维普资讯 http://www.cqvip.com

计算机科学2000Vol 27№ 10

贝叶斯网络结构学习分析*)

Learning Bayesian Network Structure

王双成 林士敏 陆玉昌
(清华大学计算机科学与技术系 北京100084)

Abstract In this paper the analysis of principle and process of Bayesian network structure learning is given. Bayesian network structure learning is a process that seeks the best network structure fitting the prior knowledge and data. The computing of posterior can be closed when data are completed and some other conditions are satisfied, while the computing is not closed when some data are missing. One solution for missing data is fill-in methods, another is to approximate the likelihood of structure, then to compute the probabilities of structure.

Keywords Bayesian networks, Scoring function, Searching method

贝叶斯网络结构学习(以下简称结构学习)的目标是寻找对先验知识和数据拟合得最好的网络结构。结构学习有两种方式,一种是模型选择,即选择一个最好的网络结构;另一种是选择性的模型平均,即选择合适数量的网络结构,以这些网络结构代表所有的网络结构。我们从限定的结构学习与非限定的结构学习两类问题来探讨模型选择。由贝叶斯公式可得 $P(S^h|D)=P(S^h)P(D|S^h)/P(D)$,我们用 S^h 表示数据库 D 是来自网络结构 S 的随机样本的假设,用 $P(S^h)$ 表示其不确定性,使 $P(S^h|D)$ 取最大值的网络结构就是要寻找的网络结构。其中 $P(D)$ 与网络结构无关,可以忽略。$P(S^h)$ 称为先验结构概率分布。在实际处理时,一种情况是认为所有的网络的先验结构是等可能的,另一种情况是认为网络结构不是等可能的(这时对每一个网络结构要指派一个先验结构概率分布)。在限定的结构学习中,让先验结构的作用小一些,有时甚至不考虑先验结构(认为是等可能的),在非限定的结构学习中,让先验结构的作用大一些,因为结构似然越高网络结构越复杂(完全网的结构似然最高,但不是我们所需要的,因为它不具有任何的条件独立性)。如果先验结构的作用过小,有可能会产生对数据的过度拟合,因此用先验结构平衡结构似然结构。结构似然 $P(D|S^h)$ 是结构学习的核心,结构似然的统计含义是:结构似然值越高,网络结构就与数据拟合得越好。结构似然的计算有两种情况,一种是结构似然的计算具有封闭性,另一种是结构似然的计算不具有封闭性,这时一般采用近似的方法。

一、非限定的贝叶斯网络结构学习

非限定的贝叶斯网络结构学习不对网络结构进行任何限定,在理论上应该搜索所有的网络结构。学习过程可以是贝叶斯的,也可以是非贝叶斯的。

1. 似然评分标准

对数似然标准:

$$P(D|S^h)=\prod_{l=1}^{N}P(x_l|x_1,\cdots,x_{l-1},S^h)$$

$$\log P(D|S^h)=\sum_{l=1}^{N}\log P(x_l|x_1,\cdots,x_{l-1},S^h)$$

其中 x_l 是数据库的第 l 个情况。交叉检验标准如下:

$$CV(S^h,D)=\sum_{l=1}^{N}\log P(x_l|V_l,S^h)$$

其中 $V_l=\{x_1,\cdots,x_{l-1},x_l,\cdots,x_N\}$

局部对数似然标准:

$$LC(S^h,D)=\sum_{l=1}^{N}\log P(a_l|F_l,x_l,S^h)$$

其中 a_l 和 F_l 是数据库中第情况类变量和属性变量的观测值,这种标准主要用于分类。这里使用的是对数似然标准。

2. 评分函数

在贝叶斯网络结构学习中有两个主要的评分函数:贝叶斯后验评分函数和基于最小描述长度原理的

*)国家重点基础研究发展计划项目,国家自然科学基金项目,"九五"国家攀登计划预选项目。王双成 副教授,访问学者,研究方向:知识工程、数据采掘与知识发现。林士敏 副教授,访问学者,研究方向:机器学习,数据采掘与知识发现。陆玉昌 教授,研究方向:知识发现,机器学习,知识工程。

