第15卷 第16期 2015年6月
1671—1815(2015)16-0077-07

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 15 No. 16 Jun. 2015
ⓒ 2015 Sci. Tech. Engrg.

一种面向纵向数据的 RE-BET 算法及应用

鲁 庆 穆志纯*
(北京科技大学自动化学院,北京 100083)

摘 要 混合效应模型是分析纵向数据的有效方法,但模型的线性结构限制了其适应现实数据的能力。提出了一种 RE-BET 算法及其变形的 RE-BEBT 算法,采用树形方法估计混合效应模型的固定效应,可以自动选择重要变量,能更好地发现和描述变量间关系;采用基于 Dirichlet 过程先验的贝叶斯方法估计混合效应模型的随机效应,使模型可以适用于小样本数据。以低合金钢和碳钢的海水腐蚀数据为例,通过与实验数据和其他算法的计算结果对比分析,验证了 RE-BET 算法可行性和有效性。

关键词 纵向数据 Dirichlet 过程 树形方法 腐蚀数据 贝叶斯
中图法分类号 TP391.9; **文献标志码** A

纵向数据是指在不同时间点上,对研究对象进行多次重复测量而获取的实验数据。由于每一个研究对象的多次观测间具有自相关性,而且不同对象间的测量数据,其分布通常存在差异,因此纵向数据不满足传统线性模型关于样本独立和分布同质的假设。

对于纵向数据的分析,主要有时间序列分析[1]、方差分析[2]和混合效应模型[3]等方法。其中时间序列分析方法适合于处理长序列的时间次序数据;方差分析方法适合于处理固定时点的纵向数据,要求每一个研究对象具有相同数量的观测,并且观测的时间间隔相同;混合效应模型则可以处理非固定时点的纵向数据,它使用随机效应描述研究对象间的差异,能够反映纵向数据的内部结构,在社会科学、生物、金融[4]等领域得到广泛应用。

尽管混合效应模型的发展取得了很大成功,但该模型将固定效应和随机效应均设定为线性,这一假设不能充分描述出数据中各变量的交互关系。针对此问题,提出了一种新的 RE-BET(random effects bayesian estimation tree)算法,其基本思想是应用树形算法估计混合模型中的固定效应部分,应用半参数贝叶斯方法估计模型中的随机效应部分,以提高模型估值的准确性,并改善小样本数据的随机效应参数估计问题。

2015年1月23日收到 国家科技基础性工作专项基金(2012FY113000)资助

第一作者简介:鲁 庆(1975—),男,博士研究生。研究方向:数据挖掘、系统建模。E-mail:luqing_qhd@hotmail.com。

* 通信作者简介:穆至纯,教授,博士生导师。研究方向:复杂系统建模与智能控制、信息自动化。E-mail:mu@ies.ustb.edu.cn。

1 混合效应模型

假设某个纵向数据集中包含 n 个实验对象,任意编号为 i 的实验对象的观测记录为 n_i 次,则混合效应模型可以表示为

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\varepsilon}_i \tag{1}$$

式(1)中 $i=1,\cdots,n$;$\boldsymbol{Y}_i = [y_{i1},y_{i2},\cdots,y_{in_i}]^{\mathrm{T}}$,为 $n_i\times 1$ 的模型响应变量;$\boldsymbol{X}_i = [x_{i1},x_{i2},\cdots,x_{in_i}]^{\mathrm{T}}$,为表示固定效应变量;$\boldsymbol{\beta} = [\beta_1,\beta_2,\cdots,\beta_p]^{\mathrm{T}}$,为固定效应;$\boldsymbol{Z}_i$ 为 $n_i\times q$ 矩阵,表示随机效应变量;$\boldsymbol{u}_i = [u_{i1},u_{i2},\cdots,u_{in_i}]^{\mathrm{T}}$,$q\times 1$ 矩阵,表示对象 i 的随机效应;$\boldsymbol{\varepsilon}_i$ 为 $n_i\times 1$ 的误差向量,为便于计算,通常假设 $\boldsymbol{u}_i \sim N_q(0,D)$,$\boldsymbol{\varepsilon}_i \sim N(0,\sigma_\varepsilon^2\boldsymbol{I}_{ni})$;实验中所有对象的观测总数量 $N = \sum_{i=1}^{n} n_i$。

混合效应模型的参数估计过程通常采用基于最大似然估计的 EM 算法来实现[5]:将 $\boldsymbol{Y}_i$ 视为已知的观测数据,将未知的 $\boldsymbol{u}_i$ 视为缺失数据,完整的观测数据为($\boldsymbol{Y}_i,\boldsymbol{u}_i$),需要估计的参数为 β、D 和 σ_ε^2。在 EM 算法的 E 步,根据当前参数的估计值,计算完整数据的似然函数期望;在 M 步,通过最大化似然函数,对参数的估计进行更新。通过多次迭代,使得模型的参数估计逼近真实参数。

