doi: 10.12068/j.issn.1005-3026.2020.02.002

基因调控网络的父节点筛选贝叶斯建模方法

曲璐渲[1], 郭上慧[1], 王之琼[1,2], 信俊昌[3,4]
(1. 东北大学 医学与生物信息工程学院, 辽宁 沈阳 110169; 2. 沈阳东软智能医疗科技研究院有限公司, 辽宁 沈阳 110179; 3. 东北大学 计算机科学与工程学院, 辽宁 沈阳 110169; 4. 辽宁省大数据管理与分析重点实验室, 辽宁 沈阳 110169)

摘 要: 在构建基因调控网络的方法中,贝叶斯网络模型可以直观地表达基因间的调控关系,但在结构学习时的复杂度极高,使得网络建模效率较低且规模有限.因此,本文提出一种基于父节点筛选的贝叶斯网络(parent node screening based Bayesian network, PS-BN)建模方法.PS-BN方法将关联模型与贝叶斯网络模型相结合,在充分利用贝叶斯网络模型结构学习搜索策略的前提下,先基于父节点筛选方法去除部分冗余信息,以达到缩减搜索空间的目的.实验结果表明,与传统的贝叶斯网络模型方法相比,PS-BN方法极大提升了基因调控网络构建效率,同时准确率有所提高.

关 键 词: 基因调控网络;父节点筛选;贝叶斯网络模型;关联模型;结构学习
中图分类号: TP 181 **文献标志码**: A **文章编号**: 1005-3026(2020)02-0158-05

Modelling of Gene Regulatory Networks by Parent Node Screening-Based Bayesian Method

QU Lu-xuan[1], *GUO Shang-hui*[1], *WANG Zhi-qiong*[1,2], *XIN Jun-chang*[3,4]
(1. School of Medicine & Biological Information Engineering, Northeastern University, Shenyang 110169, China; 2. Neusoft Research of Intelligent Healthcare Technology Co., Ltd., Shenyang 110179, China; 3. School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China; 4. Key Laboratory of Big Data Management and Analytics (Liaoning Province), Shenyang 110169, China. Corresponding author: WANG Zhi-qiong, E-mail: wangzq@bmie.neu.edu.cn)

Abstract: Among the methods for modeling gene regulation networks, Bayesian network model can intuitively express the regulatory relationship between genes. However, due to the high complexity of Bayesian network model in the structure learning, the efficiency of the gene regulation networks modeling is low and the scale of the reconstructed network is limited. Therefore, this paper proposed a method which is called the parent node screening-based Bayesian network (PS-BN). The PS-BN method combines the correlation model with Bayesian network model. Under the premise of making full use of the search strategy of structure learning in Bayesian network model, the parent node screening method is used to remove some redundant nodes, thus reducing the search space. The experimental results show that compared with the Bayesian network model, the PS-BN method greatly improves the efficiency of modeling gene regulatory networks while improving the accuracy.

Key words: gene regulatory networks; parent node screening; Bayesian network model; correlation model; structure learning

人类常见疾病,如神经退行性疾病、恶性肿瘤疾病等,究其原因都是由于基因表达异常所导致的结果.而基因并不是孤立存在的,基因间促进和抑制的调控关系形成了基因调控网络[1].通过构

收稿日期: 2019-04-17
基金项目: 国家自然科学基金资助项目(61472069,61402089); 中央高校基本科研业务费专项资金资助项目(N180101028, N180408019,N161602003,N160601001); 中国博士后科学基金资助项目(2018M641705,2019T120216); 沈阳东软智能医疗科技研究院有限公司开放课题基金资助项目(NRIHTOP1802).
作者简介: 曲璐渲(1987-),女,黑龙江哈尔滨人,东北大学博士研究生; 信俊昌(1977-),男,辽宁辽阳人,东北大学教授,博士生导师.

