

DOI:10.3876/j.issn.1000-1980.2013.06.015

基于稳健回归分析的实时多播协议性能预测方法

谭国平¹, 霍孝义¹, 谭林风¹, 倪新洋¹, 刘修泉²

(1. 河海大学计算机与信息学院通信与信息系统研究所, 江苏 南京 210098; 2. 广州番禺职业技术学院, 广东 广州 510000)

摘要: 为探讨网络编码块长度、节点移动速度、接收节点个数对基于部分网络编码的实时多播(PNCRM)协议性能的影响,提出采用稳健回归分析构建预测 PNCRM 协议性能回归模型的方法。利用迭代加权最小二乘法求解回归系数,并对该模型进行显著性检验。结果表明,该模型对 PNCRM 协议性能预测的平均相对误差仅为 4.83%,因而具有较准确的预测能力。
关键词: 稳健回归分析;迭代加权最小二乘法;预测模型
中图分类号: TP393 **文献标志码:** A **文章编号:** 1000-1980(2013)06-0554-06

A method for prediction of performance of real-time multicast protocol based on robust regression analysis

TAN Guoping¹, HUO Xiaoyi¹, TAN Linfeng¹, NI Xinyang¹, LIU Xiuquan²

(1. *Research Institute of Telecommunication and Information System Engineering, College of Computer and Information Engineering, Hohai University, Nanjing 210098, China;*
2. *Guangzhou Panyu Polytechnic, Guangzhou 510000, China*)

Abstract: To investigate the influences of the BlockSize, the speed of node movement, and the number of receivers on the performance of partial network coding-based real-time multicast protocol (PNCRM), a model for the prediction of the performance of PNCRM was built based on robust regression analysis. The iteratively reweighted least-squares algorithm was used to solve regression coefficients. A significance test was conducted for the model. The results show that the mean relative error was only 4.83% when using the model for the prediction of the performance of PNCRM, and, thus, the model has strong prediction capability.
Key words: robust regression analysis; iteratively reweighted least-squares algorithm; prediction model

网络编码自 2000 年由 Ahlswede 等^[1]提出以来,便引起国内外学者的广泛关注。网络编码推翻了网络中传输的信息不能叠加、只能存储和转发这一经典结论,通过允许网络内的节点对不同来源的信息进行编码,使得网络能够达到最大流最小割定理所确定的最大理论容量。目前,对于网络编码的研究主要集中在编码策略、网络编码的安全性以及理想模型下网络编码性能分析等方面,并提出了一系列实用的基于网络编码的多播路由机制^[2-3]。

Swapna 等^[4]分析了随机线性网络编码在单跳多播模型下的吞吐量和延时性能。通过分析不同编码长度以及不同网络规模对网络编码性能的影响,揭示了在保证一定性能前提下网络编码长度和网络规模的关系。但是这种理想状况下的理论分析没有也无法充分考虑现实情况下的各种干扰,因而无法反映物理环境下的网络真实性能。Li 等^[5]基于 RTC(real-time calculus)框架,分析了在采用网络编码的网络中单条数据分组传输路径的端到端性能以及网络编码机会对网络性能的影响,并给出理论上计算延时上下限的公式。在得到某条数据分组传输路径相关参数的情况下,利用该公式可获得此路径上数据流端到端传输延时的上下

收稿日期: 2013-03-15
基金项目: 国家自然科学基金(61001068);广州番禺区科技计划项目(2011-Z-01-07);教育部博士点基金(20100094120017)
作者简介: 谭国平(1975—),男,湖南澧县人,副教授,博士,主要从事移动自组网、无线多媒体通信、随机网络优化与控制和网络信息论等研究。E-mail: gptan@hhu.edu.cn

限。与文献[4]类似,这种基于理论的性能分析在一定程度上揭示了理想模型下的网络性能,而对于真实环境下的性能没有现实的指导意义。张建等^[6]利用梯度法对采用网络编码技术的网络延时进行预测,构建了数据块延时值预测模型,并分析了数据块大小对延时值和信道利用率的影响。该模型在网络延时值变化比较平稳时能够进行较好的预测,而在由于某种原因(如干扰)影响而导致延时剧烈波动的情况下,预测值滞后于真实的观测值。类似的,该预测模型也未考虑在物理环境下网络可能出现的状况,因此,只能在网络性能相对平稳的条件下才能有比较理想的预测结果。此外,基于网络编码的多播路由机制很好地利用了网络编码带来的性能增益,优化了数据的投递过程,但却未涉及一般场景下的协议性能分析或只进行了理想模型下的理论分析。网络建设者和管理者无法从这些多播路由机制本身获取其是否能够达到系统设计指标的信息,也无法对机制在不同场景下的性能进行预测。