在式(1)中,混合效应模型中的固定效应和随机效应与响应变量 $\boldsymbol{Y}_i$ 之间被设定为线性关系,但在实质上,这种线性关系的假设是否恰当是未知的。如果将式(1)扩展,可以得到如式(2)所示的模型

$$\boldsymbol{Y}_i = f(\boldsymbol{X}_i) + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\varepsilon}_i \tag{2}$$

式(2)中,任意函数 $f(\boldsymbol{X}_i)$ 代替了混合效应模型中的固定效应部分 $\boldsymbol{X}_i\beta$,在这个框架下,很多经典数据挖掘算法都可以用来估计函数 $f(\boldsymbol{X}_i)$,包括树形算法、

广义回归模型,神经网络、支撑向量机等等,这种扩展将提高模型发现和描述非线性关系和变量间交互效应的能力,提高模型的预测性能。在式(2)中,模型的随机效应部分 $Z_i u_i$ 仍然假定与 Y_i 为线性关系,使得模型可以反映出不同研究对象之间的差异性。

如前所述,纵向数据的每一个研究对象 i 都拥有 n_i 次观测记录,而不同研究对象之间的观测存在差异性,混合效应模型将这种差异性表达为随机效应。将式(2)变形,如式(3)所示

$$Y_i - f(X_i) = Z_i u_i + \varepsilon_i \tag{3}$$

可以看出,随机效应 $Z_i u_i$ 项的存在是必要的,否则 $Z_i u_i$ 项将被注入到模型偏差中,导致 Y_i 的估值的偏差变大。

对于如式(2)所表述的非线性混合效应模型,文献[6]提出了一种 MERT(Mixed Effects Regression Tree)算法,采用分类回归树(CART)对固定效应 $f(X_i)$ 进行拟合,应用基于最大似然的 EM 算法估计随机效应 u_i 以及 ε_i。文献[7]独立提出了 RE-EM(random effects/EM)trees 算法,其思想与 MERT 算法基本一致,但采用了基于约束最大似然估计的 EM 算法对混合线性模型随机效应部分进行参数估计。

MERT 和 RE-EM trees 算法采用 EM 算法进行参数估计,其焦点参数的推断依赖于其他未知的参数的点估计,Raudenbush 等人[8]证明,当实验对象数量 n 较大时,随着实验数据样本容量增加,采用 EM 算法的参数的估值将收敛于参数的真值;但对于小样本数据,参数估值可能相当不准确。因此,这两种算法均不能很好地处理研究对象数量 n 较小时的数据。关于混合效应模型所需的样本容量,目前尚没有定论,文献[9]提出了 30/30 准则,即要得到较可靠的估计,数据至少要包括 30 个研究对象,每个研究对象包括 30 个观测值;Menard[10]则建议研究对象数量 $n > 100$。

2 RE-BET 算法

为了改善混合效应模型在小样本情况下的性能,Seltzer 等人[11]应用完全贝叶斯方法对参数进行推断。根据贝叶斯观点,观测数据是由某些未知参数定义的随机分布产生的,对于每一个未知参数,首先要确定先验分布,然后将观测数据和先验分布联合起来,产生在这一观测数据下的联合后验分布,对于具体的未知参数的推断可以通过对其他未知量(如辅助参数)的所有可能值取平均数来得到,贝叶斯方法充分考虑到了参数点估计的不确定性,使焦点参数的估计更为稳健。

对于形如式(1)的模型,假设已知后验联合分布 $f(\beta,\sigma_\varepsilon^2,D\,|\,Y)$,如果要估计参数 β,则需要计算

$$f(\beta\,|\,Y) = \iint f(\beta,\sigma_\varepsilon^2,D\,|\,Y)\,\partial(D)\,\partial(\sigma_\varepsilon^2) \tag{4}$$

在实际问题中,很难得到贝叶斯后验分布的显式表示,高维数据的数值积分计算也存在困难。文献[11]采用马氏链蒙特卡罗(Markov Chain Monte Carlo,MCMC)方法成功解决了混合效应模型的参数估计问题。但该方法和 EM 算法存在同一个瑕疵:为便于计算,假设随机效应为正态分布,$u_i \sim N_q(0,D)$。由于随机效应 u_i 的分布实质上是未知的,这一假定限制了模型对于现实数据的适应性。

为了改善这一问题,文献[12]采用半参数贝叶斯方法对混合效应模型进行参数估计,将随机效应的先验分布从参数分布扩展到非参数的 Dirichlet 过程分布,并利用 MCMC 抽样法推断其后验分布,以期获得更为稳健的估计结果。文献[13]将半参数混合效应模型应用于绵羊种群的研究。

与文献[6,7]相比,本文提出的 RE-BET 算法借鉴了文献[12]的思想,在采用非线性混合效应模型框架,应用树形算法估计模型的固定效应的同时,应用基于 Dirichlet 过程分布的半参数贝叶斯方法估计模型的随机效应,使随机效应的估计不局限于正态分布假设,更好地适应现实数据。

