基于贝叶斯网络SP算法的改进研究

奚海荣[1],马文丽[2],梁 斌[1]
(1.上海大学 电子生物中心,上海 200072;
2.南方医科大学 基因工程研究所,广东 广州 510515)

摘 要:针对SP算法中利用优化组合处理稀疏候选集来评分得最优候选集,这样得到的每个节点的候选集为父节点集,从而容易导致最后的贝叶斯网络双向边较多,对双向边处理后还存在较多的反向边,从而提出了利用爬山算法处理稀疏候选集,得到新的算法SCHC,该算法减少了双向边的数量和提高了正确边的数量。

关键词:SP算法;稀疏候选集;贝叶斯网络;爬山算法;双向边
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2009)03-0155-03

Improvement of SP Algorithm Based on Bayesian Networks

XI Hai-rong[1], MA Wen-li[2], LIANG Bin[1]
(1. Electronic Biology Technology Research Center, Shanghai University, Shanghai 200072, China;
2. Institute of Genetic Engineering, Nanfang Medical University, Guangzhou 510515, China)

Abstract: SP algorithm used optimization to deal with sparse candidate sets, scored the optimal set as the candidate, so are the candidates for each node-parent node sets, thus easily lead to the final Bayesian network had more two-way edges, after dealed with two-way edge, there are still more reverse side. And put forward using hill-climing algorithm to deal with sparse candidate sets, getting new algorithm SCHC. And the algorithm has reduced the number of two-way edges and increased the number of correct side.

Key words: SP algorithm; sparse candidate sets; bayesian networks; hill-climing algorithm; Two-way edges

0 引 言

贝叶斯网络理论将先验知识与样本信息相结合、依赖关系与概率表示相结合,是数据挖掘和不确定知识表示的理想模型[1]。因此,贝叶斯网络是目前用于生物信息学上的一种重要的方法。研究最多的是从基因表达谱(gene expression profiles)推断和识别基因网络,主要包括:从表达数据识别基因调控网络结构;通过随机扰动,分析个体基因对全局动态网络性能的影响得出网络特性;根据大规模的数据进行基因网络分析,识别基因网络中的调控关系,获得网络参数,推断网络特征[2];通过建立静态网络,推断网络中基因之间在稳态下的相互作用机制。

由于贝叶斯网络模型对于推断基因调控网络有其自身的优越性,所以其是现在研究的热点之一。但是由于经典的贝叶斯学习算法不能直接应用到基因表达数据上,所以目前针对基因表达数据本身特点的算法需要不断提出和改进。目前在静态贝叶斯网络用于基因调控网络的有Friedman提出的SC(Sparse Candidate)算法[3],这次第一个应用于有几百个数据集的算法;有Ioannis Tsamardinos提出的MMPC(Max-Min Parents and Children)算法[4]及改进算法MMHC(Max-Min Hill-Climbing)算法[5];以及由杨庆平提出的SP(Sparse CandidateParents)算法[6],该算法是先确定每个节点的候选稀疏节点集合,选择方法多样,原文采用互信息方法,接着采用相应的优化组合算法和评分函数来找到确定的最终的前驱节点集,就得到了相应的基因调控网络。

在SP算法的基础上,利用爬山搜索[7]对SP算法进行改进,提出了一种新的算法SCHC(Sparse Candidate Hill-Climbing)算法。

收稿日期:2008-06-23
基金项目:国家自然科学基金(39880032)
作者简介:奚海荣(1982-),男,江苏张家港人,硕士研究生,研究方向为人工智能、生物信息学;马文丽,博士生导师,研究方向为生物信息学、基因芯片;梁 斌,高级工程师,研究方向为结构力学。

1 互信息与评分函数

1.1 互信息

基于信息熵的互信息一般是测定变量之间的独立性,来确定两个变量的关系。两个变量之间互信息测

试为：

$$I(A,B)=\sum_{a,b}P(a,b)\log\frac{P(a,b)}{P(a)P(b)} \tag{1}$$

和条件互信息，相应于一组条件集 C：

$$I(A,B\mid C)=\sum_{a,b,c}P(a,b,c)\log\frac{P(a,b\mid c)}{P(a\mid c)P(b\mid c)} \tag{2}$$

变量 A 和 B 之间的互信息表示在观测到 A 的值后，从 B 处能获得的期望信息量。在贝叶斯网络中，如果两个节点是依赖的，在已知一个节点的值后，将给出另一个节点值的相关信息。因此，互信息可以说明两个节点是否是依赖的和其关系的程度。

