu$  为最大父节点数,从 CPS 中搜索  $X_i$  的父节点集合;
13.  end /* for */
14. 将每个节点与其父节点之间的有向边拼接起来,得到有向图 dg;
15. while(exist cycle in dg) do
16.   find a cycle  $C_j$ 
17.   去掉  $C_j$  中对评分影响最小的一条边;
18. end /* while */ /* 将所得有向图环图记为 dag */
19. result_dag=hill_climbing(start_dag=dag);
20. return result_dag
```

3.3 评分函数的选择

评分函数有两类:一类在给定数据集情况下,选择最高后验概率的网络结构作为最佳结构,最具代表性的是 BD 评分及其变体 K2 评分^[5]和 BDeu 评分;另一类评分函数同时考虑网络结构对数据的拟合度以及结构的复杂度,其代表是贝叶斯信息准则 BIC 评分^[6]和最短信息准则 MDL。因此,评分搜索方法所得结果相对于真实的网络结构新增的边,采用第一类评分函数会多于采用第二类评分函数,相应的汉明距增大,算法准确度下降。本文中,第一类评分函数选用 K2 评分,第二类评分函数选用 BIC 评分。

Sub-BN-Merge 算法中的 4 个主要步骤中,均涉及到评分函数的选用。每个步骤的目的不同,因此选用不同的评分函数。在第一步子网构建过程中,希望挖掘出更多的节点对之间的先后顺序信息,则选用 K2 评分。在其后的 3 个步骤中,包括整体有向图构建、消环和启发式优化,其目的是能够提供一个接近真实网络的初始网络结构并进一步优化得到最优联网了,需要控制网络结构的复杂度,故选用 BIC 评分。

4 实验结果及分析

为了验证 Sub-BN-Merge 算法的性能,从通用 Benchmark 贝叶斯网络库(the Bayesian Network repository, BNR)^[20]中选择了 3 个不同规模的网络:小型网络 Asia、中型网络 Alarm 以及大型网络 Win95pts。表 1 列出了所

选用的 3 个网络的节点数、有向边数、最大父节点数以及网络规模。

表 1 算法性能验证所选用的贝叶斯网络
Table 1 Bayesian networks used in experiment

Network	Nodes	Arcs	Max Parent Number	Network scale
Asia	8	8	2	Small
Alarm	37	46	4	Medium
Win95pts	76	112	7	Large

对以上 3 个网络进行采样,分别生成数据集,样本量包括 500,1 000,5 000 和 10 000,数据缺失率涵盖 0%,1%,3%,5%,10%以及 20%。对每一样本量和缺失率生成 10 组数据集,所得结果均为这 10 组数据的平均值。

本文采用结构汉明距 (Structural Hamming Distance, SHD)和算法正确率 (Correct Rate, CR)来验证算法的性能。其中,结构汉明距 SHD 是正确的边数(the Correct Edges, CE)、多余的边数(the Additional Edges, AE)、丢失的边数(the Missing Edges, ME)以及反向的边数(the Reverse Edges, RE)之和。

结构汉明距: $SHD=AE+ME+RE$

算法正确率: $CR=\frac{CE}{CE+AE+ME+RE}$

选用主流的 4 个对比算法:HC 算法、K2 算法、MWST 算法、OBS 算法。其中,HC 算法以包含所有节点的空网络为初始网络结构;K2 算法节点序随机生成;MWST 算法通过 MWST 生成节点顺序,然后采用 K2 算法得到最终结果;OBS 算法以邻接点交换生成节点序的邻域值,采用爬山搜索获得最佳节点序,采用随机重启进行优化^[21]。对比算法均采用 BIC 评分。所有算法的最大父节点数为实际值。

4.1 数据无缺失值的实验结果

在数据无缺失值的情况下,表 2—表 4 分别列出了小型网络 Asia、中型网络 Alarm 和大型网络 Win95pts 的实验结果,加粗的字体表示几种算法中的最优。

实验结果表明,无缺失值的情况下,在结构汉明距和算法正确率方面,本文所提出的 Sub-BN-Merge 算法明显优于另外 4 种对比算法,仅在 500 例数据的 Asia 网络上,实验结果略差。K2 算法由于节点顺序随机生成,与真实的节点顺序差异较大,其结果相当糟糕。

表 2 小型网络 Asia 实验结果
Table 2 Experiment results of small network Asia
(a)structural hamming distance

Size	HC	K2	MWST	OBS	Our
500	4.3	11.5	9.9	4.8	4.5
1 000	5.3	10.8	9.4	4.7	2.9
5 000	4.8	11.0	9.4	4.2	3.1
10 000	6.6	11.0	8.6	2.9	2.1

(b)correct rate