下面对 RE-BET 算法中随机效应的估计过程进行说明。

由式(2)可知,

$(Y_i\,|\,u_i,\sigma_\varepsilon^2) \sim N_{n_i}[f(X_i) + Z_i u_i,\sigma_\varepsilon^2 I_{n_i}]$。

首先,确定各参数的先验分布:

对于 σ_ε^2,令 $\sigma_\varepsilon^{-2} \sim \text{gamma}(a_0,\lambda_0)$,其中 a_0,λ_0 为 gamma 分布的形状参数和尺寸参数。

对于随机效应 u_i,本文采用 *Dirichlet* 过程定义其先验分布。即:

$P(u_i) = G$; $\quad G \sim DP(M.\,G_0)$。

式中随机效应 u_i 的先验分布为 G,其分布为 $DP(M.\,G_0)$。M 为 Dirichlet 过程的聚集参数,G_0 为基础分布。本文假设 $G_0 = N(0,D)$,其中 $P(D) \propto$ constant。

在确定先验分布后,计算各参数的后验分布

以 $\hat{f}(X_i)$ 表示树形算法对式(2)中 $f(X_i)$ 的估计,σ_ε^2 后验分布为

$$P(\sigma_\varepsilon^2\,|\,u,D,Y) \propto \left[\prod_{i=1}^{n}\left(\frac{1}{\sigma_\varepsilon^2}\right)^{\frac{n_i}{2}}\right]\exp\left[-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i) - Z_i u_i)'(Y_i - \hat{f}(X_i) - Z_i u_i)\right] \tag{5}$$

随机效应 u_i 的后验分布采用 Polya urn 构造进

行推断。Polya urn 构造是 Dirichlet 过程的一种构造方法,其基本工作原理是首先从基础分布 G_0 中抽取 u_1;然后抽取 u_2,其中 u_2 将以概率 p_1 等于 u_1,或以概率 $p_0 = 1 - p_1$ 从分布 G_0 中抽取;当抽取 u_3 时,u_3 以概率 p_1 等于 u_1,以概率 p_2 等于 u_2,以概率 $p_0 = 1 - (p_1 + p_2)$ 从分布 G_0 中抽取;以此类推,其中概率 p_i 由 Dirichlet 过程的参数决定。关于 Dirichlet 过程和 Polya urn 构造的详细解释请参见文献[14,15]。

按照 Polya urn 构造的定义,可以得到随机效应 u_i 的后验分布如下[13]。

(1)按照概率

$$\frac{N[Y_i \mid \hat{f}(X_i) + Z_i u_j, \sigma_\varepsilon^2 I_{n_i}]}{I_i + \sum_{j \neq i} N[Y_i \mid \hat{f}(X_i) + Z_i u_j, \sigma_\varepsilon^2 I_{n_i}]} \times$$

$$P(u_i \mid \sigma_\varepsilon^2, D, u_{-i}, Y) = u_j。$$

(2)按照概率

$$\frac{I_i}{I_i + \sum_{j \neq i} N[Y_i \mid \hat{f}(X_i) + Z_i u_j, \sigma_\varepsilon^2 I_{n_i}]} \times$$

$$(u_i \mid \sigma_\varepsilon^2, D, u_{-i}, Y) \sim$$

$$N\left[\left(Z'_i Z_i + \frac{\sigma_\varepsilon^2}{D}\right)^{-1} Z_i \times [Y_i - \hat{f}(X_i)], \left(Z'_i Z_i + \frac{\sigma_\varepsilon^2}{D}\right)^{-1} \sigma_\varepsilon^2\right] \quad (6)$$

式(6)中

$$u_{-i} = (u_1, u_2, \cdots, u_{i-1}, u_{i+1}, \cdots, u_n);$$

$$I_i = M(2\pi)^{-\frac{1}{2}n_i} D^{-\frac{1}{2}} Q_i^{\frac{1}{2}} (\sigma_\varepsilon^2)^{-\frac{1}{2}n_i} \exp \frac{1}{2\sigma_\varepsilon^2}\{[Y_i - \hat{f}(X_i)] \times U_i [Y_i - \hat{f}(X_i)]\};$$

$$Q_i = \left(\frac{1}{\sigma_\varepsilon^2} Z'_i Z_i + \frac{1}{D}\right)^{-1};$$

$$U_i = \left(\frac{1}{\sigma_\varepsilon^2} Z'_i Q_i Z_i - I_{n_i}\right)。$$

按照 Polya urn 的构造过程,将具有相同 u_i 的研究对象进行归类,将形成 k 个聚类,$0 < k < n$。定义 $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_k)$ 为聚类的值,关于 D 的后验分布,文献[13]证明:

$$P(D \mid \gamma, Y) \propto \left(\frac{1}{D}\right)^{\frac{k}{2}} \exp\left(-\frac{1}{2D}\gamma'\gamma\right) \quad (7)$$

