计算机科学2002Vol.29№.2

数据缺失条件下的贝叶斯推断方法

Bayesian Inference with Missing Data

虞健飞　张恒喜　朱家元

（空军工程大学工程学院航空机械工程系　西安710038）

Abstract Recently Bayesian network(BN)becomes a noticeable research direction in Data Mining. In this paper we introduce missing data mechanisms firstly, and then some methods to do Baysesian inference with missing data based on these missing data mechanisms. All of these must be useful in practice especially when data is scare and expensive. It can foresee that Bayesian networks will become a powerful tool in Data Mining with all of these methods above offered.

Keywords Bayesian network, Probabilistic inference, Missing data mechanisms, Approximation

1. 简介

当根据先验信息或者已有的数据建立起一个贝叶斯网络后，经常需要根据这个网络模型计算一些感兴趣的事件的概率或做一些预测。但是，一般情况下，这些概率并不能够直接从网络模型中获得，而需要依据贝叶斯定理进行一些推导。通常称这些推导计算为概率推断。

在样本完整时，贝叶斯共轭分析对给定X时求Y的条件分布以及Y的边缘分布提供了一个简单的方法。假设样本来自于参数为θ的多项分布，且$\theta_{ij}=p(X=i,Y=j|\theta)$，$\theta$的共轭先验是Dirichlet分布，则$\theta$的后验分布也是Dirichlet分布[Lindley,1964]。根据θ的后验分布可以很容易地推导出给定X时Y的条件分布以及Y的边缘分布。X和Y之间联系的测度也可以推导出来。但是当样本不完整时，由于X和Y提供的信息不对称，不能够使用前述的方法。

在现实世界中，往往出现数据缺失的情况，经常由于处理方法不恰当导致错误的推断结果发生。例如，1992年4月9日，保守党第四次在英国大选中以7.6%的优势胜出。但在大选当日，四个主要的调查公司的最后一次民意测验结果仍然表明工党领先了0.9%，预测误差达到了8.5%。在后来的调查中发现引起如此大的预测误差的主要原因之一在于那些在选举中没有明确表示他们选举意图的人的影响，大约占误差的2%。由于在进行分析时假设这些缺失数据是可以忽略的，所以民意测验中剔除了拒绝回答或回答“不知道”的选民。但结果表明，这一假设对调查结果有着致命影响，因此在实际解决有关问题时必须对数据缺失采取正确、有效的处理方法。

本文在论述数据缺失模式的基础上，着重介绍了几种不同的处理数据缺失模式的方法，并对这几种方法进行了对比分析。

2. 数据缺失模式

缺失数据能够影响到抽样的随机性，这使得许多现有的统计学方法失去了应用基础，因此不完整抽样动摇了现代统计学的核心基础[Copas and Li,1997]，分析者遇到了极大的挑战。我们目前对缺失数据对样本对总体特性的真实反映程度的影响主要是基于Riubin[1976]、Littele和Rubin[1987]以及Gelman[1995]年提出的数据缺失模式的分类。设想一个样本根据两个枚举型变量X和Y的取值进行分类，变量X是独立变量，Y是响应变量，且Y有可能无响应(数据缺失)。我们对Y无响应用$Y=?$表示，样本中一个不完整的一个案例用$(X=i,Y=?)$表示。这个简单的二变量模型可以很容易地扩展到包含多个变量，并且其中一个无缺失值的变量集影响一个有可能有缺失值的变量。这是在统计应用中经常遇到的情况，例如在机器学习和数据挖掘中的观点学习、可控试验以及有监督分类等。

数据缺失模式分类取决于$Y=?$出现的概率依赖于Y和X或仅依赖于X。如果这个概率仅依赖于X，那么数据可以称为随机缺失(MAR)，此时观察到的Y值虽然不能完整表征总体，但是如果考虑X的取值则该不完整样本也认为可以表征总体。一种特殊的情况是$Y=?$的概率既不和Y有关也不和X有关。在这种情况下，数据是完全随机缺失(MCAR)，则可以认为观察到的值和缺失值的产生机理相同，因此可以认为Y的观察值是隐含的完整抽样的一个子集。当数据是MAR和MCAR时，数据缺失模式称为是可忽略模式，此时推断结果不取决于数据缺失模式。当$Y=?$同时依赖于Y和X时，数据缺失模式称为不可忽略模式(NI)，此时不完整的样本就不再具有典型意义。