1.2 评分函数

贝叶斯网络结构学习过程中需要对候选结构评分，根据不同的评分值选择最优的结构。评分函数有贝叶斯后验概率、最小描述长度和 Kullback－Leiber 熵等。

变量 $X_1,X_2,\cdots,X_n$，$x_i\in\{x_i^1,\cdots,x_i^{r_i}\}$，$r_i$ 为取值个数，π_{xi} 为 X_i 的父节点集 Π_{X_i} 的所有节点值的排列，排列的情况数量为 $q_i=\prod_{X_i\in\Pi_{X_i}}r_i$，$S$ 为取某一网络结构的假设或事件。

现在介绍一种基于似然函数的贝叶斯—迪里赫列(Bayesian Dirichlet—BD)评分标准：

$$P(D\mid S_B)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N'_{ij})}{\Gamma(N_{ij}+N'_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N_{ijk}+N'_{ijk})}{\Gamma(N'_{ijk})} \tag{3}$$

其中 N_{ijk} 是在数据库 D 中满足 $x_i=x_i^k$，且 $\pi_{x_i}=j$ 的情况数量 $N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$。

其中 $N'_{ijk}>0$ 为先验分布指系数(或超级参数)，$N'_{ij}=\sum_{k=1}^{r_i}N'_{ijk}$。

2 SP 算法及算法的改进

文献[6]提出了基因互信息的稀疏候选算法(Sparse Candidate Parents)，此算法是基于打分－搜索的贝叶斯网络结构学习算法和基于依赖分析的方法的两类算法[8]的杂交算法，首先采用互信息方法确定每个节点的候选稀疏节点集，互信息表示两个节点之间的信息依赖程度，值越大说明距离越应该近。接着采用优化组合，运用评分函数来找到确定的最终的前驱节点集，就得到了相应的基因调控网络[6]。

在 SP 算法的基础上，重新对其进行改进，对于 SP 算法中利用优化组合算法和评分函数来选择父亲节点，每次处理只能够选择每个节点的父节点集，从而容易导致最后的贝叶斯网络双向边较多，对双向边处理后还存在较多的反向边，从而提出了利用爬山算法处理稀疏候选集，同时 SP 算法在数据多的情况下会显的关系单一，最后遗漏和错误率会很高，文中的改进方法是直接对每个稀疏候选节点集用爬山法搜索，进行最优化处理，取最优解，把最优的解存入矩阵 ***SH*** 中，***SH*** 矩阵的 i 行 j 列表示 i 是 j 的父亲节点，然后对候选集依次处理，每次处理后都加入矩阵 ***SH*** 中，不断得到新的矩阵 ***SH***，直到处理完成。如果出现双向边再按照 SP 算法中的过程处理，直到处理完，得到的一个矩阵表示基因调控网络。

算法：SCHC 算法

输入：基因表达数据矩阵 ***D***，依赖性阈值 ε，最大候选个数 n

输出：节点相连的网络图

1. load $\boldsymbol{D}=\{D_1,D_2,D_3,\cdots,D_n\}$
2. initialize SC /* SC 为候选节点集 */
for each D_i
3. $SC_i=I(D_i,D_j)$
/* $i\neq j$，SC_i 中的值从大到小排列，且取最大的前 n 个数 */
4. repeat
for each SC_i
5. HillCliming();
/* 利用爬山法进行搜索 */
6. until 得到最优解，处理得新的矩阵 ***SH*** /* 矩阵 ***SH*** 不断更新 */
7. repeat
8. if(Correlation(D_i,D_j)=true)
/* Correlation(D_i,D_j)表示 D_i 和 D_j 双向边 */
if(Score(D_i)>Score(D_j))
/* Score(D_i)表示 D_i 有父节点和无父节点的评分分数之差 */
$D_i\rightarrow D_j$
else $D_j\rightarrow D_i$
9. return ***SH***

计算互信息阶段(第三步)，使用频率代替概率，使用式(1)，但并不保存所有节点对的计算值，而是在大于一个给定的阈值 ε 下，才保存相应的值，此阈值的选择可以根据实验得到，这个阈值对结果影响很大，所以通过多次实验获取，目前一般选 0.01，当大于此值说明两个节点是相连的。

用爬山法搜索(第五步)，目标是找到评分最高的模型，它从一个初始模型出发开始搜索，初始模型是一般为无边模型，在搜索的每一步都用搜索算子对当前

模型进行局部修改,得到一系列候选模型,计算每个候选模型的分数,使用式(3),并将最优候选模型与当前模型比较;若最优候选模型分数大,则以它作为下一个当前模型继续搜索;否则停止搜索,返回当前值,其中搜索算子有 3 个:加边、减边、转边。