在得到边际分布的显式表达式后,使用 Gibbs 抽样对各参数进行后验估计。

完整的 RE-BET 算法描述如下。

1)令 $m = 0$,选择 $u_i^{(0)}$、$\sigma_\varepsilon^{2(0)}$、$D^{(0)}$ 的初始值。

2)更新 $\hat{Y}_i^{(m)}$:$\hat{Y}_i^{(m)} = Y_i - Z_i u_i^{(m)}$。

3)以 $\hat{Y}_i^{(m)}$ 为响应变量,以 $X_i (i = 1, \cdots, n)$ 为预报变量,采用树形算法,如 CART,训练树模型 $\hat{f}(X_i)$,用于拟合固定效应 $f(X_i)$。注意在树的训练过程中,需使用全部数据,而非只使用本研究对象的观测数据。

4)根据式(5),从 $(\sigma_\varepsilon^2 \mid u^{(m)}, D^{(m)}, Y)$ 中抽取 $\sigma_\varepsilon^{2(m+1)}$。

5.1)根据式(6),从 $(u_i \mid \sigma_\varepsilon^{2(m+1)}, D^{(m)}, u_{-i}^{(m)}, Y)$ 中抽取 $u^{(m+1)}$。

⋮

5.n)根据式(6),从 $(u_n \mid \sigma_\varepsilon^{2(m+1)}, D^{(m)}, u_{-n}^{(m)}, Y)$ 中抽取 $u_n^{(m+1)}$。

6)根据式(7),从 $(D \mid \gamma^{(m)}, Y)$ 中抽取 $D^{(m+1)}$。

7)令 $m = m + 1$,回到第(2)步。不断重复这一过程,直至生成的马氏链收敛到平稳分布。然后,将生成链前面的 $m - 1$ 次数据作为预烧数据,舍弃不用,后面的 m 次迭代的数值保存下来作为各参数的后验样本。

8)在得到随机效应的估计 $\hat{u}_i$ 后,令 $\hat{Y}_i = Y_i - Z_i \hat{u}_i$,以 CART 算法在数据集 $(X_i, \hat{Y}_i)$ 上训练模型,并使用 1 - SE(1 标准差)原则进行剪枝,结果作为对于式(2)中固定效应 $f(X_i)$ 的估计。

按照 Dirichlet 过程的定义,聚集参数 M 反映了对于 $E(G) = G_0$ 的先验信任,当 $M \to \infty, G \to G_0$。对于聚集参数 M 的选择,可以依据先验信任程度人为设定,Pretorius[13] 将 M 认为是随机变量,令 $M \sim \text{gamma}(c_0, d_0)$,其中 c_0, d_0 为 gamma 分布的形状参数和尺寸参数,并推导了 M 的后验分布公式,详细说明请参见文献[13]。

3　数据实验

为了检验 RE-BET 算法的有效性,将其应用于金属材料在自然环境的腐蚀预测研究,所使用的数据选自青岛、舟山、榆林、厦门四地的海水腐蚀实验数据,数据包含了以 1、2、4、8、16(10)年为周期对 15 种牌号(详见表 1)的低合金钢和碳钢在不同自然环境下的实验测量记录。数据记录有部分缺失,每种牌号的材料分别有 3 ~ 5 条记录,数据集样本容量为 86 个。随机抽取其中 70 个样本作为训练集,其余作为检验集。

定义用于模型固定效应的变量(X_i)共 8 个,为实验区(Region,全浸、潮差、飞溅)、暴露时间(Time,年)、平均温度(Temperature,℃)、溶解氧(mol·L^{-1})、盐度(Salt,g·L^{-1})、pH 值、流速(mm·s^{-1})、生物附着物(%);用于模型随机效应的变量(Z_i)共

二个，为暴露时间（年）和截距；模型的输出为试件的海水腐蚀速率（$\mu m \cdot a^{-1}$）。通过设置 Z_i，考察不同牌号试件的海水腐蚀速率发展是否存在随机效应，如何发挥影响。

表 1 实验材料的牌号

Table 1 The sign of experimental materials

分类	材料牌号		
低合	09CuPTiRE	09MnNb	10CrCuSiV
金刚	15MnMoVN	15MnTi	16Mn
	10CrMoAl	12CrMnCu	14MnMoNbB
	3C	CF	D36
	E2		
碳钢	08Al	A3	

由于缺少明确的先验信息，对各参数采用了相对平坦的先验分布：令 $a_0 = 1$，$\lambda_0 = 0.001$；M 为聚集参数，令 M 等于 0.5、10、10 000 以及采用文献[13]算法的估计值，计算 M 估计值时设置 $c_0 = 1$，$d_0 = 0.001$。

在实验中，RE-BET 算法的 Gibbs 抽样过程运行 10 000 次，将前 5 000 次迭代作为预烧期去掉，对于后面的 5 000 次迭代，每间隔 5 次抽取样本，以消除相关性，得到后验样本总数为 1 000 个，在此样本基础上对各参数进行估计。采用 Geweke 等人[14]的方法检验马氏链的收敛性。如图 1 的“RE-BET”图所示，各估计参数的 Geweke 值均在（-1.96，1.96）范围之内，可以认为马氏链收敛性是可以接受的。图 2 是关于腐蚀率模型的 RE-BET 树。

