基于贝努里分布的贝叶斯网络结构学习算法*)

孙 岩[1,2,3] 吕世聘[3] 唐一源[2]

(辽宁师范大学计算机与信息技术学院 大连 116029)[1]

(大连理工大学神经信息学研究所 大连 116023)[2] (大连理工大学计算机科学与工程系 大连 116023)

摘 要 目前,学习具有丢失数据的贝叶斯网络结构主要采用结合 EM 算法的打分-搜索方法和基于依赖分析的思想,其效率和可靠性比较低。本文针对此问题建立一个新的具有丢失数据的贝叶斯网络结构学习算法。该方法首先根据贝努里分布来表示数据库中变量结点之间的关系,并用 Kullback-Leibler(KL)散度来表示同一结点的各个案例之间的相似程度,然后根据 Gibbs 取样来得出丢失数据的取值。最后,用启发式搜索完成贝叶斯网络结构的学习。该方法能够有效避免标准 Gibbs 取样的指数复杂性问题和现有学习方法存在的主要问题。

关键词 贝努里分布,KL 散度,贝叶斯网络,Gibbs 取样

Learning Bayesian Network Structure Based on Bernoulli Distribution

SUN Yan[1,2,3] LU Shi-Pin[3] TANG Yi-Yuan[2]

(Computer Science Department of Liaoning Normal University, Dalian 116029)[1]

(Neuroinformatics Institute of Dalian University of Technology, Dalian 116023)[2]

(Computer Science Department of Dalian University of Technology, Dalian 116023)[3]

Abstract At present, the algorithm of learning bayesian structure with missing data is mainly based on the search and scoring method combined with EM algorithm. The algorithm has low efficiency. In this paper , a new algorithm of learning Bayesian network structure with missing data is presented. First, we adopt the Bernoulli distribution to express the relationship between the variables in database. Second, we use KL divergence to express the similarity between the cases. Third, we draw the value of the missing data according to the Gibbs sampling. Finally, we use heuristical search to complete the learning of Bayesian network structure. This method can avoid the exponential complexity of standard Gibbs sampling and the main problems in the existing algorithm.

Keywords Bernoulli distribution, Kullback-Leibler divergence, Bayesian network, Gibbs sampling

1 引言

贝叶斯网络推理能够处理不完备数据集,传统推理是无法解决的。对于传统的推理,必须知道所有可能的数据输入,如果缺少其中的某一输入,就会对建立的模型产生偏差。贝叶斯方法可以解决这个问题,因为贝叶斯网络反映的是整个数据域中数据间的概率关系,即使缺少某一数据变量,仍然可以建立精确的模型。贝叶斯网络是根据因果关系进行推理的。在数据分析处理中,获得变量域的理解是十分重要的,而且贝叶斯网络可以在缺少插入值的情况下进行决策。它说明了联合条件分布,允许在变量的子集之间定义类条件独立性并提供了一种因果关系图形,可以在其上学习并根据学习结果进行分类。贝叶斯网络和贝叶斯概率统计是紧密相关的,这就促进了依据知识和数据域的先验知识就可以建立正确的预测模型。由于贝叶斯网络具有语义的因果关系,因而可以直接地进行因果先验知识的分析,从而在贝叶斯网络中可以获得较全面的先验知识。

贝叶斯网络是描述随机变量之间依赖关系的图形模式,被广泛用于不确定性知识表示和推理,是处理不确定性问题的有力工具。

在各种实用的数据库中,丢失数据是指不能被观察到的变量值,属性值缺失的情况经常发生甚至是不可避免的。造成数据缺失的原因是多方面的,例如有些对象的某个或某些属性是不可用的。也就是说,对于这个对象来说,该属性值是不存在的,如一个未婚者的配偶姓名、一个儿童的固定收入状况等;或者在医疗数据库中,并非所有病人的所有临床检验结果都能在给定的时间内得到,致使一部分属性值空缺出来;或者因为获取这些信息的代价太大,尤其在有关脑功能实验的研究中,由于一些原因使得一些数据无法完整获得或缺失,出于实验成本的考虑,无法将有缺失的数据全部删除。