这种划分方式为研究不完整样本提供了一个有力的框架，但是由于数据缺失模式的可忽略性不同使得包含缺失数据的案例不能够简单地从样本中剔除，贝叶斯推断过程因此也变得复杂，需要一些计算过程中的技巧。下面基于上述的数据缺失模式介绍几种贝叶斯推断过程中采用的方法。

3. 数据缺失时的贝叶斯推断

表1给出了各种数据缺失模式下的贝叶斯推断(先验为Dirichlet分布)。

CP：后验分布与先验共轭，为Dirichlet分布；CP(a)：当大样本时后验与先验共轭；EV：只有后验均值和方差可以确切计算；?：后验分布为混合Dirichlet分布，没有简单的表达式。

虞健飞　博士生，主要从事数据仓库、数据挖掘、知识发现等方面的研究。张恒喜　教授，博导，主要从事军事装备学武器装备发展规划与管理方面的教学与科研工作。朱家元　博士生，主要从事神经网络优化、管理智能决策等方面的研究。

表1 不同数据失效模式下的贝叶斯推断

	MCAR	MAR	NI
θ_{ij+}	CP	CP	CP
θ_{ijk}	CP	CP	?
θ	CP(a)	EV	?
θ_{+j}	CP(a)	EV	?

当数据缺失模式为MCAR，且样本中完整案例的数量很大时，可以将包含不完整数据的案例从样本中剔除后再采用贝叶斯共轭分析方法。

当数据缺失模式是MAR时，$p(Y|X)$的后验分布仍然和先验共轭，那么$p(X,Y)$和$P(Y)$的后验概率以及它们的后验方差都可以简单地计算出来。但是$p(X,Y)$和$P(Y)$的后验表达式很复杂，一般都采用蒙特-卡洛方法(或抽样方法)和高斯近似等方法进行计算。

当数据缺失模式是NI时，贝叶斯推断将变得十分困难。一种基于转移的分析方法首先根据无响应$\phi_{ij}=p\{Y=j|X=i,Y=?,\theta\}$的分布模拟产生缺失的数据以获得一个完整的抽样，然后采用一些统计分析方法对该完整样本进行分析。在这种情况下，由于观察到的数据并没有包含无响应分布的信息，因此在采用基于转移的分析方法时，需要关于ϕ的分布的外生信息。本文将介绍一种对任何一种数据缺失模式都有良好效果的定界塌陷方法(简称为BC方法)，该方法对使NI模式下的贝叶斯推断也很容易。

3.1 蒙特-卡洛方法

蒙特-卡洛方法是一种基于抽样的方法。如果有足够长时间使计算过程收敛的话，该方法可以得到非常准确的结果。

在此将讨论蒙特-卡洛方法中的Gibbs抽样方法，这是由Geman于1984年提出的。假设变量组$=\{X_1,\cdots,X_n\}$服从一定的联合分布$p(X)$，可以根据以下步骤用Gibbs抽样器近似函数$f(X)$对$p(X)$的期望：

·首先，对X中的每一个变量用某种方法(例如随机选择)选择初始状态；

·其次，选择某一个变量X_i，取消其初始状态，计算其在其余$n-1$个变量当前状态下的概率分布；

·然后，根据X_i的分布抽取一个值，并计算$f(X)$；

·最后，重复上述两个步骤，并记录其平均值。

如果满足下列两个条件，对平均值取极限，也就是重复次数接近无穷，记录下来的平均值就是$E_{p(X)}f(X)$：

·第一，Gibbs抽样器必须是不可缩减的。也就是概率分布必须保证无论处于何种初始状态，最终都可以抽取任何一个可能的X的组合形式。例如，如果$p(X)$不包含0概率，则Gibbs抽样器就是不可缩减的。