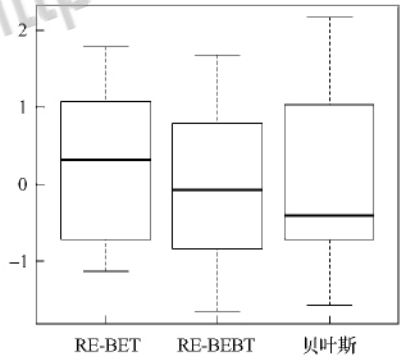

图 1 关于贝叶斯混合效应模型、RE-BET、RE-BEBT 算法的马氏链 Geweke 值分布

Fig. 1 The Geweke statistics of Markova chain about Bayesian mixed effects model、RE-BET and RE-BEBT algorithm

从图 2 中可以看出，试验区、暴露时间、海水平均温度和盐度是低合金钢和碳钢在海水中腐蚀率的主要影响因素。表 2 报告了不同 M 值的截距和暴露时间变量的随机效应参数估计。

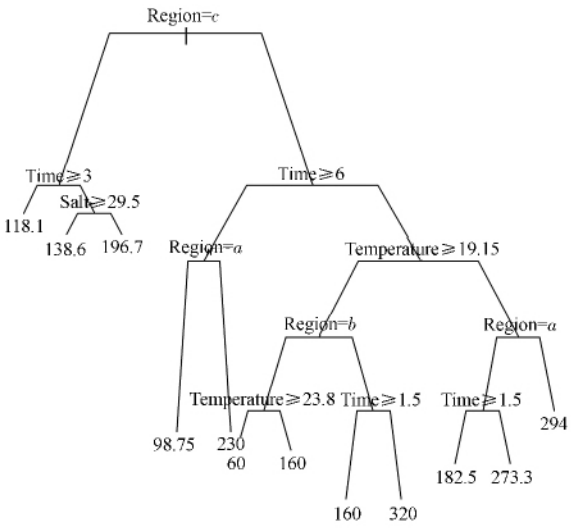

图 2 腐蚀率模型的 RE-BET 树

Fig. 2 The RE-BET tree of corrosion rate model

表 2 基于 RE-BET 算法模型的随机效应参数估计

Table 2 The random effects estimation of RE-BET model

材料牌号	随机效应							
	$M = 0.5$		$M = 10$		$M = 10\,000$		M 估计值	
	截距	暴露时间	截距	暴露时间	截距	暴露时间	截距	暴露时间
09CuPTiRE	-5.444	0.722	-21.475	1.555	-31.823	2.159	-11.620	1.116
08Al	-2.215	0.380	13.046	-1.498	20.258	-2.109	2.943	-0.233
09MnNb	-4.536	1.497	-12.747	4.474	-15.890	5.287	-8.192	2.623
10CrCuSiV	-58.943	2.980	-156.732	8.008	-159.426	8.228	-130.315	6.670
10CrMoAl	-4.394	0.684	-6.603	0.654	-6.913	0.557	-5.963	0.843
12CrMnCu	-3.997	0.840	-1.944	1.169	2.672	0.650	-4.248	1.129
14MnMoNbB	-4.740	0.822	-14.001	1.725	-19.779	2.165	-8.402	1.288
15MnMoVN	-3.579	1.113	6.427	1.632	16.605	1.018	-1.638	1.514
15MnTi	-22.733	1.150	-66.039	3.146	-70.654	3.539	-46.520	2.262
16Mn	-4.038	0.937	-2.183	1.586	1.961	1.227	-4.368	1.348
3C	28.640	-2.765	69.990	-7.354	71.241	-7.445	60.917	-6.038
A3	-4.213	0.892	-3.938	1.688	-0.283	1.342	-4.879	1.317
CF	-4.746	0.781	-15.011	1.704	-24.075	2.378	-8.312	1.191
D36	-4.780	0.726	-13.066	1.206	-16.718	1.127	-8.549	1.074
E2	36.661	-2.133	105.138	-5.905	106.809	-5.559	87.700	-5.518

从表 2 可以看出，$M = 0.5$ 时，随机效应截距的方差为 6 095.37，暴露时间斜率的方差为 26.24；$M = 10$ 时，这二个值为 45 427.31 和 192.13；$M = 10\,000$ 时，这二个值为 49 093.90 和 206.16；M 为估计值时，这二个值为 30 513.52 和 129.284。不同牌号材料的随机效应参数存在着明显的差异。因此对于本文的实验数据，随机效应项的设置对于合理描述数据结构是有益的。

按照 Dirichlet 过程的定义，较小的 M 值反映了对于基础分布的低信任度，反之，随着 M 值的增大，随机效应将接近基础分布。在本文中，按照文献[13]的方法，M 的估计值为 2.027，反映了随机效应的估计分布与基础分布（正态分布）有较大的偏离。

