维普资讯 http://www.cqvip.com

一种新的软件可靠性模型参数估计方法

曾劲涛,崔志明,陈建明
(苏州大学 智能信息处理及应用研究所,江苏 苏州 215006)

摘 要:一个软件可靠性模型建立以后,需要对模型的参数进行估计,而参数估计的准确程度将直接影响到模型的预测能力。现在对参数的估计一般采用极大似然估计法。文中提出一种加权最小二乘法,根据各故障数据点对预测贡献值的不同,给与相应的权重。并以一种重要的GO软件可靠性模型为例进行分析,实验表明,该方法获得的参数模型具有更好的预测能力。

关键词:NHPP;GO模型;失效强度;最小二乘法
中图分类号:TP311.5 **文献标识码**:A **文章编号**:1673-629X(2008)07-0209-03

A New Method on Parameters Estimation of a Software Reliabilty Model

ZENG Jin-tao, CUI Zhi-ming, CHEN Jian-ming
(Institute of Intelligent Information Processing and Application, Suzhou University, Suzhou 215006, China)

Abstract: When a software reliability model is built, the follow work will aim to estimate the parameters of the model. Nowadays, the parameters estimation of a software reliability model usually makes use of maximum likelihood estimation. In this paper, a weighted minimum least square method is proposed. In the estimation of model parameters, every data point is assigned a weight corresponding to its contribution to the prediction on the software failure tendency. And GO model which is an important software reliability model is taken as an example for applying this method. Experiments show this method brings GO model a better preditction ability.

Key words: NHPP; GO model; failure intensity; minimum least square

0 引 言

软件可靠性建模是软件可靠性工程的一个重要组成部分,对其研究具有重要意义[1]。目前已建立了数百种软件可靠性模型,但大多数模型只停留于理论研究的层面,与实际应用还有一定的距离,真正广泛应用于工业实践的可靠性模型非常少。GO模型[2]是一种具有代表性的NHPP模型,由于其计算过程简单,并具有较好的预测能力,而被广为采纳。一个模型建立以后,接下来最重要的工作就是利用数据对该模型的参数进行估计,而参数估计的准确程度将直接影响到模型的预测能力。文中以GO模型为例,采用一种新的方法对模型参数进行估计,以提高模型的预测能力。

收稿日期:2007-10-18
基金项目:2005教育部科研重点项目(205059);苏州市创新载体建设项目(SS20524)
作者简介:曾劲涛(1978-),男,江西吉安人,硕士研究生,井冈山大学讲师,主要研究方向为软件测试、软件可靠性;崔志明,教授,博士生导师,主要研究方向为智能化信息处理、计算机网络、数据库。

1 GO模型简介

GO模型是一种用NHPP描述的软件错误查出模型,也可称为软件错误查出的指数类增长模型。设$N(t)$和$m(t)$分别表示在区间$[0,t]$内的累积错误数和期望错误数,则有:

$$m(t) = a(1 - e^{-bt}), a > 0, b > 0 \quad (1)$$

其中a是最终查出的期望错误个数;b是在时刻t每个错误的错误查出率。推导过程见参考文献[3]。对参数a和b的估计一般采用极大似然估计法[4],考虑完全数据,假设观测到N个故障出现的时间$s = (S_1, S_2, \cdots, S_n)$,则$s$的联合概率密度函数是在给定$s$时,关于$a, b$的似然函数。

$$f_{s_1,s_2,\cdots,s_n}(s_1, s_2, \cdots, s_n) = (\prod_{k=1}^{n} abe^{-bs_k})\exp[-a(1 - e^{-bs_n})] \quad (2)$$

由(2)式可得方程组

$$\begin{cases} \frac{n}{a} = 1 - e^{-bs_n} \\ \frac{n}{a} = \sum_{k=1}^{n} s_k + as_n e^{-bs_n} \end{cases}$$

维普资讯 http://www.cqvip.com

解上述方程组,即得 a 和 b 的估计值。

2 加权最小二乘估计法

2.1 加权的思想

GO 模型是一种软件可靠性增长模型。随着软件测试排错的不断进行,软件的缺陷越来越少,软件的可靠性呈不断增长的趋势。另一点,软件的失效强度呈不断下降的趋势,在软件测试的前期,软件的失效强度下降的较快,而在测试后期,失效强度下降的较为平缓,并且逐渐收敛到某个值。可以从两方面来解释这种现象:

(1)在软件测试的早期,软件的缺陷较多并且很容易暴露出来,而在测试的后期,软件的缺陷逐渐变少,并且很难被查出。

(2)在测试的早期,软件故障出现的时间频度很不稳定,而在测试后期表现出较好的稳定性。

因此,可以做一种合理的假设,即随着软件测试的不断进行,在越靠后的时间段出现的故障点越能反映软件剩余缺陷的暴露趋势。因而在参数估计时,越靠后出现的故障点应给与越大的拟合权重,以期获得更好的预测能力。