维普资讯 http://www.cqvip.com

评分函数。

(1)贝叶斯后验评分函数　假设数据是完整的、变量是有限离散的、参数具有独立性、参数具有迪里赫列分布、那么结构似然的计算具有封闭性。

·先验结构。它的一种处理方法是认为所有的网络结构都是等可能的，这时可以忽略先验结构；另一种处理方法是认为网络结构不是等可能的，给每一个网络结构指派一个先验概率，可以用下面的方法指派。给定一个先验网络结构 P（认为比较可能的网络结构，作为参照结构，如果不了解可指定为空网或完全网），用 $\Pi_i(S)$ 和 $\Pi_i(P)$ 分别表示变量 X_i 在网络结构 S 和 P 中对应的父结点集合，对任意的变量 $X_i\in U$，让 δ_i 表示在 $\Pi_i(S)$ 和 $\Pi_i(P)$ 中对应的不同的结点的个数，即 $\delta_i=(\Pi_i(S)\cup\Pi_i(P))\backslash(\Pi_i(S)\cap\Pi_i(P))$，让 $\delta=\sum_{i=1}^{n}\delta_i$，取

$$P(S^k|\xi)=c\kappa^{\delta}$$

c 是正规常数，可以忽略，κ 是网络结构惩罚因子 $0<\kappa\leqslant 1$，如果愿意可以使 κ 变得更具体，使其是对一种局部结构（结点和它的父结点结构）的惩罚 $P(S^k|\xi)=c\kappa_{11}^{\delta_1}\kappa_{12}^{\delta_2}\cdots\kappa_{ij}^{\delta_{ij}}$。

·后验结构。假设：(1)变量 X_i 是离散的；(2)数据集是完整的；(3)参数矢量是相互独立的，即 $P(\theta_S|B_S^k,\xi)=\prod_{i=1}^{n}P(\theta_i|B_S^k,\xi)$，$P(\theta_i|B_S^k,\xi)=\prod_{j=1}^{q_i}P(\theta_{ij}|B_i^k,\xi)$；(4)参数矢量具有迪里克莱(Dirichelet)分布，即：

$$P(\theta_{ij}|B_S^k,\xi)=Dir(\theta_{ij}|N_{ij1},\cdots,N_{ijr_i})$$

$$=\frac{\Gamma(\sum_{k=1}^{r_i}N_{ijk})}{\prod_{k=1}^{r_i}\Gamma(N_{ijk})}\prod_{k=1}^{r_i}\theta_{ijk}^{N_{ijk}-1}。$$

记 $U=\{X_1,\cdots,X_n\}$ 为变量集，其中 $X_i\in U$ 的值域（或状态集）为 $\{x_i^1,\cdots,x_i^{r_i}\}$，$D=\{C_1,\cdots,C_N\}$ 为数据样本，其中 C_i 为一事例，让 $\theta_{ijk}=P(x_i^k|Pa_i^j,S^k,\xi)$，$\theta_{ijk}>0$，$\sum_{k=1}^{r_i}\theta_{ijk}=1$，其中 $Pa_i^1,\cdots,Pa_i^{q_i}$，$q_i=\prod_{x_t\in Pa_i}r_t$，描述了 Pa_i 的结构，$\theta_S=\bigcup_{i=1}^{n}\{\theta_i\}$，$\theta_i=\bigcup_{j=1}^{q_i}\{\theta_{ij}\}$，$\theta_{ij}=\bigcup_{k=2}^{r_i}\{\theta_{ijk}\}$。满足这些条件，结构似然的计算具有封闭性。

先验参数分布：

$$P(\theta_S|S^k,\xi)=c\prod_{k=1}^{r_i}\theta_{ijk}^{N'_{ijk}+N_{ijk}-1}\qquad(1)$$

其中 c 是正规常数。

后验参数分布：

$$P(\theta_S|D,S^k,\xi)=c'\prod_{k=1}^{r_i}\theta_{ijk}^{N'_{ijk}+N_{ijk}-1}\qquad(2)$$

其中 c' 是正规常数，N_{ijk} 是在数据库 D 中满足 $X_i=x_i^k$ 且 $Pa_i=j$ 的案例数量。

后验联合分布为：

$$P(x_{n+1}|D,S^k,\xi)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{N_{ijk}-N_{ijk}}{N_{ij}-N_{ij}}\qquad(3)$$

其中 $N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$ 和 $N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$。应用贝叶斯公式，可得：

$$P(S^k|D,\xi)=P(S^k|\xi)P(D|S^k,\xi)/P(D)\qquad(4)$$

根据对数似然标准

$$P(D|S^k,\xi)=\prod_{j=1}^{N}P(x_j|x_1,\cdots,x_{j-1},S^k,\xi)\qquad(5)$$

把(3)代入(4)再代入(5)可得：

$$P(D|S^k,\xi)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\{$$

$$\left[\frac{N_{ij1}}{N_{ij}}\cdot\frac{N_{ij1}+1}{N_{ij}+1}\cdots\frac{N_{ij1}+N_{ijk}-1}{N_{ij}+N_{ijk}-1}\right]\cdots$$