在存在双向边的情况下,处理双向边(第八步),文中采用文献[6]中的净增益思想,这不是简单地考虑一种度量标准来判断边的方向。由于贝叶斯网络符合马尔科夫假设,因此,评分函数是可分解的,即分解为每一个节点和其父亲集的分数之和。在判断边的方向时,例如,两个节点 A 和 B 是双向的,如何判别方法判断出哪个方向是对的,首先计算 A 和 B 的没有父亲的分数,即 score1 和 score2,此时的分数表示 A 和 B 不需父亲的度量,然后计算 A 包含所有父亲的分数(包括 B) 即 score3,此时的分数 A 表示需要 B 作为父亲的度量,同样的对 B 计算得到 score4 净增益是指,需要父亲的度量减去不需要父亲的度量,这样就能更好地反映出方向的特征。如果 A 大于 B,则认为 A 需要 B 作为父亲的净增益大,因此,此时的判断被接受,相反,则认为 B 需要 A 作为父亲的净增益大。这样在一定程度上就能很好地处理双向边的问题。

3　实验结果与分析

为了评价算法的好坏,需要一组已知结构的网络。然后,由于基因调控网络研究处在初步阶段,还没有现有的完全已知的网络结构,这就有必要采用模拟网络。

3.1　算法评价机制

文中在使用模拟数据进行推断算法评价时,把已知的网络结构被称为目标网络,使用敏感性(sensitivity),差异性(specificity)和 F - factor。敏感性是测试目标网络被推断出的程度,差异性表示算法的精确度。F - factor 是敏感性和差异性的平衡,定义为两个量的调和平均值。F - factor 值越大,就越能够说明算法推断正确率高。

$$\text{Sensitivity}=\frac{\text{结果网络与目标网络相同边的数目}}{\text{目标网络中边的数目}} \tag{4}$$

$$\text{Specificity}=\frac{\text{结果网络与目标网络相同边的数目}}{\text{目标网络中边的数目}} \tag{5}$$

$$\text{F}-\text{factor}=\frac{2*\text{Sensitivity}*\text{Specificity}}{\text{Sensitivity}+\text{Specificity}} \tag{6}$$

使用真实的生物数据时,由于真实的生物调控网络没有完全未知,所以不能使用上面定义的方法来评价推断结果,这给真实生物结果评价带来了很大的困难,目前的评价方法是查询现有的已经被证实的调控关系来评价推断出的结果。因此,对基因调控网络推断算法的评价主要是集中在模拟真实的调控网络来评价,即使用模拟网络来评价。

3.2　模拟数据

文中用到的数据来自 http://www.dsl-lab.org/supplements/mmhc_paper/mmhc_index.html,使用的数据为 ALARM 网络数据,此网络是由贝叶斯网络专家构建的诊断预备系统,具有 37 个离散变量,每个变量的值在 2~4 之间,具有 46 条边。此网络数据样本有 5000 个。

在实验中,$\varepsilon = 0.01$,由于变量只有 37 个,所以 $n = 37$,程序执行结果见表 1。

表 1　SP 算法与 SCHC 算法的比较

算法	正确边数	敏感性(%)	差异性(%)	F - factor	方向相反的边	关系正确率
SP	35	0.76	0.83	0.79	7	0.91
SCHC	38	0.83	0.93	0.87	3	0.89

方向相反的边是与原网络图中两个变量的关系相反,关系正确率是正确边数与反向边数之和与所有边数的比值。方向相反说明两变量的关系能够确定了,方向错说明关系确定,但因果关系错误。

从表 1 中发现,改进的算法不管从敏感度、差异性还是正确边数,都优于原先的 SP 算法,关系正确率略低于 SP 算法,其原因是可能在爬山搜索时,陷入了局部最优,使得返回的边数少于正确的结构。

本实验数据均在 4 - M CPU,608M 内存,WINDOWS XP 的操作环境下,用 Matlab 7 实现的。

4　结束语

贝叶斯网络模型应用于基因调控网络的推断,贝叶斯网络在处理基因表达数据时,有其很好的统计性和处理隐变量以及缺失值的处理等优越性,是未来发展的重点方向。从利用爬山法取代利用优化组合算法来处理稀疏候选集,提出了新的算法 SCHC,实验证明改进算法优于 SP 算法。下一步工作是拓展到大数据集、真实的基因数据中去,还可以不断改进算法来提高准确率和降低计算复杂度。

参考文献:

[1] Cowell R G, Dawid A P, Lauritzen S L, et al. Probabilistic Networks and Expert Systems[M]. [s.l.]: Springer, 1999.