由以上造成数据丢失的原因可以看出,在建立和分析许多数据库系统中都存在着丢失数据的现象[1],所以本问题的提出,具有广泛的应用价值。

目前,具有完整数据的贝叶斯网络结构学习方法已经比较成熟,可分为两类:一类是基于依赖分析的学习方法,另一类是基于打分-搜索的学习方法。两类方法各有特点:依赖分析方法过程比较复杂,学习效率较低,一些算法通过增加假设限制条件来提高算法的学习效率[2];打分-搜索方法过程比较简单、规范,但由于搜索空间大,并且要求结点有顺序来降低算法的复杂性,根据打分函数的可分解性进行局部确定或随机搜索(完全搜索是 NP 困难问题[3]),效率较低,且易于陷入局部最优结构。因此,对于有丢失数据的贝叶斯网络结构学

*)国家自然基金项目(60472017,30670699)资助课题。孙 岩 讲师,博士生,研究方向:人工智能、计算机图形学;吕世聘 博士生,研究方向:数据挖掘、决策支持系统;唐一源 教授,博士生导师,研究方向:神经信息学。

习更加困难，现有的研究主要基于打分-搜索方法。

数据的丢失导致两方面的问题出现：一方面打分函数不再具有可分解形式，不能进行局部搜索；另一方面，一些充分统计因子不存在，无法直接进行结构打分。围绕这两个问题相继发展了一些解决的方法。Hecherman 等人[4~6]给出了一些解决后一个问题的方法。这些方法对选择的贝叶斯网络结构首先基于梯度的优化（gradient-based optimization）或用 EM（expectation-maximization）算法进行最大后验参数估计，然后使用拉普拉斯近似（Laplace approximation）或贝叶斯信息标准（Bayesian information criterion）等大样本近似方法进行近似结构打分。由于搜索空间大以及存在近似打分的误差，使其学习效率较低，且结果不够可靠。

Friedman 等人[7~9]改进了上述方法，基于 EM 算法框架进行具有丢失数据的贝叶斯网络结构学习，使用期望充分统计因子代替不存在的充分统计因子，在一些假设下，可使打分函数具有可分解形式（可进行局部搜索），并且在每一次迭代中，结构都有所改进，使结构序列收敛。该方法能够一定程度地提高学习效率，但因为 EM 算法是对分布参数的局部贪婪搜索，因此对初始值敏感，易于陷入局部极值，参数迭代还可能收敛到非似然函数极值的参数空间的边界，从而产生欺骗收敛，一般是收敛到局部最优结构。

本文提出了一种新的基于贝努里分布的具有丢失数据的贝叶斯网络结构学习 BSMB 方法（Learning Bayesian network Structure with Missing data based on Bernoulli distribution）。该方法首先根据 Bernoulli 分布来表示数据库中变量结点之间的关系，并用 Kullback-Leibler（KL）散度来表示不同案例之间的相似程度，然后根据 Gibbs 取样来获得丢失数据的值。该方法能够避免标准 Gibbs sampling 的指数复杂性问题和现有学习方法存在的主要问题。

文中用 $X_1,\cdots,X_n$ 表示离散随机变量，简称为变量，$x_1,\cdots,x_n$ 为其值。数据库 D 中有 N 个记录，假设数据是独立地产生于某一概率分布 P，数据丢失是随机的。在概率模式中的变量和表示概率模式的图形模式中的结点有时不加区分。

2 丢失值的插补

2.1 用贝努里分布表示变量结点

给定有缺失数据的数据库中结点变量为 $X=\{X_1,X_2,\cdots,X_N\}\subset R^3$，我们的目标是找到丢失数据的插补值，$Y=\{Y_1,Y_2,\cdots,Y_M\}\subset R^3$，其中 $M>N$，表示对应于 X 的插补值。对于每对点 X_i 和 $X_j(i\neq j)$，用 l_{ij} 来表示两点之间的关系，如相似关系、邻接关系和距离关系等，注意 $l_{ij}\in[0,1]$。l_{ij} 为 1，表示 X_i 与 X_j 相关，记为 $X_i\sim X_j$；l_{ij} 为 0 时，表示 X_i 与 X_j 无关。通过此算法使得 Y_i 和 Y_j 保持这种关系，即 $X_i\sim X_j$ 当且仅当 $Y_i\sim Y_j$。