建相对精准的基因调控网络来研究基因间的表达关系,对人体机制和疾病治疗的探索有着重要意义. 而如何构建一个更加准确有效的网络成为基因调控网络研究的首要难题[2]. 目前,基因调控网络建模的研究方法主要包括:关联模型[3]、布尔网络模型[4]、微分方程模型[5]和贝叶斯网络模型[6]. 近年来,关联模型和贝叶斯网络模型广泛应用于基因调控网络的构建. 在关联模型方面,文献[7]基于自适应分块策略来估计互信息并构建基因调控网络. 在贝叶斯网络模型方面,Frolova 和 Wilczynski[8]提出了一种可以支持动态贝叶斯网络的 BNFinder2 软件来构建基因调控网络.

关联模型可以支持构建大规模的基因调控网络,但构建出的网络不能描述调控方向[9];而贝叶斯网络模型既可以描述调控关系又可以描述调控方向,但由于其计算复杂度很高,限制了网络构建的规模和效率[10]. 基于此,本文将两种模型的优点相结合,提出了基于父节点筛选的贝叶斯网络模型方法来构建基因调控网络. 首先,对于每个节点,利用皮尔逊相关系数计算节点间的关联程度,筛选出与该节点具有强相关性的若干节点;其次,将筛选出的节点作为贝叶斯网络模型中结构学习时的候选父节点以缩小搜索空间,并基于此构建基因调控网络. 最后,通过实验验证了该方法在准确率和计算效率上均优于传统的贝叶斯网络模型,并且该方法可以支持大规模基因调控网络的构建.

1　基于父节点筛选的贝叶斯网络建模

1.1　总体框架

贝叶斯网络(Bayesian network, BN)[10]借助有向无环图来刻画基因之间的依赖关系. 对于任一变量 X_i,通常可以找到一个与 X_i 都不独立的最小子集 $\mathrm{Parent}(X_i) \subseteq \{X_1, X_2, \cdots, X_{i-1}\}$,使得 $P(X_i \mid X_1, X_2, \cdots, X_{i-1}) = P(X_i \mid \mathrm{Parent}(X_i))$. 因此,当网络变量元组 $\langle X_1, X_2, \cdots, X_n \rangle$ 赋予具体数据值 $\langle x_1, x_2, \cdots, x_n \rangle$ 时,贝叶斯网络的联合概率分布为

$$P(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} P(x_i \mid \mathrm{Parent}(X_i)). \quad (1)$$

建立贝叶斯网络通常需要 2 个步骤:结构学习和参数学习. 结构学习的方法是针对每个节点,遍历并通过评分函数评价所有可能的结构,进而找出最好的结构作为该节点的父节点集;该方法的寻优策略可以保证所构建出的网络结构的精确性,然而,其复杂度也极高. 因此,利用传统贝叶斯网络模型的结构学习构建基因调控网络,不仅效率非常低且网络规模也有限. 本文提出一种基于父节点筛选的贝叶斯网络建模方法,该方法可以缩小结构学习的搜索空间,同时,充分利用结构学习搜索策略的优势保证所构建的基因调控网络的精确性. 网络总体框架如图 1 所示.

