基于多头绒泡菌的贝叶斯网络结构学习

林 朗 张自力
(西南大学计算机与信息科学学院 重庆 400715)

摘 要 贝叶斯网络是概率统计与图论相结合的一种图模型,已成功应用于多个领域中。然而,仅依赖专家的领域知识构建贝叶斯网络非常困难。因此,从数据中学习贝叶斯网络结构已经成为该研究领域的重点问题。针对贝叶斯网络结构学习搜索空间太大的问题,根据多头绒泡菌在觅食过程中展现出的保留重要觅食管道的特性,文中结合多头绒泡菌相关数学模型和条件互信息理论对原始搜索空间进行缩减,并将求解得到的无向图作为网络的基础骨架;之后利用爬山法确定骨架方向,并得到对应的拓扑排序;最后将节点顺序作为K2算法的输入以求得最终网络,并选用网络拓扑结构及评分作为评价指标在多个数据集上进行对比实验。实验结果表明,所提算法在网络重构及原始数据匹配上具有更高的准确度。

关键词 贝叶斯网络,结构学习,多头绒泡菌,条件互信息
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.09.030

Bayesian Structure Learning Based on Physarum Polycephalum

LIN Lang ZHANG Zi-li
(School of Computer and Information Science, Southwest University, Chongqing 400715, China)

Abstract Bayesian network is a graph model which combines probability statistics and graph theory. It has been successfully applied in many fields. However, it is very difficult to build Bayesian network only depending on the domain knowledge of experts. Therefore, learning Bayesian network structure from data has become a key issue in this field. Bayesian network structure learning is a NP-hard problem because the search space is too large. According to the characteristics of physarum polycephalum in the process of foraging to retain an important feeding pipeline, the original search space was reduced by combining the relevant mathematical model of physarum polycephalum and the theory of conditional mutual information. In this paper, the obtained undirected graph is used as the basic framework of the network, and then the mountain climbing method is used to determine the direction of the skeleton and get the corresponding topological ordering. Finally, the order of nodes is used as the input of K2 algorithm to get the final network. The network topology structure and score are selected as the evaluation index, and comparative experiments are carried out on multiple data sets. Experiments show that the proposed algorithm has higher accuracy in network reconfiguration and raw data matching.

Keywords Bayesian network, Structure learning, Physarum polycephalum, Independence condition mutual information

1 引言

贝叶斯网络是一个可用来进行概率推断的有向无环图,已经被广泛应用于大量任务中,如无人驾驶[1]、语音识别[2]等。贝叶斯网络学习包含参数学习和结构学习两大部分,结构学习是为了确定网络中哪些节点间存在边及边的方向。由于随着网络中节点数目的增长,网络结构空间呈指数级增长,因此该问题是一个NP难问题[3]。此问题引发了研究者的广泛关注,他们提出了一系列求解方法。这些方法主要分为两类:1)利用变量间依赖测试[4-5]将一些相关程度高的节点连接起来作为网络结构,但该方法对依赖测试的准确性要求较高;2)利用评分函数对每个结构进行打分[6-7](评分函数用于评价网络结构与数据集的匹配程度,如贝叶斯信息准则(BIC)[8]、Akaike准则(AIC)[9]、贝叶斯狄利克雷分数(BD)[10]等),并通过最大化评分来指导搜索过程以求取最佳网络结构。这类方法由于搜索空间大,往往需要大量时间,为了克服这种缺点,人们提出了一些利用节点顺序的算法,如K2[11]算法。虽然该类算法能高效求解问题,但不同的节点顺序会对最终结果带来极大的影响。对此,有研究者提出了几种确定顺序的算法,但效果都不是太好[12-13]。

到稿日期:2019-03-13 返修日期:2019-05-23
林 朗(1994—),男,硕士生,主要研究方向为贝叶斯网络;张自力(1964—),男,博士,教授,主要研究方向为多Agent系统等,E-mail:zhangzl@swu.edu.cn(通信作者)。