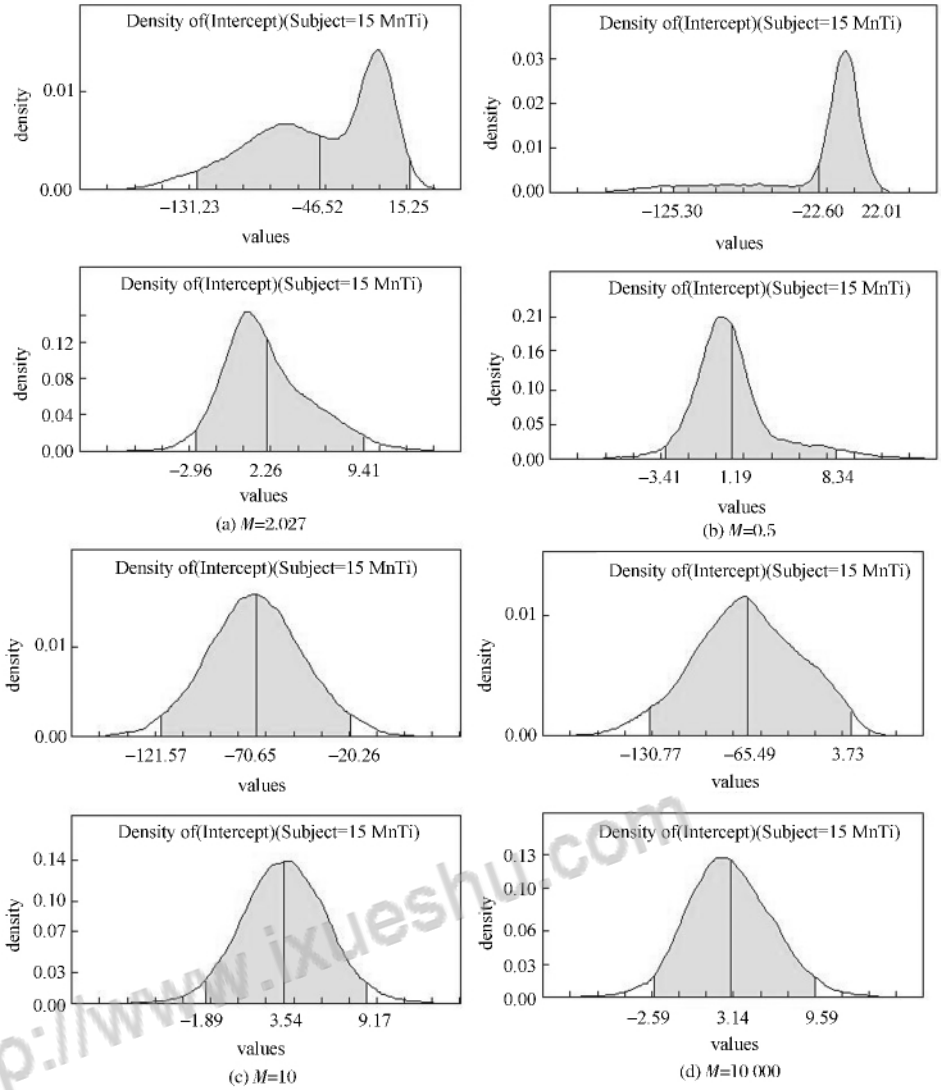

图 3　基于 RE-BET 算法模型的密度函数(15MnTi)

Fig. 3　The density function estimation of based on RE-BET model(15MnTi)

图 3 以 15MnTi 为例,报告了随机效应密度函数估计的结果,可以看出,不同的 M 值,对随机效应分布的估计影响很大。

图 4　报告了 RE-BET 算法在海水腐蚀数据集训练模型的时间。由于数据样本较小,CART 训练和 Gibbs 抽样运行均较快,迭代 10 000 次的时间为 70.24 s。

CART 算法可以用较小的运算代价训练可解释的模型,在模型训练过程中同步完成预报变量的选择,对于现实数据具有良好的包容性,但该算法的预测结果具有较高的方差,是一种不稳定的算法[17]。为改善这一问题,Friedman 等人[18]提出了梯度提升机(gradient boosting machine,GBM)算法,以分类回归树(CART)为基函数,在提升过程中进行模型集成。

在 RE-BET 算法的基础上,尝试应用 GBM 算法估计模型的固定效应部分,为便于描述,记为 RE-BEBT 算法。由于 GBM 模型通常需集成数千乃至数万个基函数,如果结合 Gibbs 抽样过程,计算所需要的时间十分庞大,因此 RE-BEBT 算法借鉴文献[3]中的策略,采用“one - step”的方式来实现计算。即在 RE-BET 算法的第(3)步中,仅完成一次基于 GBM 的固定效应估计,然后通过 Gibbs 抽样进行随机效应的估计,固定效应估计不再进行迭代。

RE-BEBT 算法的随机效应部分计算过程与 RE-BET 算法相同,其固定效应部分的估计采用 GBM 算法,该算法将输入变量与响应变量的累积相关性作为该输入变量的重要性指标,结果见表 3。