·第二，每一个X_i必须都能够抽取任意多次。在实际使用过程中经常采用各变量按一定顺序轮流抽取的方法。

Neal[1993]以及Madigan和York[1995]给出了有关Gibbs抽样和其它蒙特-卡洛方法的介绍以及初始化的方法和有关收敛性的讨论。

为了说明贝叶斯推断中的Gibbs抽样，下面给定一个不完整样本$D=\{y_1,\cdots,y_N\}$以及一个具有Dirichlet先验并针对离散变量的贝叶斯网络，近似计算对某一θ_i的概率密度$p(\theta_i|D,S^h)$。为近似计算$p(\theta_i|D,S^h)$：

①按某种方法给定样本中每一个案例中未观测变量一个初始值，从而获得一个完整的随机样本；

②选择某一在原始样本中没有观察到的变量X_{il}(第l个案例中的变量X_i)根据如下的概率分布重新给定X_{il}的值：

$$p(x'_{il}|D_c\backslash x_{il},S^h)=\frac{p(x'_{il},D_c\backslash x_{il}|S^h)}{\sum_{x^*_{il}}p(x^*_{il},D_c\backslash x_{il}|S^h)}$$

其中，$D_c\backslash x_{il}$表示数据集D_c中剔除了x_{il}，分母中的求和表示X_{il}取所有可能的值；

③对D中所有没有观察到的变量重新赋值得到新的随机完整样本D'_c；

④根据完整贝叶斯网络推断方法计算；

⑤重复步骤②～④，并以$p(\theta_i|D'_c,S^h)$的平均作为近似。

3.2 高斯近似方法

蒙特-卡洛方法虽然可以产生精确的结果，但当样本容量很大时该方法往往是难以实现的。另外一种近似方法，高斯近似，比蒙特-卡洛方法更为高效，而且在有一个相对的大样本时也可以达到准确的结果[Kass et al，1988；Kass and Raftery，1995]。

高斯近似方法的出发点在于，在大样本时，$p(\theta_i|D,S^h)\propto p(D|\theta_i,S^h)p(\theta_i|S^h)$并且可以近似为多变量高斯分布。令：

$$g(\theta_i)\equiv\log(p(D|\theta_i,S^h)p(\theta_i|S^h)) \quad (1)$$

同时定义$\tilde{\theta}$为当$g(\theta_i)$取最大值时θ_i的取值。θ_i的这个取值也最大化了$p(\theta_i|D,S^h)$，$\tilde{\theta}$称为是最大化后验(MAP)取值。将$g(\theta_i)$在点$\tilde{\theta}_i$展开，得到：

$$g(\theta_i)\approx g(\tilde{\theta}_i)-\frac{1}{2}(\theta_i-\tilde{\theta}_i)A(\theta_i-\tilde{\theta}_i) \quad (2)$$

式中A是$g(\theta_i)$在θ点的负海森矩阵。由(1)式可以得：

$$p(\theta_i|D,S^h)\propto p(D|\theta_i,S^h)\cdot p(\theta_i|S^h)$$

$$\approx p(D|\tilde{\theta}_i,S^h)p(\tilde{\theta}_i|S^h)\exp\{-\frac{1}{2}(\theta_i-\tilde{\theta}_i)A(\theta_i-\tilde{\theta}_i)'\}$$

为了计算高斯近似，必须计算$\tilde{\theta}$以及$g(\theta_i)$在$\tilde{\theta}$点的负海森矩阵。Meng和Rubin[1991]给出了一个计算展开式(2)第二部分的计算方法，Raftery[1995]介绍了如何利用许多统计学软件包中都有的相似-比率测试来计算海森矩阵的方法，Thiesson[1995]证明了对无约束多项分布，展开式的第二部分可以用贝叶斯推断进行计算。