为找到一个好的节点顺序,本文提出首先在一个较小的搜索空间中对贝叶斯网络结构进行学习,得到一个较为可靠的初始贝叶斯网络结构,并用拓扑排序对该结构的节点进行排序,以得到节点顺序。为了得到一个较小的搜索空间,本文利用多头绒泡菌在觅食过程中具有保留重要管道的特性,首先计算条件互信息以得到一张完全图;然后模拟多头绒泡菌在该图进行食物搜索的过程,得到一张只包含少量边的无向图作为结构约束,以缩小搜索空间;接着用爬山法确定该无向图中边的方向,并得到该网络对应的节点的拓扑排序;最后用得到的节点顺序作为 K2 算法的输入,以求取对应的最大分数的贝叶斯网络。

2　背景介绍

本节简要介绍贝叶斯网络结构学习的相关概念及 K2 算法的求解思路。

2.1　贝叶斯网络结构学习

贝叶斯网络表示随机变量联合概率分布的图模型,由两部分构成:1)有向无环图(DAG)$G=(V,E)$,其中 $V=\{X_1,X_2,\cdots,X_n\}$,表示随机变量集,E 是以 V 中元素为节点的有序节点对,表示 n 个变量之间的相关关系,当存在 X_i 到 X_j 的边时,称 X_i 是 X_j 的父节点;2)条件概率表(CPT),量化了节点 X_i 的父节点集 π_i 对 X_i 的影响。由马尔可夫条件可知,任意变量 X_i 在给定其父集 Π_i 的情况下,独立于其所有非后继节点。因此,贝叶斯网络的联合概率分布可表示为:

$$p(X_1,X_2,\cdots,X_n)=\prod_{i=1}^{n}p(X_i|\pi_i) \tag{1}$$

给定数据集 $D=\{D_1,\cdots,D_N\}$,贝叶斯学习的目标是评分函数最大化。本文选用 BIC 评分作为评分标准:

$$S_D(G)=\max_{\theta}\log\prod_{i=1}^{n}\prod_{j=1}^{r_{\pi_i}}\prod_{k=1}^{r_i}\theta_{ijk}^{n_{ijk}}-\frac{\log N}{2}\sum_{i=1}^{n}(r_{\pi_i}(r_i-1)) \tag{2}$$

其中,r_i 为变量 X_i 的类别数;$\theta_{ijk}=p(x_i=k|\Pi_i=j)$;$n_{ijk}$ 为数据集中变量 X_i 与其父集 Π_i 分别取值为 j,k 时对应的样本个数;N 为总的样本个数。BIC 评分度量了每个样本在给定贝叶斯网络下出现的概率以及模型的复杂程度,总的出现概率越大,模型就越简单,值也就越高。

2.2　基于 K2 的结构学习

K2 算法是一种利用节点顺序作为输入,并从数据中学习网络结构的贪心算法,最早由 Cooper[14] 提出。该算法以节点顺序为输入,并规定如果节点 X_i 的序号在 X_j 之前,则 X_j 不可能是 X_i 的父节点,换句话说,每个节点 X_i 的父集都只能取在其序号前的节点。节点顺序信息的存在大大减小了搜索空间,因此 K2 算法是非常高效的。算法初始时,每个节点的父集都被设置为空集,算法根据节点顺序对每个节点进行遍历,对于每个节点 X_i,将贪心添加在它之前的所有节点作为父集,以最大化网络评分,直到分数不再增加或者没有可添加的节点为止。K2 算法的运行效果主要取决于节点顺序,不同的节点顺序会对结果造成很大的影响。

然而,大部分情况下节点顺序是未知的,如何确定顺序是该算法的难点。为此,本文首先结合多头绒泡菌数学模型和条件互信息来求出节点顺序,之后再用 K2 算法构造贝叶斯网络。

3　基于多头绒泡菌的贝叶斯网络构造方法

基于多头绒泡菌的贝叶斯网络结构学习算法分为 4 个步骤:1)利用条件互信息得到一张完全图;2)用多头绒泡菌模型对该图去边;3)利用爬山法对简化后的图进行评分搜索,以得到初始的贝叶斯网络;4)针对初始网络,利用拓扑排序得到节点顺序,并将节点顺序作为 K2 的输入,以求得最终网络。

3.1　利用条件互信息得到一张完全图

使用多头绒泡菌求解器对原搜索空间进行约束的核心问题是如何定义变量之间的距离。本文将变量间的相关程度定义为距离,相关程度越高,距离越近,变量间的边越容易在求解过程中被保留。相关程度可由互信息描述,两变量的相关程度越高,则若已知其中一个变量的取值,就能为确定另一个变量的取值情况提供更多的信息,两者之间存在边的可能性亦越高;但由于变量之间可能存在其他条件变量,当已知条件变量取值时,两个本来相关的变量也可能条件独立,因此本文用条件互信息来反映两变量之间的相关程度。