Size	HC	K2	MWST	OBS	Our
500	0.543	0.155	0.172	0.552	0.533
1 000	0.419	0.207	0.213	0.521	0.661
5 000	0.489	0.214	0.250	0.575	0.639
10 000	0.375	0.214	0.298	0.692	0.743

表 3 中型网络 Alarm 实验结果

Table3 Experiment results of medium network Alarm

(a)structural hamming distance

Size	HC	K2	MWST	OBS	Our
500	35.0	59.3	58.2	43.7	28.6
1 000	33.7	60.3	61.6	47.0	27.5
5 000	32.2	64.1	68.9	46.9	21.4
10 000	30.2	66.1	76.0	47.5	19.7

(b)correct rate

Size	HC	K2	MWST	OBS	Our
500	0.341	0.128	0.061	0.284	0.449
1 000	0.381	0.157	0.064	0.272	0.485
5 000	0.429	0.189	0.069	0.315	0.598
10 000	0.471	0.189	0.062	0.333	0.630

表 4 大型网络 Win95pts 实验结果

Table 4 Experiment results of medium network Alarm

(a)structural hamming distance

Size	HC	K2	MWST	OBS	Our
500	104.3	145.6	152.6	141.7	94.5
1 000	103.1	154.3	161.4	141.8	83.6
5 000	102.3	173.6	209.5	163.0	63.2
10 000	96.9	183.2	230.5	155.5	59.7

(b)correct rate

HC	K2	MWST	OBS	Our	
500	0.274	0.154	0.098	0.162	0.336
1 000	0.299	0.159	0.097	0.200	0.406
5 000	0.334	0.176	0.080	0.209	0.535
10 000	0.379	0.185	0.076	0.219	0.565

4.2 数据存在缺失值的实验结果

对于基于评分搜索的结构学习算法,数据缺失值主要影响评分函数的计算。对于数据缺失值的处理,本文删除存在缺失值的行,不考虑缺失数据填充。计算评分函数的时候,缺失数据的删除有两种策略,第一种策略是若整个数据集中某一行存在缺失值,直接删除这一整行数据,这种策略简单直接,但可能会损失很多数据信息;第二种策略是针对每一次评分函数计算时的局部结构,剔除局部结构所包含节点对应数据中存在缺失值的行,这种策略能够充分利用存在缺失值的行中的非缺失数据。第一种策略在网络节点较少或者缺失率较低(不超过 5%)的时候可能效果更好,第二种策略在网络节点较多且缺失率较高(10%及以上)的时候效果更好。本文对于存在缺失值的情况,采用第二种策略。考虑数据缺失率为 1%,5%以及 10%的情况。

表 5—表 7 分别列出了 Asia 网络在缺失率分别为 1%,5%及 10%的实验结果。结果表明,本文所提出的 Sub-BN-Merge 算法在结构汉明距和算法正确率上优于其他几种算法,而且随着数据量的增多,结构汉明距呈降低趋势,算法正确率呈升高趋势。

表 8—表 10 分别列出了 Alarm 网络数据缺失率分别为 1%,5%及 10%的实验结果。结果表明,本文所提出的 Sub-BN-Merge 算法在结构汉明距和算法正确率上优于其他几种算法。数据量为 500,缺失率为 10%的情况下,本文方法的结果虽然比其他几种算法效果更好,但是正确率也较低。

表 5 Asia 网络缺失率为 1%的实验结果

Table 5 Experiment result of Asia with missing rate 1%

(a)structural hamming distance

Size	HC	K2	MWST	OBS	Our
500	4.4	11.4	10.7	5.6	3.8
1 000	4.4	10.8	10.8	4.0	2.8
5 000	6.7	11.0	11.8	4.1	2.4
10 000	4.1	11.0	12.3	3.1	2.3

(b)correct rate

Size	HC	K2	MWST	OBS	Our
500	0.516	0.149	0.102	0.465	0.584
1 000	0.499	0.207	0.140	0.595	0.661
5 000	0.307	0.214	0.115	0.599	0.711
10 000	0.607	0.214	0.066	0.679	0.719

表 6 Asia 网络缺失率为 5%的实验结果

Table 6 Experiment result of Asia with missing rate 5%

(a)structural hamming distance