$$\left[\frac{N_{ij2}}{N_{ij}+N_{ij1}}\cdot\frac{N_{ij2}+1}{N_{ij}+N_{ij1}+1}\cdots\frac{N_{ij2}+N_{ijk}-1}{N_{ij}+N_{ij1}+N_{ij2}-1}\right]\cdots$$

$$\left[\frac{N_{ijr_i}}{N_{ij}+\sum_{k=1}^{r_i-1}N_{ijk}}\cdot\frac{N_{ijr_i}+1}{N_{ij}+\sum_{k=1}^{r_i-1}N_{ijk}+1}\cdots\frac{N_{ijr_i}+N_{ijk}-1}{N_{ij}+N_{ij}-1}\right]\};$$

$$=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N_{ij})}{\prod_{k=1}^{r_i}\Gamma(N_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N_{ijk}+N_{ijk})}{\Gamma(N_{ijk})}$$

可得：$P(S^k|D,\xi)=\dfrac{P(S^k|\xi)P(D|S^k,\xi)}{\sum_S P(S^k|\xi)P(D|S^k,\xi)}$

$$=c\cdot\kappa^{\delta}\cdot\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N_{ij})}{\prod_{k=1}^{r_i}\Gamma(N_{ij}+N_{ij})}\cdot\prod_{k=1}^{r_i}\frac{\Gamma(N_{ijk}+N_{ijk})}{\Gamma(N_{ijk})}\qquad(6)$$

其中 $N_{ijk}=N'\cdot P(X_i=k,Pa_i=j|S_C^k,\xi)$，$N'$ 是等价样本大小（表示用户对先验网的信任程度），$P(X_i=k,Pa_i=j|S_C^k,\xi)$ 是利用先验贝叶斯网络来计算，N_{ijk} 是在数据库 D 中满足 $X_i=x_i^k$ 且 $Pa_i=j$ 的情况数量。

(2)基于最小描述长度原理的评分函数　假设数据是完整的、变量是有限、离散的。基于最小描述长度原理的评分函数(minimal description length，MDL)要求通过学习发现具有尽可能短的描述长度的网络结构。描述长度包括网络结构本身的描述长度和使用网络结构描述数据的长度。用 $MDL(S|D)$ 表示给定数据集 D 网络结构 S 的 MDL 评分函数，那么基于最小描述长度原理的评分函数是：

$$MDL(S|D)=1/2\cdot\log N|S|-LL(S|D)\qquad(7)$$

其中 $|S|$ 表示网络参数的数量，$1/2\cdot\log N$ 是每一个参数使用的比特数(信息单位)，第一项描述网络 S 的长度，第二项是给定数据集 D，S 的对数似然，它是基于概率分布 P_S 描述 D 所需要的比特数：

$$LL(S|D)=\sum_{i=1}^{N}\log(P_S(C_i))\qquad(8)$$

让 $\hat{P}_D(x_i,\Pi_{x_i})$ 表示在 D 中 X_i 具有状态 x_i 及 X_i 的父结点具有状态 Π_{x_i} 出现的频率，假定数据集是完整的，那么

维普资讯 http://www.cqvip.com

$$LL(S|D)=\sum_{i=1}^{n}\log(P_D(C_i))$$

$$=\sum_{i=1}^{n}\sum_{x_i\in Val(X_i)}\log(\theta_{x_i|\Pi_{x_i}})^{N\cdot P_D(x_i,\Pi_{x_i})}$$

$$=N\sum_{i=1}^{n}\sum_{x_i\in Val(X_i)}P_D(x_i,\Pi_{x_i})\log(\theta_{x_i|\Pi_{x_i}}) \quad (9)$$

其中：$N\cdot P_D(x_i,\Pi_{x_i})$为在数据集中出现的次数；$Val(X_i)$为$X_i$的取值范围。可以证明，当$\theta_{x_i|\Pi_{x_i}}=P_D(x_i,\Pi_{x_i})$时$LL(S|D)$取最大值。因为方程(7)的第一项不依赖参数的选择，因此，这个解使MDL取最小值(最小评分)。MDL评分法不依赖于结构先验，是渐近正确的。