随机变量 X,Y 的互信息定义为:

$$I(X;Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{3}$$

X,Y 在条件变量集 Z 下的条件互信息定义为:

$$I(X;Y|Z)=\sum_{z\in Z}\sum_{x\in X}\sum_{y\in Y}p(x,y,z)\log\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \tag{4}$$

本文考虑到算法的执行时间,对每对变量仅分别计算条件变量集为空和只包含一个元素情况下的条件互信息,并将求得的最小值作为两变量的实际条件互信息 cmi。为了使两变量间距离的长短表示依赖关系的强弱,即距离短的依赖关系强,本文简单地将距离 L 定义为最小条件互信息 cmi 的倒数。然后对包含 n 个变量的数据集建立一张包含 n 个节点的完全图,将节点 X_i 和 X_j 之间边的长度定义为距离 $L(i,j)$。

$$L(i,j)=1/(\min(I(i,j),\min_{0<k<n}I(i,j|k))) \tag{5}$$

其中,K 为节点集 V 中除了节点 X_i 和节点 X_j 的节点集。

3.2　利用多头绒泡菌求解器求解网络结构约束

3.2.1　多头绒泡菌求解器

多头绒泡菌的数学模型主要来源于 Nakagaki 等[15] 的迷宫求解实验。通过在迷宫中放入多头绒泡菌,在进出口放置食物源,研究者发现随着时间的推移,多头绒泡菌会首先向外扩张,用其胶体管道系统覆盖整个迷宫,之后管道内的原生质会不停流动,靠近食物源的管道内流量会逐渐增加,其余部分逐渐减少,最终找到一条连接迷宫进口和出口的最短路径。Tero 等[16] 对该机制进行了数学建模,该数学模型形式化表示如下。

用 N 表示迷宫内的分支节点,Q_{ij} 表示管道内的流量,L_{ij} 表示节点 N_i 到节点 N_j 的边,D_{ij} 表示管道的导通性,P 表示节点的压力,则由哈根-泊肃叶定律可得:

$$Q_{ij}=\frac{D_{ij}}{L_{ij}}(P_i-P_j) \tag{6}$$

假定网络中的一个节点为源节点，原生质从该点流出，流出流量大小为 1，其余节点为终点，流量流入到这些终点，由此得到多食物源模型。由基尔霍夫电流定律可知整个网络能量守恒，为此建立方程组：

$$\sum_i Q_{ij}=\begin{cases}1, & j=\text{source}\\ -\dfrac{1}{(n-1)}, & j=\text{others}\end{cases} \tag{7}$$

其中，n 为节点个数。由式(7)可求解出网络中各节点的压强，并根据式(6)求解出各管道的流量。随着管道流量的变化，各管道的传导性将按照式(8)变化，将传导性反馈给式(7)，然后循环上述过程。

$$\frac{\mathrm{d}D_{ij}}{\mathrm{d}t}=f(|Q_{ij}|)-D_{ij} \tag{8}$$

其中，函数 f 的计算如下所示：

$$f(Q)=\frac{(1+a)Q^u}{1+aQ^u} \tag{9}$$

其中，a 和 u 为函数参数。当各边的传导性收敛时，迭代结束，得到最终的网络。此时传导性仍然不为零的边被称为重要管道，这些管道构成了多头绒泡菌的一个高效性与鲁棒性并存的觅食网络系统。根据该现象，本文提出利用多头绒泡菌数学模型对上一步的完全图进行缩减。

3.2.2 多头绒泡菌求解网络结构约束

为了确定上一步得到的完全图中哪些边最有可能在原贝叶斯网络结构中存在，本文对每个节点依次运用多食物源模型，即将网络中每个节点都当作食物源，将每条边比作多头绒泡菌的管道，并将选定的节点作为流量的起点，其余点作为流量的终点，每轮循环根据管道流入各节点的流量总和得到各节点的压力值，再根据压力值更新各管道的流量及导通性，之后重复上述过程，直到相邻两轮循环的传导性矩阵 $\boldsymbol{D}$ 中元素的差值都小于 0.0001 为止。保留 $\boldsymbol{D}$ 中传导性大于 0.01 的边，将其作为以节点 i 为源点求解出的结构 E_i。所有节点求完后，将最终结构定义为每次所求结构的并集，即 $E=UE_i$。多头绒泡菌求解网络结构的约束算法的具体步骤如算法 1 所示。