假设 $l_{ij}(i\neq j)$ 为满足贝努里分布的随机变量，则其对应参数 r_{ij} 的概率分布函数 p 为

$$p(l_{ij}|r_{ij})=r_{ij}^{l_{ij}}(1-r_{ij})^{1-l_{ij}}$$

其中 $r_{ij}\in[0,1]$，为 X_i 和 X_j 之间的关系系数。同时得到：

$$q(l_{ij}|s_{ij})=s_{ij}^{l_{ij}}(1-s_{ij})^{1-l_{ij}}$$

其中 $s_{ij}\in[0,1]$，为 Y_i 和 Y_j 之间的关系系数。

我们将这个模型称为 BSMB 模型，其中 r_{ij} 为已知量，可以通过 X_i 和 X_j 求得；s_{ij} 是未知量，为 Y_i 和 Y_j 之间的连接函数。

2.2 用 KL 散度来表示不同案例的相似程度

KL 散度用来计算 $p(l_{ij}|r_{ij})$ 与 $q(l_{ij}|s_{ij})$ 之间的相似程度，我们可以得到：

$$KL(p(l_{ij}|r_{ij})\,|\,q(l_{ij}|s_{ij}))=r_{ij}\log\frac{r_{ij}}{s_{ij}}+(1-r_{ij})\log\frac{1-r_{ij}}{1-s_{ij}}$$

给定满足贝努里分布的互相独立的样本值$\{l_{ij}|I,j=1,2,\cdots,k,i\neq j\}$，它们对应的参数为 r_{ij}，我们希望通过最小化 KL 散度，即下面的函数来得到 Y_i 的坐标值：

$$L=\sum_{i\neq j}\left[r_{ij}\log\frac{r_{ij}}{s_{ij}}+(1-r_{ij})\log\frac{1-r_{ij}}{1-s_{ij}}\right]$$

这等价于最大化下面的函数：

$$F(\{s_{ij}\})=\sum_{i\neq j}[r_{ij}\log s_{ij}+(1-r_{ij})\log(1-s_{ij})] \tag{1}$$

关系系数 r_{ij} 通过高斯核来定义，即

$$r_{ij}=r(X_i,X_j,\beta)=\exp\left(-\frac{s(X_i,X_j)}{\beta}\right)$$

其中 β 为反映当数据离散化过程中为多值的变量，非二值时为反映数据特征的系数，通常与数据间的方差有关，$s(X_i,X_j)$ 为 X_i 和 X_j 之间的相似性函数。

2.3 用 Gibbs 取样来确定丢失数据的取值

Gibbs 取样是一种特殊的马尔科夫链蒙特卡罗方法，应用到 BSMB 算法中，可以把在 2.2 节中得到的 $s(X_i,X_j)$ 大于阈值结点的条件概率分布作为它们子结点的概率值。这个过程被重复多次，网络中的每个结点将都得到一个最符合条件概率分布的值。

具体做法是：首先，指定一个没有丢失值的案例作为结点变量 X 的初始值，然后用下式进行插补：

for($j=1;j<=M;j++$)

$$\begin{aligned}&\{\ X_1^{(j+1)}=\pi(X_1|X_2^{(j)},X_3^{(j)},\cdots,X_p^{(j)})\\&\quad X_2^{(j+1)}=\pi(X_2|X_1^{(j+1)},X_3^{(j)},\cdots,X_p^{(j)})\\&\qquad\vdots\\&\quad X_p^{(j+1)}=\pi(X_p|X_1^{(j+1)},X_2^{(j+1)},\cdots,X_p^{(j)})\ \}\end{aligned}$$

其中，上标 j 表示数据库中案例数的循环变量，下标表示丢失数据的序号值。这样，我们将根据上式得到丢失数据的插补值 $X_1^{(1)},X_2^{(2)},X_3^{(3)},\cdots,X_p^{(j)}$。

3 贝叶斯网络结构的学习

3.1 初始化结点变量的序列

采用最大似然树来初始化贝叶斯网络结构中结点变量的顺序。因为最大似然树是和贝叶斯网络结构具有最好拟合的树形结构，所以作为初始结点结构非常适宜，建树算法的时间复杂性是 $O(n^2)$。