2.2 加权的方法

利用最小二乘法来估计模型参数,并对每个数据点进行加权。考虑一种最简单的加权方法,函数表达式为 $w_i = c_1 i + c_2 (i = 1,2,\cdots,n)$, 其中 c_1 和 c_2 为常数,i 表示第 i 个数据点,w_i 为第 i 个数据点的权重。加权最小二乘法的表达式如下:

$$s = \sum_{i=1}^{n} w_i (m(t_i) - i)^2 \tag{3}$$

将 $w_i = c_1 i + c_2$ 和 $m(t) = a(1 - e^{-bt})$ 代入(3)式得:

$$s = \sum_{i}^{n} (c_1 i + c_2)(a(1 - e^{-bt_i}) - i)^2 \tag{4}$$

(4)式对 a 和 b 的求偏导并使其值为零,得以下方程组:

$$\begin{cases} \sum_{i=1}^{n} (c_1 i + c_2)(a - 2ae^{-bt_i} + ae^{-2bt_i} - i + ie^{-bt_i}) = 0 \\ \sum_{i=1}^{n} (c_1 i + c_2)(ae^{-bt_i} - ae^{-2bt_i} - ie^{-bt_i}) = 0 \end{cases}$$

直接求解该方程组比较困难,由于 a 和 b 的变化范围不大,可以考虑一种穷举尝试的方法求解。首先,利用极大似然估计法对模型进行参数估计,求出 a 和 b 的值,然后在以 a 和 b 为中心的上下取值范围内,对该方程组进行尝试求解。由于求精确解可能导致计算量过大,并且不能保证一定找到解,因此考虑求近似解,在求解中设定精确度即可。

3 实验分析

选取 DS1,DS2 两个数据集,DS1[3]为美国海军舰队计算机程序中心开发的海军战术数据系统收集的故障数据,DS2[5]为 Musa 数据集,分别如表 1 和表 2。

表 1 DS1

$t(i)$	$y(i)$	$t(i)$	$y(i)$	$t(i)$	$y(i)$	$t(i)$	$y(i)$
0	0	63	9	98	18	337	27
9	1	70	10	104	19	384	28
21	2	71	11	105	20	396	29
32	3	77	12	116	21	405	30
36	4	78	13	149	22	540	31
43	5	87	14	156	23	798	32
45	6	91	15	247	24	814	33
50	7	92	16	249	25	849	34
58	8	95	17	250	26		

表 1 中,$t(i)$ 表示出现失效的时间(天);$y(i)$ 表示出现第几个失效(个)。

表 2 DS2

$t(i)$	$y(i)$	$t(i)$	$y(i)$	$t(i)$	$y(i)$	$t(i)$	$y(i)$	$t(i)$	$y(i)$
5	1	2034	12	4972	23	7638	34	15169	45
78	2	2056	13	4989	24	7821	35	15885	46
219	3	2112	14	5273	25	10283	36	16489	47
710	4	2536	15	5569	26	10387	37	16489	48
715	5	2628	16	5784	27	12565	38	17263	49
720	6	3148	17	5900	28	12850	39	17519	50
748	7	4572	18	6183	29	13021	40	32156	51
886	8	4572	19	6233	30	13021	41	50896	52
1364	9	4664	20	6541	31	13664	42	52422	53
1689	10	4847	21	6820	32	14551	43		
1836	11	4857	22	6960	33	14700	44		

表 2 中,$t(i)$ 表示出现失效的时间(秒);$y(i)$ 表示出现第几个失效(个)。

对于 DS1,取前 23 个数据点拟合参数,后 11 个数据点用于比较两种参数估计方法的效果。用上文提到的极大似然估计法估计参数,得两个参数的估计值 $a_1 = 155.017$, $b_1 = 0.001$;而用加权最小二乘法估计参数,加权因子为 $w_i = 0.001i$, 估计值 $a_2 = 33.0$, $b_2 = 0.007$。在 0 到 849 的时间段内分别绘制实际故障数据曲线和用极大似然估计法、加权最小二乘法估计参数得到的拟合曲线,如图 1 所示。可以发现,在第 23 个故障点(对应时间为 798 天)后,加权最小二乘法得到的参数模型表现了较好的预测能力,明显优于极大似然估计法。对于 DS2,取前 47 个数据点拟合参数,后 6 个点用于预测效果的比较。极大似然法求出参数值分别为 $a_1 = 59.4098, b_1 = 0.000095$;加权最小二乘法求出的参数值分别为 $a_2 = 50.9100, b_2 = 0.000130$。图 2 为对应 DS2 的拟合曲线图,同样可以发现,加权最小二乘法获得了更好的预测效果。