本文以基于部分网络编码的实时多播协议 PNCRM (partial network coding based real-time multicast protocol)^[2]为例,采用在预测方面应用广泛的多元线性回归分析方法^[7-9]构建网络编码性能预测模型。通过给定的网络场景参数,预测协议的性能;反之,利用该模型,可以根据给定协议性能指标选取合适的包括编码参数和网络节点分布等在内的网络场景参数。

1 预测模型的构建及求解

1.1 构建

文献[2]给出了不同接收节点个数以及不同节点移动速度对 PNCRM 协议性能的影响关系,并分析了其产生原因,但这种基于有限仿真数据的分析只能在一定程度上反映 PNCRM 协议性能变化的趋势。为了研究各变量对 PNCRM 协议性能更精确的影响关系,本文运用多元线性回归分析方法探究其性能(performance)与编码块长度(blockSize)、节点最大移动速度(speed)、接收节点个数(number of receivers)3个变量的关系。

为方便描述,定义如下符号变量。 P ,协议性能,用于评价 PNCRM 协议的有效性、可靠性等指标; B ,编码块长度,大于零的整数,表示 PNCRM 中参与网络编码的数据块个数; S ,最大移动速度,大于或等于零的实数,表示所有节点移动速度的最大值,即各个节点的移动速度取值范围为 $(0, S] \text{ m/s}$; N ,接收节点个数,大于零的整数,表示网络编码区域内接收节点的个数。

根据多元线性回归分析理论^[10],构建 PNCRM 协议性能的通用预测模型,选用多项式函数作为各自变量与性能指标的关系类型:

$$P = b_0 + \sum_{i=1}^{p_1} b_i B^i + \sum_{j=1}^{p_2} b_{p_1+j} S^j + \sum_{k=1}^{p_3} b_{p_1+p_2+k} N^k + \epsilon \tag{1}$$

式中: $b_0, b_1, \dots, b_{p_1+p_2+p_3}$ ——待估计的回归系数; i, j, k ——自变量 B, S, N 的幂; p_1, p_2, p_3 ——各自变量的最高次幂; ϵ ——随机误差项或残差项,常假定它具有如下性质:

$$E(\epsilon) = 0 \quad \text{cov}(\epsilon) = E(\epsilon \epsilon^T) = \sigma^2 I \tag{2}$$

式中: E ——残差的期望; cov ——残差的协方差; σ^2 ——残差的方差; I ——单位矩阵。

如果模型总共测得 n 组数据,为简化计算,常用矩阵表示,则整个模型变为

$$\begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_n \end{bmatrix} = \begin{bmatrix} 1 & B_1 & B_1^2 & \cdots & N_1^{p_3} \\ 1 & B_2 & B_2^2 & \cdots & N_2^{p_3} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & B_n & B_n^2 & \cdots & N_n^{p_3} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p_1+p_2+p_3} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{3}$$

记为

$$P = X\beta + \epsilon \tag{4}$$

式中 β 为回归系数向量。

1.2 求解方法

最小二乘法是常用的求解回归方程系数的方法,其解具有无偏性,并且在一切无偏估计类中具有最小方差。但是,它是建立在随机误差项服从正态分布的假设下,而实际研究中不一定满足这一假设,从而使回归模型的精度大幅度降低。稳健回归法可以有效地解决回归模型假设条件不满足时的问题,采用极大似然估

计(M-估计)进行参数估计,目的是当样本数据存在异常值时尽可能保证回归参数求解的正确性,也就是将异常值识别和处理的思想引入回归参数求解中,使回归分析具有一定的抗差性。

在 M-估计中, β 的求解通过最小化一个关于残差的目标函数 ρ 获得,即

$$\sum_{i=1}^n \rho(\varepsilon_i) = \sum_{i=1}^n \rho(P_i - X_i \beta) \tag{5}$$

如,最小二乘法的目标函数是

$$\sum_{i=1}^n \rho(\varepsilon_i) = \sum_{i=1}^n (P_i - X_i \beta)^2 = \sum_{i=1}^n \varepsilon_i^2$$