3.2 结构搜索

如果网络结构未知，则构造贝叶斯推理模型是非常困难的事情。因为如果属性的数目是 n，那么可能的结构数目至少是 n 的指数阶[10~14]。在这样巨大的搜索空间寻找出合理的贝叶斯网络结构是十分耗时的，必须通过一些评价标准从中进行挑选。常用的评价标准有 MDL（minimum description length），BIC（Bayesian information criteria）以及 Be（Bayesian likelihood equivalent）等。MDL 和 BIC 的基础都是信息论，其核心都是要构建较短的编码长度。而 Be 尺度适用于参数符合特定分布的情况，其中最著名的有 BDe（Bayesian dirichlet likelihood equivalent）和 BGe（Bayesian Gaussian likelihood equivalent）。常用的搜索算法有贪心搜索（Greedy search）[12]、模拟退火（simulated annealing）、最好优先搜索

(best-first search)等。

综合以上经典算法,我们采用基于 BDe 的评价标准和启发式搜索算法,在给定各变量结点初始排序的情况下,可以比较成功地创建出与正确结构几乎一样的贝叶斯网络结构。

该算法的基本思想是:首先假设结点的父结点集合是空集,然后向该项集合中加入一个能使结果的概率最大的父结点。此过程一直加到父结点已经不再使结果增加为止。于是我们就可以得到一个可能的网络结构,然后通过排序算法找到最有可能的贝叶斯网络结构。

接下来,把本文提出的基于贝努里分布的有丢失数据的网络结构学习算法的主要思想描述如下:

输入:n 个结点的集合,一个结点可以有父结点的最大个数 u,一个包含 m 个案例的有缺失数据的数据库 D。
输出:最优的网络结构。

```
For(i=1;i<=n;i++)
  For(j=1;j<=n;j++)
  { p=pow(r[i][j],l[i][j]) * pow((1-r[i][j]),(1-l[i][j]));//满足贝努里分布
    q=pow(s[i][j],l[i][j]) * pow((1-s[i][j]),(1-l[i][j]));
//最小化 KL 散度
max (∑r[i][j] * logs[i][j]+(1-r[i][j])log(1-s[i][j])); }
//Gibbs 取样
for(i=1;i<=p;i++)
for( j=1;j<=M;j++)
  X[i][j+1] =π(X[i]|X[i+1][j],X[i+2][j],…,X[p][j]);
//启发式搜索
For( i=1;i<=n;i++)
 { πi=φ; // 设置每一个结点的父结点的初始值为空。
   Padd=f(i,πi);Flag=true;
   While (flag && |πi|< u )
{ z=Pred(xi) —πi;// z 是结点 xi 的所有先验结点集中去掉已经成为 xi 父结点的集合。
  Pnew=f(i,πi∪{z});
If (Pnew>Pold)
   { Pold=Pnew;
     πi=πi∪{z};}
else flag=false; }//endwhile
printf("Node xi's parent node is :",xi,πi);}//endfor
```

4 实验结果及分析

本文通过使结点变量满足贝努里分布,并且对数据库中的不同案例应用 KL 散度相似性量度进行评价,来对缺失数据进行插补。然后基于 BDe 的函数评价方法和启发式搜索方法,对插补后的实验数据进行贝叶斯网络结构的学习。

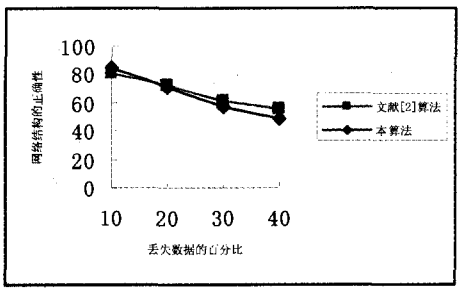

图 1 丢失数据比例对贝叶斯网络结构学习结果正确性的影响

与实验数据有关的信息见表 1,表中只是我们用于建立网络结构的训练样本集 D 中的一部分数据,目的是为了说明我们采用数据库的构成方式。文本数据库中包含五个数据域:Hippo, Thala , Amyg ,Perir ,Entor , MCI 分别表示大脑结构中的 Hippocampus(海马) , Thalamus(丘脑),Amygdala(杏仁核), Perirhinal cortex(嗅周皮质),Entorhinal (嗅内皮质) 五个结构变量和 Mild Cognitive Impairment(轻度认知障碍)一个功能变量 。其中"_"表示丢失数据。部分实验结果如图 1 所示。