维普资讯 http://www.cqvip.com

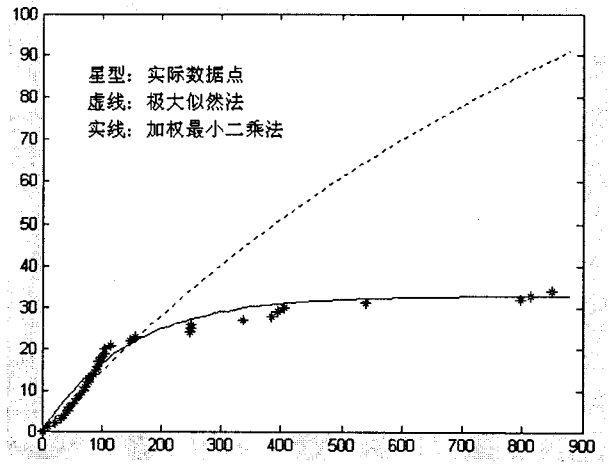

图 1　应用于 DS1 的两种参数估计方法的比较

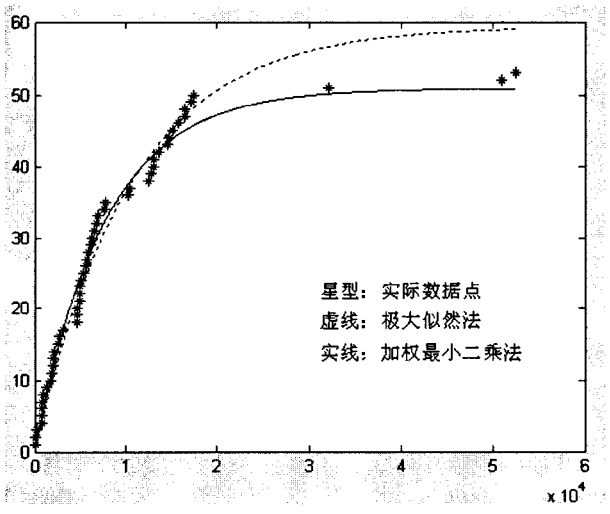

图 2　应用于 DS2 的两种参数估计方法的比较

4　结束语

一般的参数估计方法给与每个数据点相等权重,追求全局的拟合效果,忽略了各个数据点的贡献价值;而文中提出的加权方法考虑了各个数据点对预计未来趋势的贡献价值,价值越高给与的权重也越高,这在理论上是可行的,实验也表明这种方法是有效的,可以取得更好的预测效果。同时,该方法还有一些问题值得进一步研究:

(1)选用了两个失效数据集用于验证这一方法,而该方法是否适合所有失效数据集还有待进一步验证。

(2)文中给出的加权函数为线性函数,其它的函数形式例如指数函数能否取得更好的效果还值得研究。

(3)仅对 NHPP 类可靠性模型进行了研究,其它类型的可靠性模型能否采用这种方法还值得研究。

参考文献:

[1] Muda J D. 软件可靠性工程[M]. 北京:机械工业出版社,2003:17－26.

[2] Goel A L, Okumoto K. Time dependent error detection rate model for software reliability and other performance measures [J]. IEEE Trans. Rel, 1979, 28(3):206－211.

[3] 徐仁佐. 软件可靠性模型及应用[M]. 北京:清华大学出版社,1994:66－78.

[4] 刘荣官. 分组数据下 G－O 模型的参数估计[J]. 上海师范大学学报,2001,30(1):27－31.

[5] 邓　虹. 变点方法在软件可靠性模型中的应用研究[D]. 武汉:武汉大学,2004:47－48.

(上接第 208 页)

4　结束语

将粗糙集理论与神经网络方法相结合,使用了 RoughSet－NN 算法实现了 RoughSet－NN 模型的建立,将粗糙集约简后的规则送神经网络进行学习训练。运用 Matlab7.0 程序设计语言,生成了一个林业决策的 RoughSet－NN 模型系统,本系统可以实现江淮地区林业生长情况预测,预测结果表明,该方法具有很强的可用性。

参考文献:

[1] Pawlak Z. Rough sets－theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic Publishers, 1991.

[2] Jelonek J. Rough set reduction of attributes and their domain for neural networks[J]. Computational Intelligence, 1995, 11(2):339－347.

[3] Peng C. Muti－vailued neual network and the knowledge acquisitiong method by the Rough sets ambiguous recognition problem[C]// Proc of the IEEE international Conference on system, Man and Cybemetics. Beijing: [s. n.], 1996: 736－740.

[4] Slowinski R. Rough Set Approach to Decision Analysis[J]. AI Expert Magazine, 1995, 10:18－25.

[5] 范　明,孟小峰. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2001.

[6] 史忠植. 知识发现[M]. 北京:清华大学出版社,2002.

[7] 邸凯昌. 空间数据挖掘与知识发现[M]. 西安:西安交通大学出版社,2001.

[8] 吴云志. 基于粗糙集与神经网络方法结合的知识发现应用研究[D]. 合肥:合肥工业大学,2006.

[9] 朱　嫱,李章华. 排序和分级同步的综合评价模型[J]. 清华大学学报:自然科学版,1999,39(6):116－120.

[10] 赵玉杰. 基于 Matlab 6.x 的 BP 人工神经网络的土壤环境质量评价方法研究[J]. 农业环境科学学报,2006,25(1):186－189.