当样本容量增大时，先验$p(\theta_i|S^h)$的影响逐渐变弱，此时可以用$p(D|\theta_i,S^h)$获得最大时的θ_i来代替$\tilde{\theta}_i$，称为最大似然(ML)取值。有关θ_i的ML或MAP取值已经有许多成熟的算法，例如基于坡度的优化方法(Buntine，1994)和最大期望法[Dempster et al，1977]等。

3.3 定界塌陷方法

BC方法的基本特征是将样本和假定的数据缺失模式提供的信息显示地、分别地表示出来。该特征使得可以建立一个与数据失效模式无关的通用估计方法。

BC方法第一步(定界)在没有有关数据缺失模式信息时，使一个不完整样本通过极端分布来界定一个区间以包含所有可能的估计。在该方法中，一个完整的样本能够提供足够信息以将界定区间限定到一个点。当有关缺失数据模式的信息已知时，这些信息可以表示为可能的无响应样式并且用来从区间中选取某一个估计。

BC方法的第二步(塌陷)是通过根据假定的数据缺失模式确定的权值将极端估计凸联合，以将该界定区间塌陷成一个估计值。这些点估计可以用来近似感兴趣的参数的后验分布。根据完整案例建立的无响应模型可以容易地表示出不同的数据缺失模式，例如MAR和MCAR，并且此时BC方法可

(下转第50页)

全局变量 ill_g_head 指向链表的每个 ill_s 项与低层的 device 一一对应，有多少个同时打开的网络设备(每个设备可以打开多次)，则有多少个 ill_s 项，也就是说有多少个流。在启动加载安全引擎模块时，调用 dev_ops 的 attach 函数完成初始化工作，在我们的实现中主要完成 qif 链表和 ill_s 链表的钩链，任一个 qif 链表项将 ill_s 对应项的模块接入点指针 q_qinfo(实际上队列定义的相关操作入口点)保存下来，然后再让它指向新的处理程序，fr_qin()和 fr_qout()，由这两个处理程序完成各种安全处理。

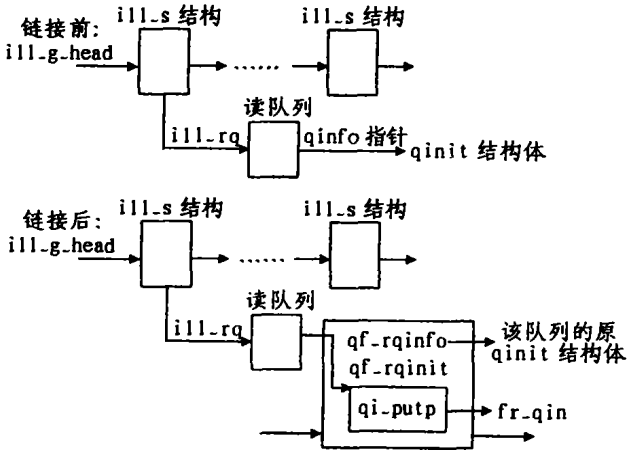

图2 ill_s 和 qif 链表的链接示意图

从图2我们看到，在打开的网络设备上加载安全引擎后，实际上将在 IP 模块中建立的 ill_s 链表与在安全引擎模块中建立的 qif 链表进行了钩链，在 qif 结构中的 qf_rqinfo 中保存 ill_s 结构中定义的读队列 ill_rq 的队列操作入口点指针 q_qinfo，然后，让其指向在 qif 中定义的新的读队列 qinit 结构 qf_rqinit，而所有安全引擎的相关操作入口都通过 qf_rqinit 进行。

对写队列也进行上述操作，只不过是修改 ill_s 结构对应的下邻接队列。

4 SecureGuard 的特点和典型应用

SecureGuard 与我们在 NT 环境实现的 NetGuard[3]功能相同，从应用的角度来看，归纳起来，系统具有如下特点：

·它是在系统内核实现网络安全，与应用程序无关，对用户操作透明。这样，不论是原来的程序还是新的程序均可照常使用，不用任何改动，程序的操作也和以前一样。但具备了对所有应用程序在私网范围内安全，不必对各个应用程序再采取特殊的安全措施。