令 $\psi=\rho'$ 为目标函数关于 β 的导数,在稳健回归法中称为影响函数。由于目标函数导数为零的解即为目标函数取得极小值的解,则对式(5)求关于 β 的导数,得

$$-\sum_{i=1}^n \psi(\varepsilon_i) X_i = -\sum_{i=1}^n \frac{\psi(\varepsilon_i)}{\varepsilon_i} \varepsilon_i X_i = -\sum_{i=1}^n w_i (P_i - X_i \beta) X_i \tag{6}$$

其中

$$w_i = \frac{\psi(\varepsilon_i)}{\varepsilon_i}$$

式中 w_i 称为权重函数。则回归系数的求解方程可写为

$$\sum_{i=1}^n w_i (P_i - X_i \beta) X_i = 0 \tag{7}$$

最小二乘法的求解方程为

$$\sum_{i=1}^n (P_i - X_i \beta) X_i = 0 \tag{8}$$

与式(8)对比,式(7)只多了一个权重函数,即所谓的加权最小二乘问题。

由以上分析可知,权重由残差求得,残差由回归系数求得,而回归系数的求解又与权重有关,所以式(7)的解用迭代法求得,称为迭代加权最小二乘法(iteratively reweighted least-squares, IRLS),求解步骤如下。

- a. 求初始估计量 $\hat{\beta}^{(0)}$ (通常用最小二乘估计 $\hat{\beta}=(X^T X)^{-1} X^T P$ 代替),得到初始残差 $\varepsilon_{i,0}=P_i-X_i \hat{\beta}^{(0)}$ 。
- b. 根据 $\varepsilon_{i,0}$ 算出初始权重 $w_{i,0}=\frac{\psi(\varepsilon_{i,0})}{\varepsilon_{i,0}}$ 。
- c. 利用加权最小二乘法解得第 1 步 β 的稳健估计 $\hat{\beta}_R^{(1)}$:
$$\hat{\beta}_R^{(1)} = (X^T W_0 X)^{-1} X^T W_0 P \tag{9}$$

式中 W_0 矩阵是一个对角阵,对角元素为 $w_{i,0}$ 。

- d. 用式(9)解得的 $\hat{\beta}_R^{(1)}$ 返回并代替第 1 步的 $\hat{\beta}^{(0)}$,又可得到新的残差 $\varepsilon_{i,1}$,从而得到新的权重 $w_{i,1}$ 。

再返回第 c 步,计算第 2 次迭代的稳健估计 $\hat{\beta}_R^{(2)}$ 。如果用 $\hat{\beta}_R^{(j)}, \hat{\beta}_R^{(j+1)}$ 分别表示第 j 步和第 $j+1$ 步的稳健回归系数向量,则可规定一个迭代收敛的误差标准 e 。当相邻两步之间回归系数差的最大绝对值小于 e 时,迭代结束,即

$$\max(|\hat{\beta}_R^{(j+1)} - \hat{\beta}_R^{(j)}|) < e \tag{10}$$

此时 $\hat{\beta}_R^{(j+1)}$ 即为 β 的解。

选用的目标函数是 bisquare 函数,其形式如下:

$$\rho(\varepsilon_i) = \begin{cases} \frac{k^2}{6} \left\{ 1 - \left[1 - \left(\frac{\varepsilon_i}{k} \right)^2 \right]^3 \right\} & |\varepsilon_i| \leq k \\ \frac{k^2}{6} & |\varepsilon_i| > k \end{cases} \tag{11}$$

对应的权重函数为

$$w(\varepsilon_i) = \begin{cases} \left[1 - \left(\frac{\varepsilon_i}{k} \right)^2 \right]^2 & |\varepsilon_i| \leq k \\ 0 & |\varepsilon_i| > k \end{cases} \tag{12}$$

式中 k 为调和常数,为了能够提供 95% 的有效性,同时排除异常值的干扰, $k=4.685\sigma$ (σ 为残差的标准差,在实际应用中必须用估计量代替)。由于方差、均方差的稳健性都不好,其分布的细微改变就会较大改变方