Size	HC	K2	MWST	OBS	Our
500	5.1	10.9	9.9	5.8	4.3
1 000	5.7	10.9	11.0	3.7	2.9
5 000	5.9	11.0	11.7	3.8	2.0
10 000	5.4	11.0	12.0	3.3	2.4

(b) correct rate

Size	HC	K2	MWST	OBS	Our
500	0.439	0.162	0.133	0.454	0.542
1 000	0.361	0.189	0.114	0.626	0.666
5 000	0.390	0.214	0.145	0.615	0.750
10 000	0.452	0.214	0.128	0.659	0.713

表 7 Asia 网络缺失率为 10%的实验结果

Table 7 Experiment result of Asia with missing rate 10%

(a) structural hamming distance

Size	HC	K2	MWST	OBS	Our
500	5.7	11.7	11.6	6.1	5.0
1 000	4.7	10.8	11.1	5.7	3.5
5 000	4.7	11.3	11.0	4.6	3.1
10 000	4.7	11.0	10.9	3.1	2.5

(b)correct rate

Size	HC	K2	MWST	OBS	Our
500	0.353	0.108	0.026	0.415	0.435
1 000	0.475	0.177	0.084	0.438	0.595
5 000	0.512	0.200	0.144	0.556	0.639
10 000	0.503	0.214	0.171	0.674	0.701

表 8 Alarm 网络缺失率为 1% 的实验结果

Table 8 Experiment result of Alarm with missing rate 1%

(a)structural hamming distance

Size	HC	K2	MWST	OBS	Our
500	34.0	60.1	56.7	44.9	30.4
1000	34.1	59.1	57.8	44.8	28.5
5000	30.5	63.4	64.3	48.2	20.9
10000	31.0	64.7	72.4	46.0	19.6

(b)correct rate

Size	HC	K2	MWST	OBS	Our
500	0.339	0.103	0.087	0.255	0.404
1 000	0.364	0.147	0.110	0.284	0.456
5 000	0.442	0.161	0.108	0.301	0.602
10 000	0.460	0.192	0.078	0.337	0.629

表 9 Alarm 网络缺失率为 5% 的实验结果

Table 9 Experiment result of Alarm with missing rate 5

(a) structural hamming distance					
Size	HC	K2	MWST	OBS	Our
500	45.1	63.0	62.4	55.9	38.1
1000	42.8	60.9	61.1	51.9	38.7
5000	32.7	59.8	60.8	47.8	24.7
10000	32.6	59.0	64.1	48.1	20.1

(b) correct rate					
Size	HC	K2	MWST	OBS	Our
500	0.145	0.068	0.046	0.130	0.237
1000	0.179	0.080	0.044	0.162	0.239
5000	0.375	0.126	0.059	0.242	0.509
10000	0.392	0.167	0.052	0.265	0.600

表 10 Alarm 网络缺失率为 10% 的实验结果

Table 10 Experiment Result of Alarm with Missing Rate 10%

(a) structural hamming distance					
Size	HC	K2	MWST	OBS	Our
500	46.4	76.7	78.1	75.7	33.0
1000	46.9	71.0	73.6	70.6	30.7
5000	44.2	62.4	64.7	55.9	26.3
10000	41.4	61.1	62.2	49.8	23.3

(b) correct rate					
Size	HC	K2	MWST	OBS	Our
500	0.004	0.027	0.010	0.034	0.359
1000	0.015	0.043	0.018	0.055	0.400
5000	0.144	0.070	0.015	0.134	0.500
10000	0.182	0.077	0.016	0.179	0.563

表 11—表 13 列出了大型网络 Win95pts 在数据缺失率分别为 1%, 5% 及 10% 的实验结果。

表 11 Win95pts 网络缺失率为 1% 的实验结果