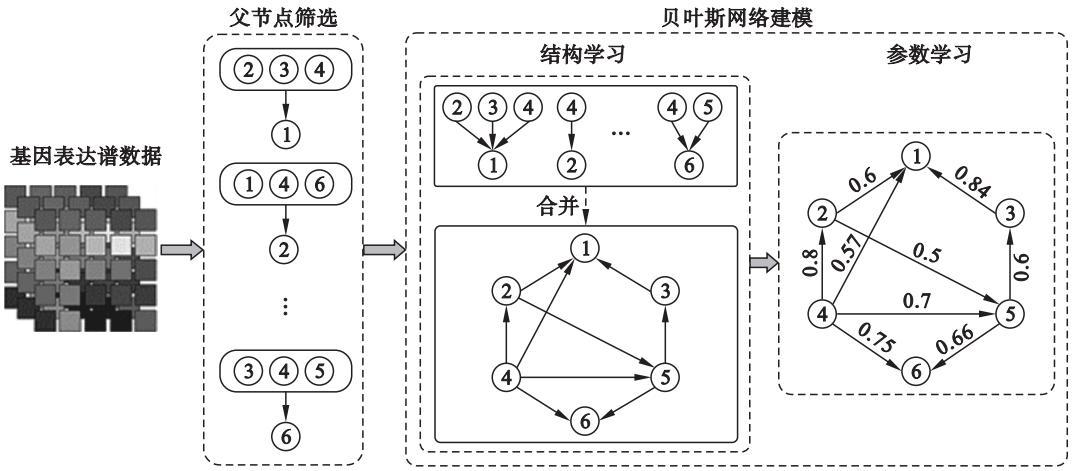

图 1　基于父节点筛选的贝叶斯网络建模总体框架

Fig. 1　Overall framework of modeling by parent node screening-based Bayesian network

由图 1 可以看到,基于父节点筛选的贝叶斯网络建模方法在贝叶斯网络结构学习前,先利用皮尔逊相关系数方法按照节点间的关联程度,针对每个节点筛选出若干个候选父节点,然后,将这些父节点作为结构学习时的搜索空间,但不改变结构学习的搜索策略,从而在结构学习过程中准确、高效地得出父节点集;经过结构学习后即可得到贝叶斯网络模型所构建出的基因调控网络. 最后,通过参数学习得到节点间边的权重,进而得出既有调控方向又有概率值的基因调控网络.

1.2 父节点的筛选方法

父节点筛选方法采用皮尔逊相关系数对节点间的关系进行描述. 皮尔逊相关系数衡量了两个变量 X_i 和 Y_i 的相似度,可由式(2)定义:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2}\sqrt{\sum (Y_i - \bar{Y})^2}}. \quad (2)$$

式中 $\bar{X}$ 和 $\bar{Y}$ 分别表示两个变量 X_i 和 Y_i 的平均值.

根据式(2)得到两个节点的相关程度后,将具有一定相关程度的父节点作为候选父节点. 基于父节点筛选的贝叶斯网络建模过程如算法 1 所示.

算法 1 基于父节点筛选的贝叶斯网络建模方法

输入:基因表达数据集合 $\{X_i\}_{i=1}^{n}$,父节点筛选比例 α,父节点集个数 m.

输出:基因调控网络矩阵 $\boldsymbol{G}$.

算法描述:

1) 父节点筛选个数 $\lambda = \alpha \times n$;
2) for $i = 1$ to n do
3) $\{Y_j\}_{j=1}^{n-1} = \{X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n\}$;
4) for $j = 1$ to $n - 1$ do
5) 计算 X_i 和 Y_j 的平均值 $\bar{X}, \bar{Y}$;
6) 根据式(2)计算皮尔逊相关系数 r_j;
7) 将 $[r_j, Y_j]$ 添加到父节点集合 P_{father};
8) 对 $P_{\text{father}} \in \{\text{Pa}_k\}_{k=1}^{n-1}$ 根据 r_j 由大到小排序;
9) for $k = 1$ to λ do
10) 将 Pa_k 中的 Y_j 存入 X_i 的候选父节点集合 P_i 中;
11) 计算 X_i 父节点集为空的 BDE 分数 emptyscore;
12) 设置最优父节点集 optimal_set = $[X_i, [\ ]]$;
13) 设置最优父节点分数 optimal_score = emptyscore;
14) for $p = 1$ to m do
15) 计算可能的父节点组合数 $\text{pc} = \frac{\lambda!}{p!\ (\lambda - p)!}$;
16) for $q = 1$ to pc do
17) 在候选父节点集合 P_i 中计算父节点集组合 Pa_q 的 BDE 分数 score;
18) if score > optimal_score
19) optimal_set = $[X_i, [\text{Pa}_q]]$;
20) optimal_score = score
21) 将 [optimal _ set, optimal _ score] 保存到 G_i;
22) G_i 添加到 $\boldsymbol{G}$;
23) 输出 $\boldsymbol{G}$.