算法 1 多头绒泡菌求解网络结构约束的算法

输入：节点信息 V 及边长度信息 L，输出求得的网络结构约束 E

Input：V，**L**

Output：**E**

%构建完全图的邻接矩阵，$\mathbf{I}_n$ 为单位阵

$\mathbf{E}=\text{ones}(n)-\mathbf{I}_n$；

%对流量、压力、导通性矩阵进行初始化

Q=**E**；**P**=[]；**D**=**E**；

For each $X_i\in V$

repeat

根据式(6)、式(7)计算 t 时刻每个节点的压力 $\mathbf{P}^t$，并求出流量矩阵 $\mathbf{Q}^t$

根据式(8)、式(9)对传导性矩阵 $\mathbf{D}_i^t$ 进行更新

Until $\mathbf{D}_i^{t+1}-\mathbf{D}_i^t<0.0001$

$\mathbf{E}_i=\text{find}(\mathbf{D}_i>0.01)$

End

$\mathbf{E}=\text{union}(\mathbf{E}_1,\mathbf{E}_2,\cdots,\mathbf{E}_n)$；

3.3 运用爬山法对边确定方向

根据 3.2 节得到的网络结构约束，简单地利用爬山法求取对应的贝叶斯网络，即确定约束结构中哪些边存在及这些边的方向，由于上一步得到的图中只包含少量的边，因此搜索空间较原始空间大幅减小。带约束的爬山算法在算法初始时为一张只有节点的空图，之后每轮循环在添边、翻转边、删边操作中选择使网络结构分数增加最大的一种操作，其中添边操作只能选择在原始图中存在的边，算法运行到分数不再增加为止。

3.4 对节点进行排序并通过 K2 确定最终结构

3.3 节求得的贝叶斯网络大部分情况下是不完全的，但能确定大部分边，因此可以得到一个较为准确的节点顺序。本文针对 3.3 节的结果，利用拓扑排序对该网络的节点进行排序，求得节点顺序，并将其作为 K2 算法的输入。

4 实验结果

本文选用 R 语言贝叶斯学习工具包官网上的 7 个网络[1)]作为原始网络，并用其工具包对每个网络生成了 2000 条与原始网络同分布的数据。每个数据集的节点数、边数以及原始网络对应的 BIC 评分信息如表 1 所列。多头绒泡菌的参数 $I=1,u=1,a=4$。首先对边的导通性进行分析，以 sachs 数据集中的任意一个节点为源点，其余节点为终点，运行多头绒泡菌求解器，直至传导性收敛。其余数据集有类似结果，如图 1 所示。

表 1 数据集属性及原始网络对应的 *BIC* 分数

Table 1 Data set attributes and *BIC* score corresponding to original network

数据集	节点数	边数	*BIC*
Asia	8	8	−4577.5
Sachs	11	17	−15027.0
Child	20	25	−25206.0
Insurance	27	52	−29622.0
Water	32	66	−31737.0
Alarm	37	46	−22603.0
Hailfinder	55	66	−106840.0

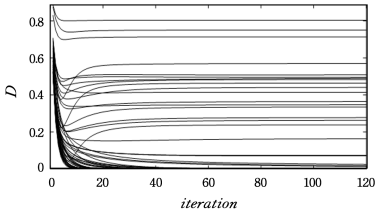

图 1 边导通性与多头绒泡菌求解器运行轮数的关系

Fig. 1 Relationship between edge conductivity and number of running wheels of Phytophthora multicephalus solver

1) http://www.bnlearn.com/bnrepository/

可见在一轮循环结束后，网络中边的导通性值可粗略地分为 3 类：第一类最少，位于顶部，表示该边连接的是一个与该节点紧密联系的节点；第二类稍多，表示连接的是与节点存在间接联系的节点；第三类最多，其值基本接近零，表示与该节点联系少的节点。因此，本文简单地将阈值设置为 0.01，只有大于阈值的边才能存活。