差、均方差的计算值,而中位数是稳健性较好的量。一种常见的选择方法是

$$\hat{\sigma} = \frac{m}{0.6745} \tag{13}$$

式中: m ——残差 ε_i 的中位数^[11]。

2 仿真数据及模型应用

2.1 仿真数据

根据编码块长度、节点最大移动速度、接收节点个数 3 个变量来构造不同的仿真场景,通过在 NS2 中仿真得到预测模型的原始数据。仿真场景中各参数具体设置如下:仿真平台为 NS2.35;总节点数为 50;发送节点个数为 1;编码块长度分别为 2,4,6,8,10;移动模型为随机路径模型;仿真时间为 82 s;MAC 协议为 IEEE802.11;传输负载为 10 包/s;节点信号覆盖距离为 250 m;节点分布范围为 1 000 m×1 000 m;接收节点数分别为 5,15,25,35,45;节点最大移动速度分别为 4 m/s,8 m/s,12 m/s,16 m/s;停留时间为 0 s;每组参数仿真次数为 100;数据包大小为 512 Bytes;信道容量为 2 Mbps。

为了合理有效地评估 PNCRM 协议性能,选取 2 个指标进行性能评估及预测:

a. 协议总开销(TO, total overhead),反映了协议的有效性,TO 越小,协议的效率越高。

$$\text{协议总开销} = \frac{\text{传输的控制包数} + \text{传输的数据包数}}{\text{实际接收的数据包数}} \tag{14}$$

b. 数据包投递率(PDR, packet delivery ratio),反映协议的可靠性,PDR 越高,协议的可靠性越好。

$$\text{数据包投递率} = \frac{\text{实际接收的数据包数}}{\text{发送节点发送的数据包数} \times \text{接收节点数}} \tag{15}$$

根据仿真场景参数各自变量的排列组合,共得到 100 组仿真场景。每组场景仿真 100 次求平均值得到 PNCRM 协议的 TO, PDR 原始数据。

2.2 模型参数确定

由于不确定式(1)中各自变量(B, S, N)的最高阶次,因此在预测模型求解之前先确定 p_1, p_2, p_3 的值。下面以 BlockSize 与 TO 为例说明确定最高阶次的过程。

首先,根据仿真数据求得 BlockSize 分别取 2,4,6,8,10 时对应的 TO 平均值为 2.093 6, 2.100 2, 2.116 7, 2.126 6, 2.175 4。然后,采用最小二乘法来确定其函数关系式。在创建多个拟合的情况下,可以用残差平方和、调整的判定系数和均方根误差作为对比拟合效果优劣的依据,即残差平方和越小,均方根误差也越小,调整的判定系数则越大,可认为拟合的效果越好。最后,通过比较不同拟合的效果最终得到 TO 与 BlockSize 的函数关系式为

$$P_{TO} = 2.111 + 0.017\,08B + 0.013\,87B^2 + 0.009\,54B^3 \tag{16}$$

式中 BlockSize 的最高次项为 3 次方,即 p_1 的取值为 3。拟合效果如图 1 所示。

同样地,可求得 p_2, p_3 的取值均为 3。这样就得到了 TO 的预测模型:

$$P_{TO} = b_0 + b_1B + b_2B^2 + b_3B^3 + b_4S + b_5S^2 + b_6S^3 + b_7N + b_8N^2 + b_9N^3 + \varepsilon \tag{17}$$

2.3 模型求解及检验

利用 Matlab 对式(14)构建的 TO 预测模型进行求解,得到回归系数的估计值以及模型检验所需的数据,结果如表 1 所示。

为检验所构建模型的有效性,选用两种检验方法:(a)回归模型的效果是否显著,由表 2 中 F 检验值检验;(b)回归模型能否简化,即是否存在某个自变量与预测变量无关或它能被其他自变量代替,因而回归模

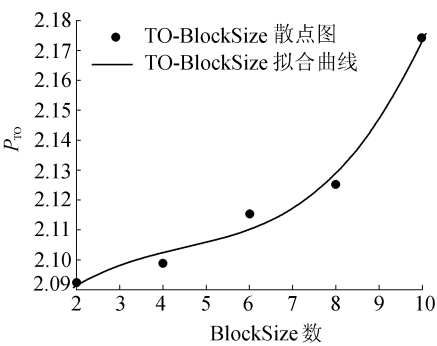


图 1 TO 与 BlockSize 的拟合效果
Fig. 1 Fitting of TO and BlockSize

型中可以删去这个自变量,由表1中 t 检验值检验^[12]。

查 t 分布表和 F 分布表得: $t_{0.025}(90)=1.98667, F_{9,90}(0.01)<2.72$ 。由表1可知, F 检验值为560.1010,远远大于 $F_{9,90}(0.01)$,可以认为此回归模型高度显著。由于自变量参数中有4项的 t 检验值的绝对值小于 $t_{0.025}(90)$,故可认为此4项参数对PNCRM的TO性能没有显著影响,可以删除。