算法分两部分:候选父节点筛选过程(第 1 ~ 10 行)和贝叶斯调控网络结构学习(第 11 ~ 23 行). 在父节点筛选过程中,先根据输入的父节点筛选比例计算父节点筛选个数 λ(第 1 行);其次,取基因表达数据集合 $\{X_i\}_{i=1}^{n}$ 中的一个节点 X_i,将该节点以外其余所有节点设置为 $\{Y_j\}_{j=1}^{n-1}$(第 2 ~ 3 行);再次,计算节点 X_i 与其余所有节点的皮尔逊相关系数,并将相关系数及其表达数据存储到父节点集合 P_{father} 中(第 4 ~ 7 行);最后,根据皮尔逊相关系数的取值对 P_{father} 中的行从大到小进行排序,并取前 λ 行中的数据 Y_j 作为 X_i 节点所对应的候选父节点集合(第 8 ~ 10 行);在贝叶斯网络结构学习过程中,根据 X_i 节点对应的候选父节点及其基因表达数据,先计算父节点集为空时的 BDE 分数,并且将该结构及其对应的 BDE 分数作为 X_i 节点的父节点集的最优组合和最优分数初始值(第 11 ~ 13 行);然后,根据输入的父节点集个数和父节点筛选个数 λ,计算从 1 个父节点作为父节点集到 m 个父节点作为父节点集的所有组合数,即为可能的父节点组合数(第 14 ~ 15 行);接着,依次在父节点筛选时所筛选的集合 P_i 中计算所有可能父节点的 BDE 分数并与最优分数进行比较,并根据计算结果更新最优组合和最优系数,直至得到 X_i 节点最终的最优组合和最优分数并保存到 G_i 中(第 16 ~ 22 行);最后,重复上述所有步骤,计算出所有节点的结构及其 BDE 分数并输出基因调控网络矩阵 $\boldsymbol{G}$.

2 实验结果及分析

2.1 实验设置

实验选用的数据为 GeneNetWeaver 上获取的大肠杆菌基因调控网络. 选取了其中 20 个基因的子网络,对父节点筛选的贝叶斯网络模型(PS-BN)和传统的贝叶斯网络模型(BN)的性能进行评价;还分别选取了包含 100,200,300,400 和 500 个基因的子网络对 PS-BN 的大规模基因调控网络构建效率进行评价.

在对 PS-BN 和 BN 进行性能评价时,选用了准确率、精确率、召回率、F 值 4 项评估指标,另外还对两种方法的运行时间进行了比较,验证了 PS-BN 的效率.

准确率描述了构建基因调控网络时判断为有边或无边的总体的准确率，可由式(3)计算：

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}. \tag{3}$$

精确率描述了判断为有边的结果中，预测正确的概率，可由式(4)计算：

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{4}$$

召回率描述的是在所有标签为有边的金标准中，实验结果也判断为有边的概率，可由式(5)计算：

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{5}$$

F 值是一个综合评价指标，表示的是精确率和召回率的调和平均评估指标，可由式(6)计算：

$$F=\frac{2\times\mathrm{recall}\times\mathrm{precision}}{\mathrm{recall}+\mathrm{precision}}. \tag{6}$$

上述评估指标中，TP 表示调控网络真实为有边，构建结果也为有边；TN 表示真实为无边，结果也为无边；FP 表示真实为无边，结果却为有边；FN 表示真实为有边，结果却为无边.

2.2　实验结果

将父节点筛选比例分别设置为总节点个数的 10% ,20% ,30% ,40% 和 50% ,BN 和 PS-BN 两种方法的评估指标与运行时间的实验结果对比如表 1 所示. 可以看出，对于 PS-BN 方法，父节点筛选比例高的评估指标优于筛选比例较低的；当筛选比例大于或等于 30% 时，各项评估指标值相同. 与 BN 方法相比，PS-BN 方法筛选比例大于或等于 20% 时的实验结果均优于 BN 方法. 在运行时间方面，各个比例的 PS-BN 方法的运行时间均远远少于 BN 方法. 因此，PS-BN 方法在保证评估指标优于 BN 方法的前提下，计算效率得到了显著提高.