·它与具体的通讯设备无关。凡是 Solaris 系统支持的网卡，调制解调器等物理接入设备都可使用。

·灵活性。可根据用户要求配置安全策略：譬如根据不同地址，不同服务和不同用户设置准入控制；选择加解密算法，选择保密范围，定制保密级别等。NetGuard 系统框架结构设计灵活，功能容易扩充。

·性能/价格比高。它是纯软件系统实现，比目前主要采用过滤方法的防火墙系统有更优越的性价比。

(下转第19页)

(上接第123页)

以得到一个广义的极大似然估计。在一般的数据缺失模式下，BC 估计是否就是期望的贝叶斯估计取决于所确定的无响应 ϕ 样式。

界定过程中，所有可能的估计边界表示由于缺失数据给估计过程带来的不确定性，而界定区间宽度可以度量估计时不完整样本所提供的信息的质量。通过计算边界可以将缺失数据带来的不确定性保留在分析过程中，这样由于确实缺失数据导致的抽样过程的变化和不确定性可以被独立地进行计算并表示出来。BC 方法的另一个优势就是其计算开销，对每一种条件分布，BC 方法在界定过程和塌陷过程中的凸联合针对响应变量的每一个取值进行逐步更新，减少了分析过程中的计算的复杂性。因此 BC 方法的计算复杂程度只是响应变量 Y 的可能取值个数的函数，而和缺失数据的数量无关。

结论 Gibbs 抽样、基于转移的方法以及高斯近似方法都有三个主要的缺点：

·都假设数据缺失模式是能够知道的，但实际情况并不总能够满足这一要求。

·由于没有完全考虑缺失数据，提供的估计结果可靠性程度的度量方法使抽样中的不确定性和无响应带来的外生不确定性混合在一起。

·这两种计算方法的计算代价都是缺失数据数量的函数。由于 Gibbs 抽样将缺失数据看作未知参数，因此该方法的收敛速度随着缺失数据的增多而下降。基于转移的方法的精度随着模拟产生的样本数量而增加，但每一次模拟的计算开销随着缺失数据的数量增加而增加。

与建立了数据缺失模式模型的现有的一些方法相比，BC 方法在塌陷阶段利用了一个无响应样式，该样式是数据缺失模式的函数，并且随后的推断都是基于这种无响应样式。由于 BC 方法计算的简单性，不同的无响应样式都可以被表示出来，从而各种不同数据缺失模式对推断结果的影响的敏感程度可以很快地计算出来。因此，BC 方法提供了一个通用的对不完整样本进行分析的框架，Kadane[1993]以及 Kadane 和 Terrin[1997]采用该方法所做的几个实际案例都证明了这一点，并且可以对不同无响应样式分别计算并求其平均来获得边际推断。

参 考 文 献

1 Bayesian H D. Networks for data mining[J]. Data Mining and Knowledge Discovery, 1997, 1: 79～119

2 Geiger D, Heckerman D. A characterization of the Dirichlet distribution with application to learning Bayesian networks[A]. In: Proc. of Eleventh Conf. on Uncertainty in Artificial Intelligency[C], Montreal, QU, 1995. 196～207

3 Dagum P, Luby M. Approximating probabilistic inference in Bayesian belief networks is NP-hard[J]. Artificial Intelligence, 1993, 60: 141～153

4 Paola S, Marco R. Baysian inference with missing data using bound and collapse. [KMI-TR-58], The open university research report, 1997

5 Gilks W R, Roberts G O. Strategies for improving MCMC. In: Gilks W R, Richardson S, Spiegelhalter D J, eds. Markov Chain Monte Carlo in Practice, pp. 89～114, Chapman and Hall, London

6 林士敏，田凤占，陆玉昌. 贝叶斯学习、贝叶斯网络与数据挖掘. 计算机科学，2000，27(10)