之后选用向前看爬山算法 LAGD[17] 作为对比算法。为了比较算法求得的贝叶斯网络与原始网络之间的结构差异，将算法学习到的边分为 3 类：1）在不考虑边的方向时，原始网络及学得的网络中同时存在的边 L；2）在不考虑边的方向时，原始网络中没有但学得的网络中学到的边 A；3）学得的网络中边的方向与原始网络中边的方向相反的边 R。对这三者的数量进行统计可知：L 越大，A 和 R 越小，则网络结构与原始网络越接近。同时，对生成的网络进行 BIC 评分，该评分值越大，表示与原数据集的分布越接近，如表 2 所列。表格中每行评价指标上方的是本文所提算法的结果，下方是 LAGD 算法在取参数(3,5)情况下的结果。

可见，本文提出的算法在除 Child 数据集以外的其他数据集上求得的分数要优于 LAGD 算法。由于原始网络生成的数据集并不能完全反映原始的概率分布，因此算法所求分数一定程度上高于原始网络的 BIC 分数。在所学结构上，本文算法正确学到的边 L 要多于 LAGD 学到的边，而在多学到的边 A 及错误反转的边 R 上要略少于 LAGD 算法，由此可见本文算法在学得的结构上也更为准确。图 2 展示了 Hailfinder 数据集的原始网络，图 3 展示了本文算法在该数据上的运行结果。从图中可知：所提算法学得的网络能展现出与原网络类似的因果关系。

表 2　本文算法与 LAGD 算法的比较

Table 2　Comparisons between the proposed algorithm and LAGD algorithm

数据集	L	A	R	BIC
Asia	7	0	1	−4573.9
	7	0	1	−4573.9
Sachs	14	2	9	−15008
	13	2	9	−15103
Child	24	4	6	−23430
	23	2	5	−22903
Insurance	34	8	13	−27582
	33	12	13	−27639
Water	25	72	6	−25878
	22	72	6	−26191
Alarm	41	8	13	−22222
	38	8	13	−22425
Hailfinder	55	10	19	−99660
	51	12	20	−99684