表 1　评估指标与运行时间比较

Table 1　Comparison of the evaluation indices and the operating time

方法	父节点筛选比例/%	准确率	精确率	召回率	F 值	运行时间/s
BN		0.971	0.4	0.444	0.421	10 500
PS-BN	10	0.971	0.33	0.222	0.264	2
PS-BN	20	0.982	0.583	0.778	0.667	4
PS-BN	30	0.984	0.615	0.889	0.727	20
PS-BN	40	0.984	0.615	0.889	0.727	100
PS-BN	50	0.984	0.615	0.889	0.727	355

在准确率方面，BN 方法与筛选比例为 10% 的 PS-BN 方法的准确率相同. 因为，在经过父节点筛选后，贝叶斯网络搜索空间中的绝大部分冗余信息都被过滤掉，因而假阳边非常少；但由于筛选比例过小而导致部分真阳边也被过滤掉. 因此，准确率是在牺牲真阳边数量的前提下得以保证的. 筛选比例为 20% 和 30% 时，准确率逐渐升高，且均高于 BN 方法. 随着筛选比例的增高，搜索空间中对应的真阳边越来越多，使得所构建的基因调控网络真阳边也增加，而大部分冗余信息仍会被过滤掉；因此在保证真阳边数量的前提下，假阳边的比例少于 BN 方法，从而提升其准确率. 而继续增加父节点筛选比例后，由于有效节点已经被筛选进搜索空间，而增加少部分冗余信息并没有使假阳边增多，因此，其准确率与 30% 时相比不变.

精确率和召回率的评估考虑的均是真阳边的预测情况. 当 PS-BN 方法的父节点筛选比例设置为 10% 时，精确率和召回率均低于 BN 方法，说明该比例所得到的真阳边数量少于 BN 方法. 当筛选比例增加至 20% 及以上时，精确率和召回率得到提升并高于 BN 方法，也就是说，筛选比例增加，真阳边数量也随之增多，并且超过了 BN 方法所得到的真阳边. 当筛选比例高于 30% 时，真阳边数量不再增加，并且网络结构没有变化，因此，精确率和召回率保持不变.

评估指标 F 值是平衡精确率和召回率的一项综合指标，因此该指标与精确率和召回率的相关程度较高，F 值的变化趋势与精确率和召回率的变化趋势一致. 筛选比例为 10% 的 PS-BN 方法在精确率和召回率上均低于 BN 方法，因此其 F 值也较低. 当筛选比例为 20% 和 30% 时，两项评估指标均高于 BN 方法，因而其 F 值也所有提高. 当筛选比例为 40% 和 50% 时，两项评估指标均与 30% 相同，故其 F 值也与 30% 相同.

在运行时间方面，PS-BN 方法均优于 BN 方法. 因为父节点经过筛选后，贝叶斯网络的搜索空间大幅缩小，在搜索空间中进行父节点集遍历搜

索的结构学习时,搜索的节点数量和各节点的组合数目都大量减少,因此,PS-BN 方法的运行时间大大缩短. 随着筛选比例的增加,贝叶斯网络模型的搜索空间逐渐增大,不但导致需要搜索的节点数量增加,还导致需要进行评分计算的组合数目大量增多,使结构学习耗费的时间呈指数增长. 因此,随着筛选比例的增加,运行时间也呈指数增长.

由表 1 可知,在评估指标值较高的筛选比例中,比例为 30% 时运行时间最短. 因此,在对大规模基因调控网络的构建进行效率评估时,选用的筛选比例为 30% . 大规模基因调控网络构建运行时间如图 2 所示,其中带标记的实线为运行时间,虚线表示运行时间增加的趋势线. 由图 2 可知,随着基因数量的增加,运行时间呈幂增长趋势. 这是由于基因数量增加,同样的筛选比例所筛选出的基因数量也随之增加;但由于筛选比例较小,基因数量增加有限,因此,在该搜索空间进行结构学习时,运行时间呈幂增长趋势.