表 1 部分文本数据库

Hippo	Thala	Amyg	Perir	Entor	MCI
1	—	0	0	1	1
0	0	—	1	1	1
1	1	0	1	0	1
—	0	1	0	1	1
1	0	1	—	0	1
1	1	0	0	—	1

由图 1 可以看出,丢失数据的比例对网络结构学习结果正确性的影响,是随丢失数据比例的增加而增大,经过插补后的实验数据,基本能够反映变量之间的因果关系。同时我们也看出:

在丢失数据的比例为 20%的范围内,我们的算法是优于文[2]中提出的算法。但随着丢失数据比例的不断增加,本文提出算法的准确性略低于文[2]算法,分析其可能原因为:

本算法只考虑了结点变量之间的概率分布关系,以及数据库中各案例的相似程度,没有考虑结点变量之间的拓扑关系,所以导致算法的准确性略有下降。因此,下一步我们将在本算法的基础上加入结点之间的拓扑关系,以提高算法的准确度。

算法的复杂性分析:本文提出的新方法,只是在完整数据的贝叶斯网络结构学习算法的复杂性的基础上,增加 $O(n^3)$ 的代价,但却解决了数据丢失的问题。

结论 本文基于变量之间的概率分布关系、数据库中各个案例的相似性程度分析,结合 Gibbs 取样来确定丢失数据的取值,并用最大似然树和启发式搜索算法完成了有缺失数据的贝叶斯网络结构的学习,有效避免打分-搜索结构学习方法所导致的指数复杂性,同时也为离散数据的聚类提供一种有效的方法。

参 考 文 献

1 Heckerman D. Bayesian networks for data mining[R]:[Technology Report]. MSR-TR-97-02. Microsoft Research, Redmond, 1997
2 王双成,苑森淼. 具有丢失数据的贝叶斯网络结构学习研究[J]. 软件学报,2004, 15(7):1042～1048
3 Chickering D M, Herkerman D, Meek C. Large-sample Learning of Bayesian Networks is NP-Hard[J]. Journal of Machine Learning Research, 2004,5: 1287～1330
4 Binder J, Koller D, Russell S, et al. Adaptive probabilistic networks with hidden variables[J]. Machine Learning, 1997,29(2-3):213～244
5 Chickering D M , Heckerman D. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables [J]. Machine Learning, 1997,29(2-3):181～212
6 刘大有,王飞,卢奕南,等. 基于遗传算法的 Bayesian 网结构学习研究[J]. 计算机研究与发展,2001,38(8):916～922
7 Frideman N. Learning belief networks in presence of missing values and hidden variables[C]. In: Proc. of the 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers,1997. 125～133
8 Friedman N. The Bayesian structural EM algorithm[C]. In: Proc. of the 14th International Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers,1998. 129～138
9 Herskovits E H,Gerring J P. Application of a data-mining method based on Bayesian networks to lesion-deficit analysis[C]. NeuroImage 19, 2003. 1664～1673
10 Frey B J,Jojic N. A Comparison of Algorithms for Inference and Learning in Probabilistic Graphical Models[J]. IEEE Transactions on Pattern Anal Ysis And Machine Ineligence,2005,27(9): 1392～1416
11 Cheng J,Greiner R,Kelly J,et al. Learning Networks from Data: an Information-Theory Based Approach[J]. The Artificial Intelligence Journal ,2002,137:43～90
12 Cooper G F,Herskovits E. A Bayesian Method for the Induction of Probabilistic Networks from Data [J]. Machine Learning, 1992,9:309～347
13 Yang T Y, Lee J C. Bayesian nearest-neighbor analysis via record value statistics and nonhomogeneous spatial Poisson processes [J]. Computational Statistics & Data Analysis, 2006
14 Feelders A D,van der Gaag L C. Learning Bayesian network parameters under order constraints[J]. International Journal of Approximate Reasoning,2006,42:37～53