从表 3 可以看出,暴露时间、试验区、平均温度和盐度是低合金钢和碳钢在海水中腐蚀率的主要影

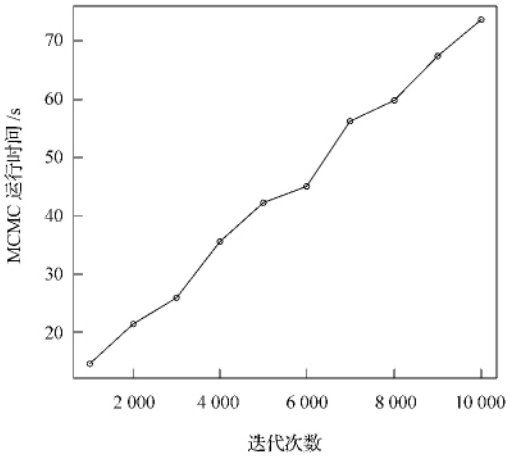

图 4 RE-BET 算法在海水腐蚀数据集训练模型的时间

Fig. 4 The model training time of RE-BET algorithm on seawater corrosion data

表 3 GBM 算法计算的变量重要性系数

Table 3 Importance factors of variables calculated by GBM algorithm

变量名	变量说明	重要性系数
Time	暴露时间	34.44
Region	实验区	33.93
Temperature	平均温度	17.45
Salt	盐度	11.18
Velocity	流速	1.55
Oxygen	溶解氧	1.44
pH	pH	0
Biofouling	生物附着物	0

响因素。这与 RE-BET 算法的结果是一致的。通过计算偏依赖函数，GBM 算法可以探索输入变量与响应变量之间的关系。图 5 报告了暴露时间、试验区等四个海水腐蚀主要影响因素的偏依赖函数。可以看出，RE-BEBT 模型能够发现随暴露时间延长腐蚀率逐渐降低、飞溅区腐蚀率较高等规律[18]。

为了检验 RE-BET 算法的性能，应用不含随机效应项的 CART 和 GBM、MERT、REEM 以及基于半参数贝叶斯方法的混合效应模型共 5 种算法在同样的数据集上训练模型并进行比较，采用的模型性能指标为绝对预测偏差的均值 mape（mean of absolute prediction error），定义为

$$\text{mape} = \frac{1}{m}\sum_i |y_i - y_i^*|。$$

式中 m 为数据集的样本容量，$y_i - y_i^*$ 是实际值和模型预报值的偏差。

作为对比的算法，CART 采用 R 程序（3.1.1 版）的 rpart 包、GBM 采用 gbm 包、REEM 采用 RE-EMtree 包、贝叶斯混合线性模型采用 DP

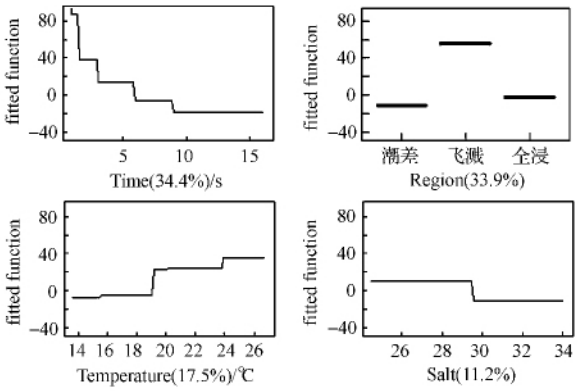

图 5 腐蚀因素的偏依赖函数

Fig. 5 The partial dependence function of corrosion factors

package 包。

海水腐蚀试验数据集中包含了定性变量和定量变量，MERT 算法和本文提出的 RE-BET 等算法采用树形算法拟合模型的固定效应部分，可以方便地处理这类混合数据。半参数贝叶斯方法估计混合线性模型参数时，无法直接处理定性变量，采用设置哑变量的方式进行处理，其 Geweke 指标与 RE-BEBT 的 Geweke 指标一同报告于图 1 中，各马氏链基本收敛。各算法的 mape 结果如表 4 所示。

表 4 不同算法的 mape 指标

Table 4 The mape value of different algorithms

算法	训练集	检验集
CART（不含随机效应）	37.52	54.97
GBM（不含随机效应）	32.03	47.21
混合线性模型（贝叶斯）	57.63	72.04
REEM	18.86	65.10
MERT	17.86	67.56
RE-BET	36.84	50.22
RE-BEBT	29.87	43.73

从表 4 可以看出，MERT 和 REEM 所训练模型在训练集中的 mape 指标低，但在测试集的 mape 指标高，表现出了较弱的泛化能力，这可能是由于算法对于小样本数据处理的不适性造成的。在测试集上，RE-BET 算法的 mape 指标比不含随机效应的 CART 模型减少了 8.6%，比基于半参数贝叶斯方法的混合效应模型减少了 30.3%，比 REEM 模型减少了 22.8%，比 MERT 模型减少了 25.67%。而 RE-BEBT 模型在比较中表现出的性能最好，其 mape 指标比不含随机效应的 GBM 模型减少了 7.3%。

