文章编号:1006-3080(2017)04-0559-04 **DOI**:10.14135/j.cnki.1006-3080.2017.04.016

一种二元响应变量模型的分布式贝叶斯估计方法

吴 磊, 钱夕元
(华东理工大学理学院,上海 200237)

摘要:在海量数据背景下,传统的基于单个计算节点的算法很难满足分析要求。考察了一种分布式贝叶斯估计方法,通过在每台机器上单独运行蒙特卡洛抽样并做加权平均可以有效地解决算法效率问题。将该方法应用于基于广义极值模型的二元响应变量回归分析,并探讨其实用性。模拟研究表明分布式算法比传统方法更有效。

关键词:海量数据; 分布式贝叶斯方法; 极值模型
中图分类号:TP301.6 **文献标志码**:A

A Distributed Bayesian Regression Method for Binary Response Massive Data

WU Lei, *QIAN Xi-yuan*
(*School of Science*, *East China University of Science and Technology*, *Shanghai* 200237, *China*)

Abstract: In the background of massive data, it is difficult to meet the analysis requirements for traditional one-node based algorithm. This paper considers a distributed Bayesian estimation method to solve the GEV based general linear regression model by running a separate Monte Carlo algorithm on each machine. The method is applied to regression analysis of binary response variables based on generalized extreme value model. The results show that the proposed distributed Bayesian regression algorithm is much faster than the traditional algorithm in the simulated data sets studying.

Key words: massive data; distributed Bayesian regression; GEV model

逻辑回归模型是在处理二元响应变量数据时最为常用的一种广义线性模型,它采用逻辑分布作为连接函数,可以实现利用连续型解释变量来说明0-1二元响应变量的变化。该模型一般假设潜在变量的概率响应曲线是对称的,即0-1二元响应变量中的各类样本数目基本均衡,但当样本数存在明显不平衡时,逻辑回归模型会严重违背对称性的假设,带来连接函数设定错误,使得模型参数估计存在较大的偏差和均方误差[1]。

不平衡数据在实际应用中并不少见,它一般来源于某类稀有事件或现象发生概率较小的环境,且相关属性的数据会具有明显的偏度特征。为此,学者们提出了大量改进的连接函数用来灵活处理此类数据。最近,Wang等[2]提出了以广义极值(GEV)分布作为连接函数的二元响应变量回归模型,该连接函数比传统GEV分布增加了一个形状参数,新增的形状参数不但没有取值约束,而且可以更大幅度地调节偏度,使得该模型对非对

收稿日期:2016-10-31
基金项目:国家高科技研究发展("863")计划(2015AA20107);上海市经信委"软件和集成电路产业发展专项资金"(140304)
作者简介:吴 磊(1992-),男,上海人,硕士生,主要研究方向为统计计算。
通信联系人:钱夕元,E-mail:xyqian@ecust.edu.cn

称或对称的响应曲线都可以进行拟合,具有了更广泛的灵活性,可以更好地处理二元不平衡数据。

值得注意的是模型的灵活性带来了经典参数估计方法(极大似然估计)求解的困难性。随着马尔科夫链蒙特卡罗(MCMC)方法的发展,贝叶斯估计方法得到了更加广泛的应用。贝叶斯估计方法可以有效地利用先验信息,对小到中型样本问题可以有效改善估计精度,但如何有效地将贝叶斯方法应用到海量数据分析成为近期一个研究热点。随着数据爆炸式增长,单个的处理器已经很难满足人们的需求,一个可以想到的解决办法是将数据分发到多个处理器上,但随之带来的问题就是如何解决各个节点间的信息交换,如何协调好各个处理过程,否则很容易出现死锁或者串行化等问题。贝叶斯方法中广泛采用的 MCMC 方法是基于马氏链构造的,其当前状态转移的概率依赖于前一个状态,这实际上和分布式的思想是有冲突的,因为马氏链要求串行化而分布式要求的是并行化[3]。

本文首先给出了基于极值理论的二元响应变量回归模型及其贝叶斯估计,针对海量数据应用环境,给出了分布式贝叶斯估计算法,设计模拟数据验证了算法的有效性。

1 二元响应变量模型及其贝叶斯估计(Be)

令向量 $\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^{\mathrm{T}}$ 表示 n 个独立的二元 0-1 响应变量,$\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)^{\mathrm{T}}$ 表示独立的解释变量矩阵,其中每个观测值对应的解释变量向量为 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ik})$,向量 $\boldsymbol{\beta}_c=(\beta_1,\beta_2,\cdots,\beta_k)^{\mathrm{T}}$ 表示 k 个回归系数参数。一般地,设定响应变量 Y_i 服从伯努利分布,使得 $Y_i=1$ 的概率为 p_i,而 $Y_i=0$ 的概率为 $1-p_i$。因此,在二元响应变量回归模型中

$$p_i=Pr(Y_i=1\mid\boldsymbol{\beta}_c,\boldsymbol{x}_i)=\Psi(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_c)$$