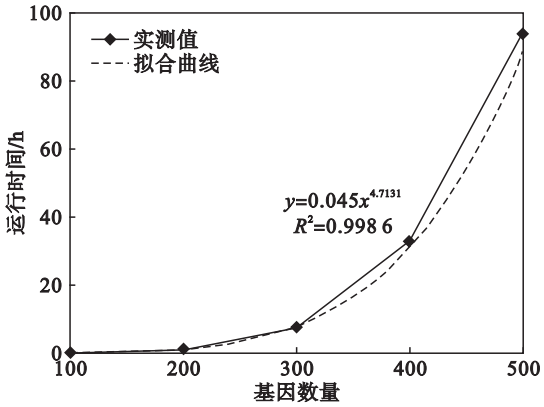

图 2 大规模基因调控网络构建的运行时间

Fig. 2 Operating time for modeling large-scale gene regulatory networks

3 结 语

为解决贝叶斯网络模型在构建基因调控网络时效率低且构建规模有限的问题,本文提出了基于父节点筛选的贝叶斯网络(PS-BN)建模方法. PS-BN 方法充分利用传统贝叶斯网络建模方法在结构学习中的搜索策略,同时,通过筛选父节点缩小了搜索空间且去除了部分冗余信息,从而在极大提高效率的同时,准确率等 4 项评估指标也均有所提升. 实验证明了上述结论.

参考文献:

[1] Banf M, Rhee S Y. Computational inference of gene regulatory networks: approaches, limitations and opportunities [J]. *Biochimica et Biophysica Acta*, 2017, 1860 (1): 41 – 52.

[2] 王沛,吕金虎. 基因调控网络的控制:机遇与挑战[J]. 自动化学报,2013,39(12):1969 – 1979.
(Wang Pei, Lyu Jin-hu. Control of genetic regulatory networks: opportunities and challenges [J]. *Acta Automatica Sinica*, 2013, 39(12): 1969 – 1979.)

[3] Liu A N, Wang L L, Li H P, et al. Correlation between posttraumatic growth and posttraumatic stress disorder symptoms based on Pearson correlation coefficient: a meta-analysis[J]. *Journal of Nervous and Mental Disease*, 2017, 205(5): 380 – 389.

[4] Dorier J, Crespo I, Niknejad A, et al. Boolean regulatory network reconstruction using literature based knowledge with a genetic algorithm optimization method [J]. *BMC Bioinformatics*, 2016, 17(1): 410.

[5] Shamarova E, Chertovskih R, Ramos A F, et al. Backward-stochastic-differential-equation approach to modeling of gene expression[J/OL]. *Physical Review E*, 2017, 95(3): 032418 [2019 – 02 – 25]. https://doi. org/10. 1103/PhysRevE. 95. 032418.

[6] Gendelman R, Xing H, Mirzoeva O K, et al. Bayesian network inference modeling identifies TRIB1 as a novel regulator of cell-cycle progression and survival in cancer cells [J]. *Cancer Research*, 2017, 77(7): 1575 – 1585.

[7] Lachmann A, Giorgi F M, Lopez G, et al. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information[J]. *Bioinformatics*, 2016, 32 (14): 2233 – 2235.

[8] Frolova A, Wilczyński B. Distributed Bayesian networks reconstruction on the whole genome scale[J/OL]. [2019 – 03 – 10]. https://doi. org/10. 7717/peerj. 5692.

[9] Barman S, Kwon Y K, Enrique H L. A novel mutual information-based Boolean network inference method from time-series gene expression data[J/OL]. *PLoS ONE*, 2017, 12(2): e0171097 [2019 – 03 – 10]. https://doi. org/10. 1371/journal. pone. 0171097.

[10] 赵建喆,李凯. 一种改进的多模块贝叶斯网络局部推理算法[J]. 东北大学学报(自然科学版),2015,36(9):1251 – 1255.
(Zhao Jian-zhe, Li Kai. An improved local inference algorithm for multiply sectioned Bayesian networks [J]. *Journal of Northeastern University* (*Natural Science*), 2015, 36(9): 1251 – 1255.)