根据上述检验结果删除相应自变量,重新构建TO的预测模型,求解结果见表2。查 t 分布表和 F 分布表知: $t_{0.025}(94)<1.986, F_{5,94}(0.01)<1.95$ 。由于此时 F 检验值为1008.1,远远大于 $F_{5,94}(0.01)$,可以认为此回归模型高度显著。各自变量参数 t 检验值的绝对值均大于 $t_{0.025}(94)$,故可认为各自变量参数对因变量参数TO有显著影响。故最终得到关于TO的预测模型为

$$P_{TO}=6.0087+0.0716B-0.0037B^2-0.3949N+0.0106N^2-0.0001N^3$$

(18)

与求解TO的预测模型过程类似,可求得关于PDR的预测模型为

$$P_{PDR}=0.86662+0.00385B-0.00003B^3-0.00138S+0.00371N-0.00004N^2$$

(19)

3 预测效果及残差分析

为了检验所构建预测模型的预测效果,以TO预测模型为例,通过在NS2中设置不同的场景参数,得到接收节点个数为10,20,30,40时的仿真数据,将观测值与通过预测模型得到的预测值进行对比,选取两个场景做出预测效果图和总体的残差图。

由图2、图3可以看出,观测值曲线和预测值曲线基本吻合,且残差图上点的分布是无规律的,围绕0值上下波动。经计算可得TO预测模型的平均相对误差为4.83%,即所构建的模型能较准确地预测PNCRM的TO性能。

4 结 语

为了研究编码块长度、节点移动速度以及接收节点个数等因素对PNCRM协议性能的影响,采用稳健回归分析的方法,对PNCRM协议构建了TO的预测回归模型,该方法也适用于构建的预测PDR等性能的回归模型。通过迭代加权最小二乘法得出该模型的回归效果高度显著,且预测值与观测值的平均相对误差仅为4.83%,具有很好的预测性。利用所构建的预测模型,可根据给定的性能指标设定相应的网络参数,达到优化系统性能的目的。因PNCRM协议性能不仅与编码块长度、节点移动速度、接收节点个数有关,还受诸如接收节点密度、编码包缓冲区长度等其他因素的影响,为此,在将来的工作中还需探讨这些因素对PNCRM性能的影响,构建更加精确的预测模型。

表1 预测模型结果1

Table 1 Results of prediction model 1

自变量	估计值 b	标准误差	t 检验值
常数项	5.9161	0.2685	22.0375
B	0.2365	0.0938	2.5209
B^2	-0.0349	0.0174	-2.0033
B^3	0.0017	0.0010	1.7829
S	-0.0444	0.0829	-0.5358
S^2	0.0048	0.0092	0.5231
S^3	-0.0001	0.0003	-0.4884
N	-0.3994	0.0125	-32.0437
N^2	0.0108	0.0006	18.4882
N^3	-0.0001	0.0000	-13.2164

注: F 检验值为560.1010。

表2 预测模型结果2

Table 2 Results of prediction model 2

自变量	估计值 b	标准误差	t 检验值
常数项	6.0087	0.0803	74.8001
B	0.0716	0.0207	3.4577
B^2	-0.0037	0.0017	-2.1893
N	-0.3949	0.0108	-36.4662
N^2	0.0106	0.0005	20.9469
N^3	-0.0001	0.0000	-14.9255