$\Psi(\cdot)$是一个取值在 0 和 1 之间的非负函数,标准情况下,设定 $\Psi(\cdot)$为累积分布函数,而称 $\Psi^{-1}(\cdot)$为连接函数。

常用二元响应变量模型的连接函数有 logit 连接 $\Psi^{-1}(p_i)=\lg\{p_i/(1-p_i)\}$,probit 连接 $\Psi^{-1}(p_i)=\Phi^{-1}(p_i)$,$\Phi^{-1}$为标准正态分布的反函数,cloglog 连接 $\Psi^{-1}(p_i)=-\lg\{-\lg(p_i)\}$等。上述连接函数都是对称的,在处理不平衡数据时会出现较大的偏差和均方误差。

Wang 等[2]提出了基于广义极值分布的二元响应变量回归模型,其连接函数采用如下累积分布函数,具体表示为:

$$G(x)=\exp\left[-\left\{1+\xi\frac{(x-\mu)}{\sigma}\right\}_+^{-1/\xi}\right]$$

ξ为形状参数,用以改变模型分布的偏度和尾部厚度。根据 ξ的不同,该模型既可以表现出对称性,也可以表现出非对称性,可以很好地用来处理非平衡样本数据[4]。本文采用基于 MCMC 算法的贝叶斯方法进行模型的参数估计,其 Metropolis-Hastings (M-H)抽样算法描述如下:

Step 0 选取待估参数的初始值$\boldsymbol{\theta}_0$;

Step 1 从产生候选参数的密度函数 $q(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^*)$中获得候选参数 $\boldsymbol{\theta}^*$;

Step 2 计算候选参数被接收的概率 $\alpha(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^*)$;

Step 3 以 $\alpha(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^*)$的概率设 $\boldsymbol{\theta}^{(s)}=\boldsymbol{\theta}^*$,或者以 $1-\alpha(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^*)$的概率设 $\boldsymbol{\theta}^{(s)}=\boldsymbol{\theta}^{(s-1)}$;

Step 4 重复 Step 1,Step 2 和 Step 3 S 次;

Step 5 以 $\boldsymbol{\theta}_0,\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_S$ 作为参数 $\boldsymbol{\theta}$ 的近似后验密度,作均值估计得参数 $\boldsymbol{\theta}$ 的点估计。

其中接受概率

$$\alpha(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^*)=\min\left[\frac{p(\boldsymbol{\theta}=\boldsymbol{\theta}^*\mid\boldsymbol{y})q(\boldsymbol{\theta}^*;\boldsymbol{\theta}=\boldsymbol{\theta}^{(s-1)})}{p(\boldsymbol{\theta}=\boldsymbol{\theta}^{(s-1)}\mid\boldsymbol{y})q(\boldsymbol{\theta}^{(s-1)};\boldsymbol{\theta}=\boldsymbol{\theta}^*)},1\right]$$

式中,$p(\boldsymbol{\theta}=\boldsymbol{\theta}^*\mid y)$表示后验密度在点 $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ 的取值,$q(\boldsymbol{\theta}^*;\boldsymbol{\theta}=\boldsymbol{\theta}^{(s-1)})$表示随机变量 $\boldsymbol{\theta}$ 的密度函数在 $\boldsymbol{\theta}=\boldsymbol{\theta}^{(s-1)}$处的取值。

2 分布式贝叶斯估计(DBe)方法

2.1 模型参数估计

在海量数据背景下,上述估计方法将变得十分困难。本文提出借助分布式贝叶斯方法对模型参数进行估计。该方法的主要思想是根据现有的计算资源,在确保每份数据集的样本容量足够的情况下合理地把样本数据拆分,为每份数据分配一个独立的计算节点做蒙特卡洛抽样,从而得到参数的贝叶斯后验分布,最后根据一定的方式把每份数据的后验分布整合成一个全局的后验分布,其主要过程描述如下:

记 $\boldsymbol{y}$ 为全部的样本数据(样本数为 n),$\boldsymbol{y}_s$ 是第 s 份数据(样本数为 n_s),记 $\boldsymbol{\theta}$ 为待估参数。假设数据

集间相互独立，则根据贝叶斯公式：

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \prod_{s=1}^{S} p(\boldsymbol{y}_s \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})^{1/S} \tag{1}$$