3. 局部搜索算法

这种局部搜索算法基于评分函数的可分解性，即网络结构的测度能被分解成局部测度(只与该结点和它的父结点有关)的积，$P(D|S^h,\xi)=\prod_{i=1}^{N}s(x_i|\Pi_i)$。贝叶斯后验评分函数和基于最小描述长度原理的评分函数都是可分解的。局部搜索算法的搜索过程是：(1)选择初始网络结构(可以是空结构，随机指定的结构或先验网络结构等)作为起点；(2)通过增加、删除及转向操作使得局部最大化(由于评分函数是可分解的，弧的变化的测度计算就可以局部化，使其产生局部最大值)，再逐渐扩展到整个网络最大化；(3)可以适当地使用重复爬山法(随机地弄乱网络结构，重新搜索)或模仿退火法摆脱局部最大值的束缚。

4. 结构似然的近似

当数据量很大或数据不完整时，结构似然的计算比较困难，因此常采用近似的方法。拉普拉斯方法是一种非常有效的近似方法，在某些条件下相对误差是$O(1/N)$，计算复杂度是$O(d^2,N)$

$$\log P(D|S^h)\approx\log P(D|\bar{\theta}_s,S^h)+\log P(\bar{\theta}_s,S^h)$$

$$+\frac{d}{2}\log(2\pi)-\frac{1}{2}\log|A| \quad (10)$$

其中d是$g(\theta_S)$的维度，$d=\prod_{i=1}^{n}q_i(r_i-1)$，$\bar{\theta}_s$是使$g(\theta_s)\equiv\log(P(D|\theta_s,S^h)\cdot P(\theta_s|S^h))$取最大的值$\theta_S$的配置(具有最大后验结构概率的配置)，$A=\left(\frac{\partial g(\theta_S)}{\partial\theta_{x_i|\Pi_{x_i}}}\right)_{|\theta_S=\bar{\theta}_s}$是$g(\theta_S)$在$\theta_S$具有配置$\bar{\theta}_s$的负海森(Hessian)矩阵，$\Pi_{x_i}$是$X_i$父结点的集合，$(\theta_S-\bar{\theta}_s)'$是$(\theta_S-\bar{\theta}_s)$的转置。$\bar{\theta}_s$可用EM算法求得，其计算过程是：

(1)为θ_s指派一个初值(可以随机地指派)；

(2)计算关于数据集的充分统计量的期望值

$$E_{P(\cdot|D,\theta_S,S^h)}(N)=\sum_{l=1}^{N}P(x_i^k,Pa_i^j|x_l,\theta_S,S^h) \quad (11)$$

当X_i和在Pa_i中的变量在情况x_l中没有数据丢失时，这时的概率值是1(在x_l中有x_i^k,Pa_i^j配置时)或0(在x_l中没有x_i^k,Pa_i^j配置时)。否则，可以用概率预测来求其值；

(3)计算参数最大值，如果是似然最大值，那么

$$\theta_{ijk}=\frac{E_{P(\cdot|D,\theta_S,S^h)}(N_{ijk})}{\sum_{k=1}^{r_i}E_{P(\cdot|D,\theta_S,S^h)}(N_{ijk})} \quad (12)$$

如果是后验最大值，那么

$$\theta_{ijk}=\frac{N_{ijk}+E_{P(\cdot|D,\theta_S,S^h)}(N_{ijk})}{\sum_{k=1}^{r_i}(N_{ijk}+E_{P(\cdot|D,\theta_S,S^h)}(N_{ijk}))} \quad (13)$$

$|A|$的计算也常采用近似的方法，一种方法是用主对角线元素之积近似行列式的值，另一种方法是用准对角线子行列式乘积近似行列式的值。

二、限定的贝叶斯网络结构学习

限定的贝叶斯网络结构学习要对网络结构作某种限定，搜索空间是限定的网络空间。非限定的贝叶斯网络结构学习中的方法均适用于限定的贝叶斯网络结构学习。贝叶斯网络分类器的结构学习是典型的限定的贝叶斯网络结构学习。TAN贝叶斯网络分类器的结构学习就是其中一个例子。学习过程是非贝叶斯的。

结束语 贝叶斯网络结构学习是贝叶斯网络学习的重要组成部分。学习方法可以是贝叶斯方法，也可以是非贝叶斯方法。对具有较弱限定的网络结构一般要用先验结构平衡结构似然，对具有较强限定的网络结构一般可以不考虑先验结构。提高贝叶斯网络结构学习效率的主要因素是评分函数和搜索算法，优化评分函数和搜索算法便是贝叶斯网络结构学习的核心问题。

参 考 文 献

1 Heckerman D Learning Bayesian Networks. [Technical Report MSR-TR-95-02] Microsoft Research, Microsoft Corporation, 1995

2 Friedman N. Bayesian Network Classifiers. Machine Learning, 1997, 29: 131～163

3 Heckerman D. Bayesian Networks for Data Mining. Data Mining and Knowledge Discovery, 1997, 1: 79～119