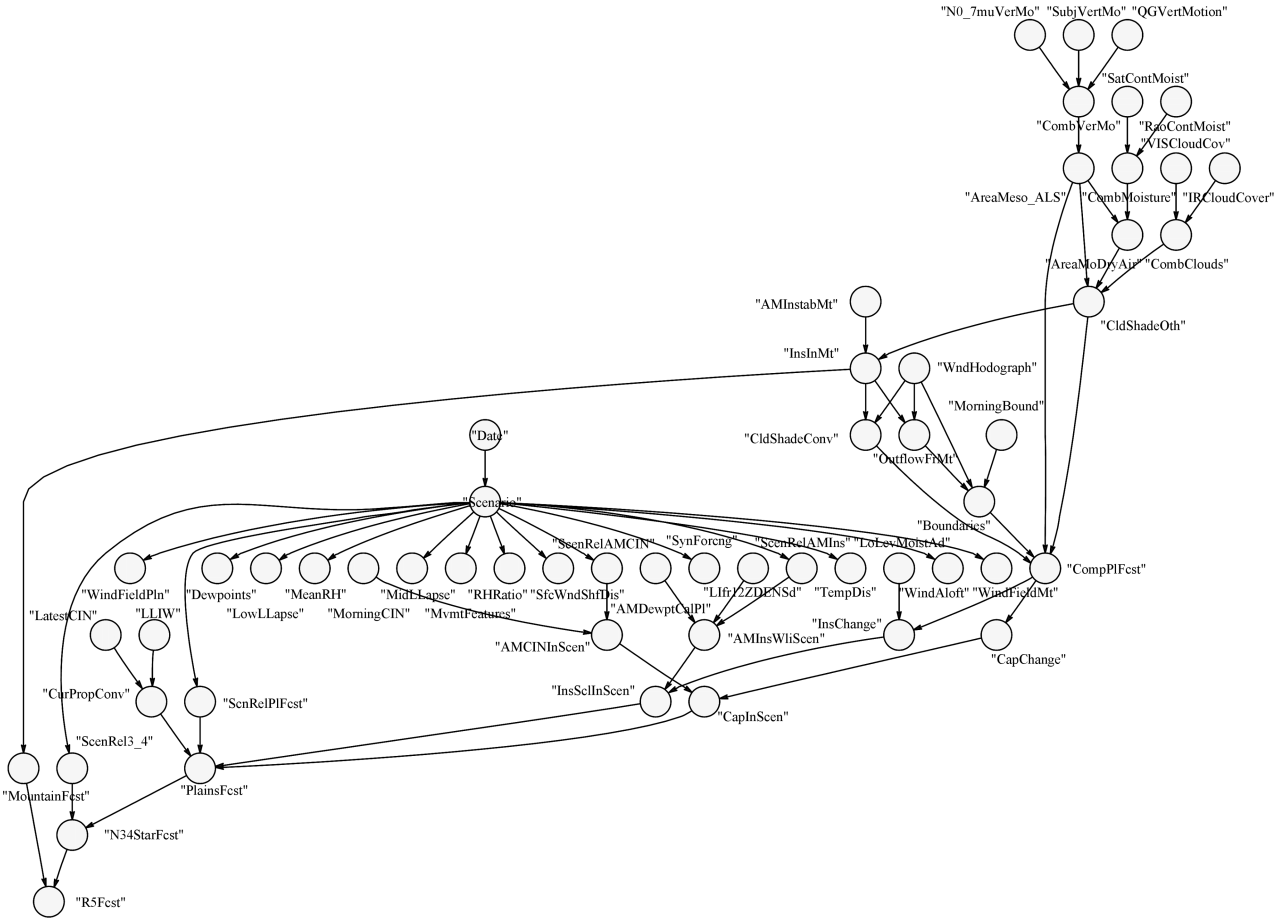

图 2　Hailfinder 原始网络

Fig. 2　Primitive network of Hailfinde

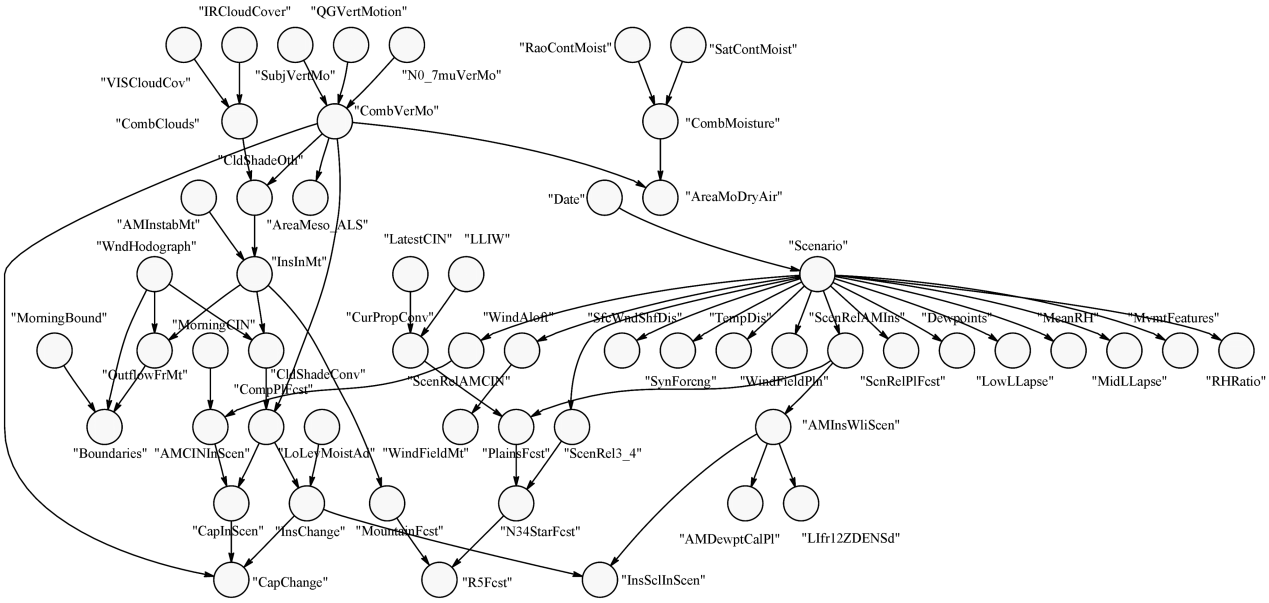

图 3 本文算法在 Hailfinder 数据集上求得的网络

Fig. 3 Network obtained by this algorithm on Hailfinder dataset

结束语 本文利用多头绒泡菌在觅食过程中保留重要觅食管道的特点,对原始搜索空间进行了缩减,在这个缩减后的结构空间中利用爬山法和拓扑排序得到了节点顺序,并将该顺序作为 K2 算法的输入,之后将本文算法与 LAGD 算法做比较。实验表明,本文算法在求得的网络结构准确性和对数据集的匹配程度上得到了更好的结果。然而本文对多头绒泡菌的运用还比较简单,如何能将该智能生物在构建高效交通网络上表现出的优良特性[18]更好地运用在贝叶斯网络构造问题上,是笔者今后研究的重点。

参 考 文 献

[1] THRUN S, MONTEMERLO M, DAHLKAMP H, et al. Stanley: The robot that won the DARPA Grand Challenge[J]. Journal of field Robotics, 2006, 23(9): 661-692.

[2] JURAFSKY D, MARTIN J H. Speech and language processing [M]. London: Pearson, 2014.

[3] CHICKERING D M, HECKERMAN D, MEEK C. Large-sample learning of Bayesian networks is NP-hard[J]. Journal of Machine Learning Research, 2004, 5(Oct): 1287-1330.

[4] KALISCH M, BÜHLMANN P. Estimating high-dimensional directed acyclic graphs with the PC-algorithm[J]. Journal of Machine Learning Research, 2007, 8(Mar): 613-636.

[5] MADSEN A L, JENSEN F, SALMERÓN A, et al. A parallel algorithm for Bayesian network structure learning from large data sets[J]. Knowledge-Based Systems, 2017, 100(117): 46-55.

[6] NIE S, DE CAMPOS C P, JI Q. Efficient learning of Bayesian networks with bounded tree-width[J]. International Journal of Approximate Reasoning, 2017, 80(C): 412-427.

[7] YUE K, FANG Q, WANG X, et al. A Parallel and Incremental Approach for Data-Intensive Learning of Bayesian Networks [J]. IEEE Transactions on Cybernetics, 2017, 45(12): 2890-2904.

[8] SUZUKI J. Learning Bayesian belief networks based on the minimum description length principle: an efficient algorithm using the B & B technique[C]// Proceedings of the Thirteenth International Conference on International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., 1996: 462-470.