[2] Wessels L F A, vanSomeren E P, Reinder M J T. A comparison of genetic network models[C]//In: Pacific Symposium on Biocomputing. Belligham: ETATS - UNIS, 2001: 508 - 519.

(下转第 192 页)

下面是在 ADS 下，一种常用的实现模型：

```
IMPORT |Image$$RO$$Base| ;RO 区开始地址
IMPORT |Image$$RO$$Limit| ;RO 区末地址后面的地址即 RW 数据源起始地址
IMPORT |Image$$RW$$Base| ;RW 区在 RAM 里的执行区起始地址
IMPORT |Image$$RW$$Limit| ;RW 区末地址后面的地址
IMPORT |Image$$ZI$$Base| ;ZI 区在 RAM 里面的起始地址
IMPORT |Image$$ZI$$Limit| ;ZI 区在 RAM 里面结束地址后面的一个地址
LDR r0,=|Image$$RO$$Limit|
LDR r1,=|Image$$RW$$Base|
LDR r3,=|Image$$ZI$$Base|
CMP r0,r1
BEQ %F1
0 CMP r1,r3
LDRCC r2,[r0],#4
STRCC r2,[r1],#4
BCC %B0
1 LDR r1,=|Image$$ZI$$Limit|
MOV r2,#0
2 CMP r3,r1
STRCC r2,[r3],#4
BCC %B2
```

程序实现了 RW 数据的拷贝和 ZI 区域的清零功能。程序先把 ROM 里以|Image$$RO$$Limt|开始的 RW 初始数据拷贝到 RAM 里面|Image$$RW$$Base|开始的地址，当 RAM 这边的目标地址到达|Image$$ZI$$Base|后就表示 RW 区的结束和 ZI 区的开始，接下去就对这片 ZI 区进行清零操作，直到遇到结束地址|Image$$ZI$$Limit| 为止。

3.5 跳转到主应用程序

当系统初始化工作完成之后，就需要把程序流程转入主应用程序。

最简单的一种情况是直接跳到自定义的主函数，函数名由用户定义，例如：

```
IMPORT Main
BL Main
```

在 ARM ADS 环境中，还另外提供了一套系统级的呼叫机制。

```
IMPORT__main
BL__main
```

其中__main()是编译系统提供的一个函数，负责完成库函数的初始化和初始化应用程序执行环境，最后自动跳转到 main()函数。这种情况下用户程序的主函数名必须是 main。

至此，系统软硬件已进入到合适的状态，用户可以进行主应用程序的开发了。

4 结束语

在嵌入式系统软件设计中，BootLoader 的编写至关重要。性能良好的 BootLoader，可以大大提高系统的实时性和稳定性。文中开发的 BootLoader 已成功应用于一款无线通信设备，所介绍的原理和设计与实现的方法，对设计和移植到其他类型的嵌入式系统有一定的参考价值。

参考文献：

[1] 万永波，张根宝. 基于 ARM 的嵌入式系统 Bootloader 启动流程分析[J]. 微计算机信息，2005，21(11—2)：90－92.

[2] 李亚锋，欧文盛. ARM 嵌入式 Linux 系统开发从入门到精通[M]. 北京：清华大学出版社，2007：48－84.

[3] ATMEL Corporation. ARM920T－based Microcontroller AT 91 RM9200 Guide[M]. [s.l.]：[s.n.]，2006：83－94.

[4] 潘 浩，马艳敏. Bootloader 在 AT91RM9200 系统中的实现[J]. 微计算机信息，2007，23(1－2)：168－170.

[5] 王宇行. ARM 程序分析与设计[M]. 北京：北京航空航天大学出版社，2008：213－232.

(上接第 157 页)

[3] Friedman N, Nachman I, Pe'er D. Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm[C]//In: Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI). [s. l.]: Morgan Kaufmann, 1999：206－215.

[4] Tsamardinos I, Aliferis C F, Statnikov A. Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations[C]//In: The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003). [s.l.]: ACM, 2003：673－678.

[5] Tsamardinos I, Brown L E, Aliferis C F. The Max－Min Hill－Climbing Bayesian Network Structure Learning Algorithm [J]. Machine Learning, 2006, 65：31－78.

[6] 杨庆平. 基于贝叶斯网络的基因调控网络构建算法的研究[D]. 哈尔滨：哈尔滨工业大学，2006.

[7] 张连文，郭海鹏. 贝叶斯网引论[M]，北京：科学出版社，2006：186－188.

[8] 张剑飞. 贝叶斯网络学习方法和算法研究[D]. 长春：东北师范大学，2005.