从式(1)中可以看到，每个部分的先验分布变成了总体先验的 S 次方根，这是为了保持整个系统中的先验信息保持不变。另外，由于对先验信息并不是很了解，本文采用了方差较大的无信息正态先验。根据模型可知，似然函数为：

$$p(\boldsymbol{y}_s \mid \boldsymbol{\theta}) = \prod_{i=1}^{n_s} y_i F_{\mathrm{GEV}}(\beta_0 + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_c; \xi) + (1 - y_i) \cdot [1 - F_{\mathrm{GEV}}(\beta_0 + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_c; \xi)]$$

由于后验分布比较复杂，本文采用自适应的正态随机游走来逼近该目标后验，其优势是可以提高 Metropolis-Hastings 迭代过程的接受率，根据上一次的参数是否被接受来调整随机游走的步长，从而自适应地保证更高的接受率[5]。

假定在第 s 个计算节点上得到了总共 G 个 Metropolis-Hastings 抽样，以下述加权平均的方法得到总共 S 个计算节点的全局后验分布参数估计：

$$\boldsymbol{\theta}_g = \left(\sum_{s=1}^{S} \boldsymbol{W}_s\right)^{-1} \sum_{s=1}^{S} \boldsymbol{W}_s \boldsymbol{\theta}_{sg},\ g = 1, 2, \cdots, G$$

式中，$\boldsymbol{W}_s$ 称为权重矩阵，一般可以取样本方差的逆或在参数 $\boldsymbol{\theta}$ 维数较高时作简单的平均即可。DBe 估计的步骤如下：

Step 1　把样本数据 $\boldsymbol{y}$ 分片为 $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_s$；

Step 2　由分离先验信息 $p(\boldsymbol{\theta})^{1/S}$，重复 S 次独立的 M-H 抽样(该部分算法步骤如上)得 $\boldsymbol{\theta}_{sg} \sim p(\boldsymbol{\theta} \mid \boldsymbol{y}_s)$，$g = 1, \cdots, G$；

Step 3　对 $\boldsymbol{\theta}_{sg}$，$s = 1, \cdots, S$；$g = 1, \cdots, G$ 加权，得到全局的后验分布参数 $\boldsymbol{\theta}_g$，$g = 1, \cdots, G$；

Step 4　根据接收率 α，过滤掉马尔科夫链的前$(1-\alpha)$ G 个参数；

2.2　其他非参数估计策略

根据样本方差加权具有简单稳定，符合直观感觉的优点。但根据贝叶斯公式(参见式 1)，可以利用诸如核密度估计(Kernel density estimate)等非参数估计方法直接求出每份数据的经验估计 $p(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}_s)$，从而得到全局的经验后验密度 $p(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) \propto \prod_{s=1}^{S} p(\hat{\boldsymbol{y}_s} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})^{1/S}$。这种方法在简单模型中比较有优势，但是随着参数 $\boldsymbol{\theta}$ 维数的增加，核密度估计的可靠性会下降[3]。

3　模拟研究

取样本容量 $n = 100\,000$，解释变量取 3 个，均由服从均值为 0、方差为 1 的正态随机数生成，模型参数的真值为 $\beta_0 = 1$，$\beta_1 = 2$，$\beta_2 = 3$，$\beta_3 = 4$，分布参数的真值 $\xi = 2$，产生模拟数据集 50 份。根据模拟数据集的样本容量以及计算资源，设定 $S = 12$，$G = 1\,250$，$\alpha = 0.8$。

数据分布式处理环境如下：Hadoop2. x：1 个 Master 节点和 3 个 Slave 节点，集群中 CPU 的数量 Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00 GHz * 8 * (1[M]+3[S])=32core，内存大小 32 (1[M]+3[S])=128 GB，JVM 参数设定-Xmx768 M，其中 Container 的大小为 1 GB[6]。

表 1 给出了分布式贝叶斯方法和传统分析方法的对比结果。从中可以看到，分布式贝叶斯方法优势明显，在本文的计算节点上，计算效率提升了近 3 倍。但由于计算时间取决于最慢的节点，且 Hadoop 环境下的 Map 阶段(将数据拆分成 12 份)并没有并行化而是仅使用了一个节点，因此计算效率并不呈线性提升。但是随着数据量的进一步增大，分布式贝叶斯方法的优势会更趋明显。另外，在准确性和稳定性上，当 M-H 抽样的初值为 0，自适应随机游走的步长分别是 0.005 和 0.1 的情况下，50 次的重复实验中分布式贝叶斯方法和传统分析方法的参数估计精度差别不大，参见图 1 的比较结果。

表 1　算法对比分析结果

Table 1　Comparative analysis of distributed Bayes vs. classical method

MH Sampling	CPU Time/s	ξ	β_0	β_1	β_2	β_3
DBe	46	1.924 6 (0.066 9)	0.938 8 (0.036 0)	1.731 8 (0.044 9)	2.616 3 (0.069 5)	3.488 3 (0.091 5)
Be	124	1.925 7 (0.050 8)	0.947 7 (0.036 4)	1.741 8 (0.044 1)	2.632 2 (0.068 2)	3.510 1 (0.088 3)