[9] AKAIKE H. Information theory and an extension of the maximum likelihood principle [M] // Selected papers of hirotugu akaike. New York: Springer, 1998: 199-213.

[10] HECKERMAN D, GEIGER D, CHICKERING D M. Learning Bayesian networks: The combination of knowledge and statistical data[J]. Machine Learning, 1995, 20(3): 197-243.

[11] COOPER G F, HERSKOVITS E. A Bayesian method for the induction of probabilistic networks from data[J]. Machine Learning, 1992, 9(4): 309-347.

[12] AGHDAM R, GANJALI M, ZHANG X, et al. CN: a consensus algorithm for inferring gene regulatory networks using the SORDER algorithm and conditional mutual information test[J]. Molecular BioSystems, 2015, 11(3): 942-949.

[13] ABELLÁN J, CASTELLANO J. Improving the Naive Bayes classifier via a quick variable selection method using maximum of entropy[J]. Entropy, 2017, 19(6): 247.

[14] TABAR V R. A Simple Node Ordering Method for the K2 Algorithm based on the Factor Analysis[C]// International Conference on Pattern Recognition Applications and Methods. 2017: 273-280.

[15] NAKAGAKI T, YAMADA H, TO'TH A. Maze-solving by an amoeboid organism[J]. Nature, 2000, 407(6803): 470.

[16] TERO A, KOBAYASHI R, NAKAGAKI T. A mathematical model for adaptive transport network in path finding by true slime mold[J]. Journal of Theoretical Biology, 2007, 244(4): 553-564.

[17] ABRAMOVICI M, NEUBACH M, FATHI M, et al. Competing fusion for bayesian applications[C]// Proceedings of Information Processing and Management of Uncertainty. 2008: 379.

[18] TERO A, TAKAGI S, SAIGUSA T, et al. Rules for Biologically Inspired Adaptive Network Design[J]. Science, 2010, 327(5964): 439-442.