注: F 检验值为1008.1。

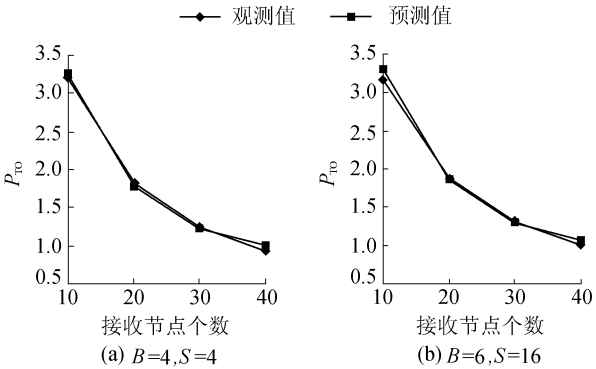


图2 不同场景下的TO预测效果

Fig. 2 Effect of TO prediction in different scenarios

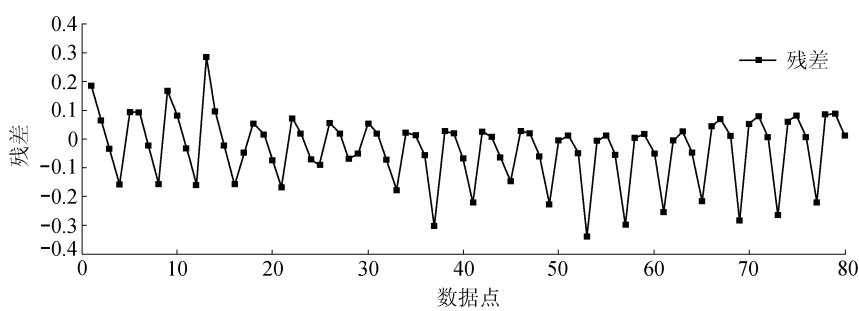


图 3 TO 预测残差
Fig. 3 Residuals plot of TO prediction

参考文献:

[1] AHLISWEDE R,CAI N,LI S Y R,et al. Network information flow[J]. Information Theory,IEEE Trans,2000,46(4):1204-1216.

[2] 谭国平,彭新华,倪新洋,等. 基于部分网络编码的移动自组网实时多播协议研究[J]. 微电子学与计算机,2012,29(9):39-41. (TAN Guoping,PENG Xinhua,NI Xinyang,et al. Research on partial network coding based real-time multicast protocols in mobile ad-hoc networks[J]. Microelectronics & Computer,2012,29(9):39-41. (in Chinese))

[3] 谭国平,倪新洋,季敏,等. 一种基于网络编码的移动自组网实时多播协议[J]. 微电子学与计算机,2011,28(8):12-14. (TAN Guoping,NI Xinyang,JI Min,et al. NCRM:a network coding based on real-time multicast protocol in mobile ad-hoc networks[J]. Microelectronics & Computer,2011,28(8):12-14. (in Chinese))

[4] SWAPNA B T, ERYILMAZ A, SHROFF N B. Throughput-delay analysis of random linear network coding for wireless broadcasting[C]//FRANK K. Toronto:IEEE International Symposium on Network Coding (NetCod),2010:1-6.

[5] LI H Z,BAN D S,YANG W,et al. Per-flow performance analysis in network coding:an approach based on real-time calculus [C]//ZHANG H M. Sanya:International Symposium on Instrumentation & Measurement, Sensor Network and Automation (IMSNA),2012:240-243.

[6] 张健,于纯武,梅峰,等. 网络编码块时延预测与控制[J]. 武汉大学学报:理学版,2012,58(4):366-369. (ZHANG Jian, YU Chunwu,MEI Feng,et al. The prediction and control of block delay under network coding[J]. Journal of Wuhan University: Natural Science Edition,2012,58(4):366-369. (in Chinese))

[7] 邓军,邢震,马砺. 多元回归分析在煤自燃预测中的应用[J]. 西安科技大学学报,2011,31(6):645-648. (DENG Jun,XING Zhen,MA Li. Application of multiple regression analysis in coal spontaneous combustion prediction[J]. Journal of Xi'an University of Science and Technology,2011,31(6):645-648. (in Chinese))

[8] 王萍,李仁杰,徐海洋. 基于多元回归的牡丹花期预测[J]. 农业网络信息,2008(3):139-142. (WANG Ping,LI Renjie,XU Haiyang. The forecasting of peony florescence based on multivariate regression analysis[J]. Agriculture Network Information,2008(3):139-142. (in Chinese))

[9] 施能,王建新. 稳健回归的反复加权最小二乘迭代解法及其应用[J]. 应用气象学报,1992,3(3):353-358. (SHI Neng, WANG Jianxin. A method of iteratively reweighted least square and its application[J]. Quarterly Journal of Applied Meteorology,1992,3(3):353-358. (in Chinese))

[10] JOHNSON R A,WICHERN D W. 实用多元统计分析[M]. 6 版. 陆璇,叶俊,译. 北京:清华大学出版社,2008:280-281.

[11] HUBER P J,RONCHETTI E M. Robust statistics[M]. 2nd. New York:Wiley,2009:46-55.

[12] 朱道元,吴诚鸥,秦伟良. 多元统计分析与 SAS[M]. 南京:东南大学出版社,1999:257-270.