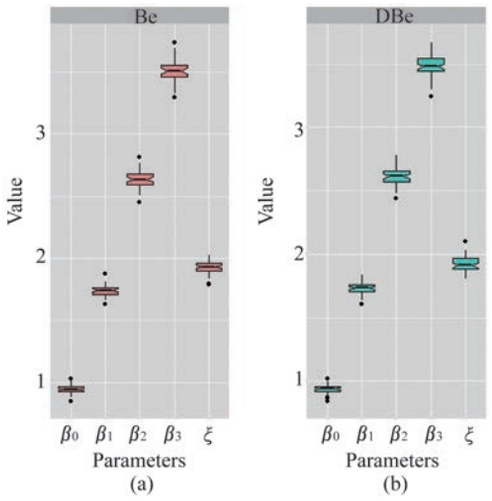

图1 分布式贝叶斯方法和传统方法参数估计精度对比分析结果

Fig. 1 Comparative analysis of distributed Bayes vs. classical method

4 结束语

本文针对不平衡二元响应变量的海量数据,基于广义极值分布进行回归分析,借助于分布式贝叶斯方法(Metropolish-Hastings抽样算法)进行参数估计,模拟研究表明该方法具有一定的计算优势,应用于模拟数据分析中取得了较好的效果。今后可以进一步推广到基于Gibbs抽样的分布式贝叶斯分析中。

参考文献:

[1] CZADO C, SANTNER T J. The effect of link misspecification on binary regression inference [J]. Journal of Statistical Planning and Inference, 1992, 33(2): 213-231.

[2] WANG X, DEY D K. Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption [J]. The Annals of Applied Statistics, 2010, 4(4): 2000-2023.

[3] STEVEN S L, BLOCKER A W, BONASSI F V. Bayes and big data: The consensus Monte Carlo algorithm [J]. International Journal of Management Science and Engineering Management, 2016, 11(2): 78-88.

[4] GHOSH S K, MUKHOPADHYAY P, LU J C. Bayesian analysis of zero-inflated regression models [J]. Journal of Statistical Planning and Inference, 2006, 136(4): 1360-1375.

[5] CHIPMA H A, GEORGE E I, MCCULLOCH R E. BART: Bayesian additive regression trees [J]. The Annals of Applied Statistics, 2010, 4(1): 266-298.

[6] DEAN J, GHEMAWAT S. Mapreduce: Simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51(1): 107-113.

(上接第532页)

[10] GANDOMI A H, YANG X S. Chaotic bat algorithm [J]. Journal of Computational Science, 2014, 5(2): 224-232.

[11] NIKNAM T, BAVAFA F, AZIZIPANAH-ABARGHOOEE R. New self-adaptive bat-inspired algorithm for unit comi--tment problem [J]. IET Science Measurement Technology, 2014, 8(6): 505-517.

[12] WANG G G, CHU H C E, MIRJALILI S. Three-dimensional path planning for UCAV using an improved bat algorithm [J]. Aerospace Science & Technology, 2016, 49: 231-238.

[13] AFRABANDPEY H, GHAFFARI M, MIRZAEI A, *et al*. A novel bat algorithm based on chaos for optimization tasks [C]//Iranian Conference on Intelligent Systems. USA: IEEE, 2014: 1-6.

[14] RAGHAVAN S, MARIMUTHU C, SARWESH P, *et al*. Bat algorithm for scheduling workflow applications in cloud [C]// International Conference on Electronic Design, Computer Networks & Automated Verification. USA: IEEE, 2015: 139-144.

[15] YıLMAZ S, KÜÇÜKSILLE E U. A new modification approach on bat algorithm for solving optimization problems [J]. Applied Soft Computing, 2014, 28(5): 259-275.

[16] YOUNG G F, SCARDOVI L, CAVAGNA A, *et al*. Starling flock networks manage uncertainty in consensus at low cost [J]. PloS Computational Biology, 2013, 9(1): e1002894.

[17] NETJINDA N, ACHALAKUL T, SIRINAOVAKUL B. Particle swarm optimization inspired by starling flock behavior [J]. Applied Soft Computing, 2015, 35(C): 411-422.

[18] XIE J, ZHOU Y, CHEN H. A novel bat algorithm based on differential operator and Lévy flights trajectory [J]. Computational Intelligence & Neuroscience, 2013(2013): 453-812.

[19] 彭泓,丁玉成. 基于遗传交叉因子的蝙蝠算法的改进[J]. 激光杂志, 2015(2): 23-26.

[20] 刘金琨. 先进PID控制MATLAB仿真[M]. 北京:电子工业出版社, 2004.