文章编号:1671-6833(2003)02-0093-03

模型选择中的Bayes 方法

严于鲜,易建华

(西南交通大学理学院,四川 成都 610031)

摘 要:在以往关于模型选择的研究中,一般都是先假定选定一个模型,然后对由此模型确定的分布族进行比较,求出最优的分布函数和数值特征,忽略了模型本身的不确定性.介绍了Bayes 方法在模型选择中的方法及应用,举例说明了用Bayes 方法选择模型,不仅能够减小模型选择中模型不确定性的影响,而且可以根据实际情况和问题认识程度的深化,对模型进行扩展.

关键词:模型选择;Bayes 方法;不确定性

中图分类号:O 212.8 **文献标识码**:A

0 引言

在常见的统计模型选择过程[1]中,一般遵循以下原则:首先根据实际情况和数据特征决定模型和未知参数的个数,然后根据模型和参数个数选定一族分布函数族(称其为可行分布族),然后在这些可行分布族中选取满意的分布函数族和最优分布函数.通常采用以下两种方法选取可行分布族和最优分布函数.

(1) 基于零假设检验的分布族:首先确定数据的一般明显特征,选取所有满足这些特征且包含上述未知参数的分布族为可行分布族,然后在这些分布族中进行检验,看其是否和数据的其它特殊特征一致.此时可以确定一个检验指标,如果不满足检验指标,则舍去此分布函数族,检验下一族函数分布族,直至找到满意的分布函数族和最优的分布函数.如果最后仍没有满意的分布族,可进一步观察数据,减少一般特征的个数(可把减下来的一般特征看作特殊特征),重新确定可行分布函数族,进行检验.如此反复,直到满意为止.

(2) 基于矛盾的分布族:其主要思想是选出分布族,该分布族和实际背景假设、样本大小、问题的具体要求等实际情况相符.然后在此基础上,检验这些分布族与实际情况差距的大小,选出差距最小的分布族作为最终的分布函数族.该方法具有很大的灵活性,因为操作者可根据实际要求或自己的直观感觉,将模型的特征排序,选出某一特征作为最重要的判断指标,在此条件下选择分布函数.

但一般情况下,在对某问题选择模型时,总是先选定一个模型,相应地就确定了可行分布族的类型,并认为该模型是所有可用于该问题的模型中最好的,然后在该模型及分布族的基础上求最优分布函数和各种数值特征.但这样做却首先排除了模型本身的不确定性.在实际情况中,很难一次性地准确确定问题的模型,如果在进行模型选择时不考虑模型的不确定性的影响,任意选定一个模型,则会导致过低估计所需要的特征值(兴趣量)[2],甚至可能产生错误的结论.

1 Bayes 方法

用Bayes 方法可以很好地解决模型不确定性影响兴趣量的问题.设△表示兴趣量,$A=\{M_1, M_2, \cdots, M_n\}$是所有可用于某问题的模型,在给定观察数据 D 的情况下,Bayes 方法用:

$$p(\triangle/D)=\sum_{i=1}^{n} p(\triangle/M_i, D)\, p(M_i/D) \quad (1)$$

表示兴趣量.它是以后验模型概率为权,对兴趣量在各个模型中的后验分布加权求平均,这样就尽量减小了模型不确定性对兴趣量的影响.而且式(1)的预测数据的能力比仅选一个模型的情况强.但这种方法却产生了新的问题,即在模型数目增加时,式(1)的计算量会大大增加,因此必须考虑对式(1)中的模型加以处理.

收稿日期:2003-01-25;**修订日期**:2003-03-20

作者简介:严于鲜(1977—),男,四川省简阳市人,西南交通大学硕士研究生.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1.1 减少模数量

(1) 当模型预测数据的能力远远小于提供最好预测能力的模型时,则在舍去该模型后,对整个兴趣量的影响不大,故可以舍去该模型.即首先选定一个数$c>0$,所有属于

$$B=\{M_k: p(M_k/D)/\max_l\{p(M_l/D)\}\geqslant c\}$$

的模型应舍去.

(2) 当一个模型的子模型预测数据能力大于该模型时,这时子模型更好,从而可以舍去该模型,即在$A-B$中所有属于

$$\{M_l: M_k\ \exists\ M_l, p(M_k/D)/p(M_l/D)\geqslant 1\}$$

的模型应舍去.

通过上述两个步骤,在式(1)中所要考虑的模型数目将大大减少,一般不超过20个.

1.2 马尔可夫链蒙特卡罗方(Monte Carlo)法

设$\overline{M}$为所有上述要考虑的模型构成的模型空间,则我们可以在$\overline{M}$中构造一个马尔可夫链$\{M(t), t=1,2,\cdots,n\}$,让它遍历任何定义在$\overline{M}$上的函数$g(M_t)$,则可在有限不可约马尔可夫链中应用遍历理论得平均值

$$K=\frac{1}{n}\sum_{t=1}^{n}g(M(t))$$

是$E(g(\overline{M}))$的一个估计.对于求兴趣量,仅需取$g(M_t)=p(\triangle/M_t, D)$即可.