Table 11 Experiment result of Win95pts with missing rate 1%

(a) structural hamming distance					
Size	HC	K2	MWST	OBS	Our
500	113.1	142.0	140.0	142.4	97.8
1000	111.0	146.5	145.7	139.9	88.8
5000	109.3	161.6	177.2	150.0	68.6
10000	106.7	169.8	195.5	143.8	56.4

(b) correct rate					
Size	HC	K2	MWST	OBS	Our
500	0.188	0.128	0.113	0.124	0.274
1000	0.225	0.140	0.113	0.161	0.352
5000	0.294	0.171	0.094	0.201	0.494
10000	0.319	0.178	0.098	0.244	0.570

表 12 Win95pts 网络缺失率为 5% 的实验结果

Table 12 Experiment result of Win95pts with missing rate 5%

(a) structural hamming distance					
Size	HC	K2	MWST	OBS	Our
500	118.8	170.4	173.2	171.1	103.9
1000	114.8	163.7	166.3	162.3	96.2
5000	109.7	143.0	146.8	141.2	75.2
10000	105.6	137.4	142.4	136.5	75.2

(b) correct rate					
Size	HC	K2	MWST	OBS	Our
500	0.035	0.036	0.024	0.030	0.249
1000	0.061	0.045	0.035	0.052	0.302
5000	0.128	0.098	0.061	0.103	0.442
10000	0.179	0.118	0.069	0.127	0.453

表 13 Win95pts 网络缺失率为 10% 的实验结果

Table 13 Experiment result of Win95pts with missing rate 10%

(a) structural hamming distance					
Size	HC	K2	MWST	OBS	Our
500	112.0	182.8	185.1	182.0	118.4
1000	112.0	183.0	184.5	184.1	103.9
5000	112.0	182.3	184.2	182.0	84.4
10000	112.0	179.0	180.0	177.0	77.0

(b) correct rate					
Size	HC	K2	MWST	OBS	Our
500	0.000	0.011	0.005	0.010	0.164
1000	0.000	0.011	0.005	0.007	0.233
5000	0.000	0.012	0.006	0.012	0.378
10000	0.000	0.020	0.012	0.024	0.429

结果表明, 本文所提出的方法优于对比算法。网络 Win95pts 共 76 个节点, 112 条有向边。当缺失率为 10% 的时候, 从表 13 可以看出, 实际上爬山算法所得结果就是初始网络结构, 即包含所有节点的空网络。同 Alarm 网络一样, 虽然 Sub-BN-Merge 方法的实验结果优于对比算法, 但是其正确率并不高, 这是网络节点多且数据量少的必然结果。

结束语 从数据中学习贝叶斯网络结构, K2 算法要求实际中难以满足的正确节点顺序且节点顺序的正确性对最终结果影响甚大, 爬山算法 HC 对初始网络结构依赖性强且容易陷入局部最优。为此, 本文提出了基于子网融合的贝叶斯网络结构学习算法 Sub-BN-Merge, 该算法为每个节点构建一个子网, 然后通过 Voting 的方式融合得到每个节点的候选父节点集, 并以贪婪搜索策略获得真正的父节点集, 再通过消环和启发式搜索进一步优化得到最终的网路结构。在小型网络 Asia、中型网络 Alarm 和大型网络 Win95pts 上通过模拟数据验证了算法的有效性, 以结构汉明距和算法正确率分析了算法的性能。实验结果表明 Sub-BN-Merge 算法有着良好的性能。然而, 在网络节点较多的情况下, 特别是存在缺失数据的时候, Sub-BN-Merge 算法的性能有待进一步提升, 这将是后续的研究内容之一。

参 考 文 献

[1] KYRIMI E, MOSSADEGN S, TAI N, et al. An incremental explanation of inference in Bayesian networks for increasing model trustworthiness and supporting clinical decision making [J]. Artificial Intelligence In Medicine, 2020, 103.