实验表明，本文提出的 RE-BET 和 RE-BEBT 能够准确的描述和预测材料在海水中的腐蚀率，模型具有较好的可解释性；通过与其他算法对比，这二种

算法表现出较好的性能,在处理小样本纵向数据方面,具有一定的优势。

4　结论

提出一种基于树形算法和贝叶斯估计的 RE-BET 算法:采用树形算法估计混合效应模型的固定效应;采用基于 Dirichlet 过程先验的贝叶斯方法估计混合效应模型的随机效应,相对于传统的混合效应模型,在现实数据分析中具有较好的灵活性和适应性。RE-BET 及 RE-BEBT 算法可以合理描述纵向数据的结构,性能优于不含随机效应的 CART 、GBM 等算法,比 REEM、MERT 等算法更适合小样本纵向数据的分析。将 RE-BET 算法应用于自然环境的腐蚀预测及类似领域是可行的。

参　考　文　献

1　Sham N M, Krishnarajah I, Shitan M, *et al*. Time series modelon hand, foot and mouth disease insarawak, malaysia. Asian Pacific Journal of Tropical Disease, 2014; 4(6): 469—472

2　Ntoumanis N. Analysing longitudinal data with multilevelmodeling. The European Health Psychologist, 2014; 16(2): 40—45

3　Kern M L, Hampson S E, Goldberg L R, *et al*. Integrating prospective longitudinal data: modeling personality and health in the *terman* life cycle and Hawaii longitudinal studies. Developmental Psychology, 2014; 50(5): 1390—1406

4　Verbeke G, Fieuws S, Molenberghs G, *et al*. The analysis of multivariate longitudinal data: a review. Statistical Methods in Medical Research, 2014;23(1): 42—59

5　Heagerty P, Liang K Y, Zeger S. Analysis of longitudinal data(2nd Ed). Oxford: Oxford University Press, 2013:67—94

6　Hajjem A. Mixedeffects trees and forests for clustered data. Montreal: HEC Montreal, Department of Management Sciences ,2010

7　Sela R J, Simonoff J S. Re-em trees: a data mining approach for longitudinal and clustered data. Machine Learning, 2012;86(2): 169—207

8　Raudenbush S W, Bryk A S. Hierarchical linear models: applications and data analysis methods. Los Angeles: Sage, 2002:210—231

9　Jula N M. Multilevel model analysis using R. Romanian Statistical Review, 2014;62(2): 55—66

10　Menard S. Longitudinal research. Los Angeles :Sage, 2002:65—78

11　Seltzer M H, Wong W H, Bryk A S. Bayesian analysis in applications of hierarchical models: Lssues and methods. Journal of Educational and Behavioral Statistics, 1996; 21(2): 131—167

12　Kleinman K P, Ibrahim J G. A semiparametric Bayesian approach to the random effects model. Biometrics, 1998;54(3): 921—938

13　Pretorius A L. Bayesian estimation in animal breeding using the dirichlet process prior for correlated random effects. Genet Sel Evol, 2003; 35: 137—158

14　Hannah L A, Blei D M, Powell W B. Dirichlet process mixtures of generalized linear models. The Journal of Machine Learning Research, 2011; 12: 1923—1953

15　Suarez A, Ghosal S. Bayesian clustering of functional data using local features. Bayesian Analysis, 2013; 1(1): 1—15

16　Reisser J, Proietti M, Sazima I, *et al*. Feeding ecology of the green turtle (Chelonia mydas) at rocky reefs in westernsouth atlantic. Marine Biology, 2013;160(12): 3169—3179

17　Breiman L. Bagging predictors. Machine Learning, 1996; 24(2): 123—140

18　Hastie T, Tibshirani R, Friedman J, *et al*. The elements of statistical learming(2nd Ed). New York: Springer, 2009:219—227

19　杨晓明,陈明文. 海水对金属腐蚀因素的分析及预测. 北京科技大学学报, 1999; 21(2): 185—187
Yang Xiaoming, Chen Mingwen. Analysis and forecasting for main factors of corrosion to metals in oceanenvironment. Journal of University of Science and Technology Beijing, 1999; 21(2): 185—187

A RE-BET Algorithm for Longitudinal Data and Its Application

LU Qing, MU Zhi-chun

(School of Automation & Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, P. R. China)

[Abstract] Mixed effects model is a useful methodology for longitudinal data, but the linear structure restricts the model ability to handle the data from real world. a RE-BET algorithm and its deformation named RE-BEBT were proposed. The algorithm used tree-based method to estimate the fixed effects of mixed effects model so that it could select important variables automatically and could discover the relationship between variables. In order to apply model to small sample size data, a Bayesian method based on Dirichlet process prior was used to estimate the random effects of mixed effects model. The RE-BET algorithm was appeied to the corrosion data of low alloy steel and carbon steel in Seawater and compared the result with other algorithms or experimental data, showing that the RE-BET algorithm is feasible and effective.

[Key words] longitudinal data Dirichlet process　tree-based method　corrosion data　Bayesian method