下面在$\overline{M}$中构造马尔可夫链.对于任何$M\in\overline{M}$,我们定义M的邻域

$$\mathrm{nbd}(M)=\{M_t\in\overline{M}: M_t\text{中的随机变量之间比}M\text{中的随机变量之间多(少)一个关系或}M_t=M\}.$$

这样就为$\overline{M}$中的元素定义了一种关系,从而可以构成一个马尔可夫链.其转移矩阵q为:当$M_t\in\mathrm{nbd}(M)$时,$q(M\to M_t)\neq 0$,当$M_t\in\mathrm{nbd}(M)$时,$q(M\to M_t)=0$.

2 应用举例

用于模型选择的Bayes方法主要有以下几种.

2.1 双抽样方法

即在同一个样本空间中的两个不同范围,用两种不同的方法抽取样本,然后用式(1)避免模型不确定性的影响,并考虑其数值特征.应用双抽样方法,我们可以在考虑模型的不确定性时,利用先验知识,处理传统方法不能处理的许多问题.如从1970年起,挪威在全国范围内开展了对新生儿先天患道恩斯(Down's)综合症情况的统计,数据由各医院收集,这些数据由医生观察所得,且可能有误.为了保证数据的完整性,在Hordaland地区(约占全国新生儿的15%)采用仪器检查,可假设其不会产生错误,1985~1988年的数据如下表1所示[3].

表1 1985~1988年的数据表

Tab.1 The Data of 1985~1988

项目	医生得到的患者数据R_1	医生得到的正常者数据$\overline{R}_1$
仪器测得的患者数据R_2	8	9
仪器测得的正常者数据$\overline{R}_2$	13	17 847

在该例中,Bayes方法克服了其它模型在分析数据中的两个困难:①虽然其它方法都能很好地解释数据,但它们的期望和方差相差很大,对数据的预测能力也较差.Bayes方法可用式(1)克服模型不确定性对兴趣量的影响.②虽然新生儿患病情况与母亲的年龄有很大关系,但由于引入新的变量会使分析更加复杂,Lie等没有考虑母亲年龄的因素,而Bayes方法可以很容易地考虑母亲年龄对新生儿患病情况的影响.在模型中添加母亲年龄这个随机变量,结合式(1),Bayes方法对道恩斯综合症的数据分析具体结果如表2所示.其中A表示母亲的年龄;S表示道恩斯综合症.箭头上的星号表示R_2可正确检查出而R_1没能检查正确的情况.从该例可以看出,Bayes方法不仅能尽量减小模型不确定性对兴趣量的影响,增加模型预测数据的能力,而且很容易地根据实际情况和对问题认识的不断深入,在模型中添加或减少随机变量,特别是在考虑那些具有隐藏变量的情况下,Bayes方法显示出易于扩展模型的能力.

3.2 复制行动方法

设有一批数据要输入到计算机中,由A,B两个人来完成该任务,则可以对比检查两组数据及其正确性,从而达到复制行动的目的.

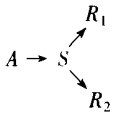
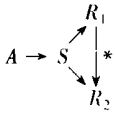
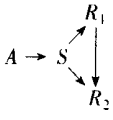
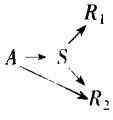
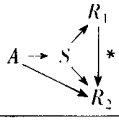
3.3 复制检查方法

如果要检查输入数据的正确性,设正确的数据经检查仍是正确的,让A,B对输入的数据进行核查,又会得到出现错误的数据,如此反复,可得到多组数据,以避免不确定性的影响.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 2　各种模型方法与 Bayes 方法的比较

Tab.2　The comparison between all models and Bayesian method

模型	后验概率	$1000\times p(S)$			$p(\bar{R}_1\mid S)$		$p(\bar{R}_2\mid S)$	
		频 度	均 值	标准差	均 值	标准差	均 值	标准差
	0.282	1.81	1.92	0.292	0.376	0.085	0.555	0.092
	0.385	1.49	1.51	0.129	0.223	0.053	0.470	0.083
	0.269	1.60	1.70	0.252	0.312	0.088	0.513	0.089
	0.03	1.71	1.78	0.226	0.333	0.076	0.518	0.090
	0.016	1.50	1.52	0.129	0.226	0.054	0.517	0.080
Bayes 方法	—	1.54	1.69	0.289	0.292	0.099	0.508	0.095

3　结论

在模型选择中应用Bayes 方法，不仅能够减少模型不确定性的影响，而且可以扩展模型；结合图论的有关知识，可以很容易地考查各随机变量间的关系，增加模型选择的准确性. 因此，这是一个值得深入研究的新方法.

参考文献：

[1] LINHART H. Model Selection[M]. New York: John Wiley & Sons Inc, 1986.

[2] DRAPER D. Assessment and propagation of model uncertainty[J]. Journal of the Royal Statistical Society, 1995, 57(1): 38～40.

[3] LIE R T. A temporary increase of Down syndrome among births of young mothers in Norway: An effect of risk unrelated to maternal age[J]. Genetic Epidemiology, 1991, (8): 217～230.

Application of the Bayesian Method in Model Selection

YAN Yu-xian, YI Jian-hua

(School of Science, Southwest Jiaotong University, Chengdu 610031, China)

Abstract: In many papers about model selection, a common approach is that a model is selected firstly and so a kind of distributions is determined. Then the best distribution is chosen from them. But the uncertainty of model is ignored. This paper presents the theory and application of the Bayesian method in model selection. It proved that Bayesian method can diminish the influence of the uncertainty of model, and can also expand model easily on the basis of the actual use and further insight into the question.

Key words: model selection; Bayesian method; uncertainty

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net