[2] ADABOR E S, ACQUAAH-MENSAH G K. Restricted-derestricted dynamic Bayesian Network inference of transcriptional regulatory relationships among genes in cancer [J]. Computational Biology and Chemistry, 2019, 79: 155-164.

[3] CHATRABGOUN O, HOSSEINIAN-FAR A. Approximating non-Gaussian Bayesian networks using minimum information vine model with applications in financial modelling [J]. Journal of Computational Science, 2018, 24: 266-276.

[4] ADEDIPE T, SHAFIEE M. Bayesian Network Modelling for the Wind Energy Industry: An Overview [J]. Reliability Engineering and System Safety, 2020, 202.

[5] COOPER G F, HERSKOVITS E. A Bayesian Method for the Induction of Probabilistic Networks from Data [J]. Machine Learning, 1992, 9(4): 309-347.

[6] SCUTARI M, GRAAFLAND C E. Who learns better Bayesian

- network structures; Accuracy and speed of structure learning algorithms [J]. International Journal of Approximate Reasoning, 2019, 115: 235-253.
- [7] SPIRTER P, GLYMOUR C. Causation, Prediction, and Search (2nd edition) [M]. The MIT Press, 2000.
- [8] CHICKERING D M, HECKERMAN D. Large-Sample Learning of Bayesian Networks is NP-Hard [J]. Journal of Machine Learning Research, 2004, 5: 1287-1330.
- [9] TABAR V R, ESKANDARI F. Finding a set of candidate parents using dependency criterion for the K2 algorithm [J]. Pattern Recognition Letters, 2018, 111: 23-29.
- [10] BEHJATI S, BEIGY H. Improved K2 algorithm for Bayesian network structure learning [J]. Engineering Applications of Artificial Intelligence, 2020, 91.
- [11] YAN ZHI, ZHANG P. Research on Bayesian Network Structure Learning Algorithm Based on Jaya [J]. Computer Engineering and Applications, 2019, 55(19): 173-177, 184.
- [12] LIU H R, ZHANG L Y. Bayesian Network Structure Learning Based on Improved Whale Optimization Strategy [J]. Journal of Electronics & Information Technology, 2019, 41 (6): 1434-1441.
- [13] LIU B, FAN R X. Bayesian network structure learning algorithm based on hybrid binarysalp swarm-differential evolution algorithm [J]. Journal on Communications, 2019, 40 (7): 151-161.
- [14] LI J W, LI G N. Application of CS-PSO algorithm in Bayesian Network structure learning [J]. Journal of Measurement Science and Instrumentation, 2020, 11(1): 94-102.
- [15] CHICKERING D M. Optimal Structure Identification With Greedy Search [J]. Journal of Machine Learning Research, 2002, 3: 507-552.
- [16] ALONSO-BARBA J I, OSSA L D L. Scaling up the Greedy Equivalence Search algorithm by constraining the search space of equivalence classes [J]. International Journal of Approximate Reasoning, 2013, 54: 429-451.
- [17] ALONSO J I, OSSA L D L. On the use of local search heuristics to improve GES-based Bayesian network learning [J]. Applied Soft Computing, 2018, 64: 366-376.
- [18] FRIEDMAN N, KOLLER D. A Bayesian Approach to Structure Discovery in Bayesian Networks [J]. Machine Learning, 2003, 50: 95-125.
- [19] LIU B, WANG H Y. Learning Bayesian Network Structure from Node Ordering Searching Optimal [J]. Journal of Electronics & Information Technology, 2018, 40(5): 1234-1241.
- [20] MARCO S. bnlearn-an R package for Bayesian network learning and inference [EB/OL]. [2017-09-06]. <http://www.bnlearn.com/bnrepository>, 2013.
- [21] TEYSSIER M, KOLLER D. Ordering-Based Search: A Simple and Effective Algorithm for Learning Bayesian Networks [J]. arXiv:1207.1429.



ZHONG Kun-hua, born in 1984, doctoral student. His main research interests include machine learning, big data intelligent analysis and medical application of AI.



QIN Xiao-lin, born in 1980, Ph.D, professor, Ph.D supervisor. His main research interests include automated reasoning, swarm intelligence and big data intelligent analysis.