第 10 卷　第 30 期　2010 年 10 月
1671—1815(2010)30-7483-05

科　学　技　术　与　工　程
Science Technology and Engineering

Vol. 10　No. 30　Oct. 2010

变点监测和控制图在数据业务监控中的应用

纪春芳　卓新建

（北京邮电大学理学院，北京 100876）

摘　要　为实现对数据业务支撑系统的有效监控，通过剖析数据业务特点，提出了用控制图和变点监测发现数据业务中存在的异常。变点监测在处理三种类型的时间序列时存在不足，通过增加控制参数——变化率，对变点检测算法进行了改进，改进后的方法适应性更高，并且误报率大大降低。最后比较了控制图和变点检测方法的优缺点及各自的应用场景。

关键词　控制图　　变点检测　　数据业务监控

中图法分类号　O231.3；　　**文献标志码**　A

3G 时代，随着数据业务的大量发展，数据业务占运营商总收入的比例将逐渐升高，甚至超过 30%。传统的网络设备级监控已经远远不能满足需求。一方面，由于各种原因，可能导致某类型的错单增加；另一方面，SP 也利用各种业务平台和计费系统信息同步的漏洞来进行恶意的自消费，采用短时间内大量业务定购欠费后弃卡，从中获利。必须有效地实现对业务支撑系统的监控，才能解决上述问题。

各种业务数据的产生是一个连续不断的过程，系统产生的大量数据流中隐含了用户行为，网络性能，市场动向等方面的模式和特征。通过对数据业务的分析，剖析其存在的各种业务异常，然后针对业务特点，提出控制图[1,2]和变点检测[3]的方法。

1　控制图

控制图，如图 1 所示，是一种有控制界限的图，用来判断过程是否处于受控状态。

目前已发展了多种控制图法[1,2]，即控制线的计算方式多种多样，有些控制线是通过极差[4]计算得来的，有些控制线是通过标准差计算得到的。其中最经典，应用最为广泛的是休哈特控制图。

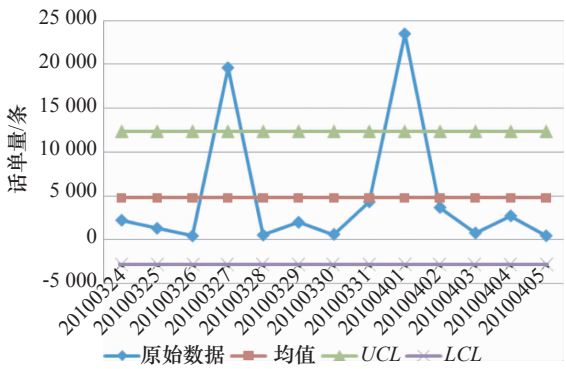

图 1　控制图

可通过当前的数据是否处于 *LCL*（控制下限）与 *UCL*（控制上限）之间来判断数据业务是否出现了异常。

对于一些处于成熟阶段的业务，其业务量相对比较稳定，可以直接用控制图法来发现异常。

该方法的控制线可以随新增数据的改变而改变。控制图在发现单点（也称孤立点）异常方面有较优的性能，简单且易实现。

2　变点定义及变点检测常用方法

变点问题是变点问题自 20 世纪 70 年代以来

2010 年 8 月 23 日收到

第一作者简介：纪春芳（1985—），女，山东人，北京邮电大学硕士生，研究方向：运筹学及其在通信中的应用。E-mail：xiaoji0106@126.com。

一直是统计中的一个热门话题，它广泛应用于工业质量控制、经济、金融、地震灾害预测等领域中。

设 $X_1, X_2, \cdots, X_n$ 是随机变量，一般情况下，假定 $X_1, X_2, \cdots, X_n$ 相互独立，在数理统计上对变点的定义如下：

如果 $X_1, X_2, \cdots, X_r$ 同分布于 F，$X_{r+1}, X_{r+2}, \cdots, X_n$ 同分布于 G，其中 $1 \leqslant r \leqslant n-1$ 是未知的正整数，称 r 为突变的变点（Abrupt Change—Point）。

关于独立随机变量序列中的变点问题，许多学者都进行了研究。研究最多的两类变点为均值变点和方差变点。

目前关于均值变点的检测大致有两种方法：局部比较方法以及 CUSUM 方法。

文献[6]中介绍的 CUSUM 方法是典型方法之一，但是只对恰有一个变点的检测最有效。

文献[7]中提出了关于刻度参数变点的非参数估计推断，该方法基于的思想为局部比较法。该方法的优点是，对总体分布的矩没有任何要求，但是缺点也很明显，就是该方法只对至多有一个刻度参数变点模型中的一维刻度参数进行估计和检验。

3 变点检测在数据业务监控中的应用

3.1 变点检测法的引入

有时数据业务的波动尽管在控制图的 *UCL* 和 *LCL* 控制范围内，但是波动的数据仍然隐含了某种变化的信息，我们把这种有变化的数据点称为“变点”，用变点检测的方法就可以发现上述变化。对表 1 中的的原始数据，用休哈特控制图法没有发现异常，如图 2 所示，但是仔细看图不难发现，尽管 2010-3-31 日之后的所有数据都落在控制线内，但是从 2010-4-1 日开始，数据量较前些日期明显上升。

在检测此种类型的数据异常方面，控制图有些无能为力。变点检测可以发现这一重要的变化。

表 1 数据和控制界限

日期	原始数据	均值	标准差	*UCL*	*LCL*
20100324	310	828	547.67	1 358.75	257.4
20100325	392	828	547.67	1 358.75	257.4
20100326	331	828	547.67	1 358.75	257.4
20100327	477	828	547.67	1 358.75	257.4
20100328	402	828	547.67	1 358.75	257.4
20100329	340	828	547.67	1 358.75	257.4
20100330	320	828	547.67	1 358.75	257.4
20100331	640	828	547.67	1 358.75	257.4
20100401	1270	828	547.67	1 358.75	257.4
20100402	1272	828	547.67	1 358.75	257.4
20100403	1354	828	547.67	1 358.75	257.4
20100404	1192	828	547.67	1 358.75	257.4
20100405	1356	828	547.67	1 358.75	257.4
20100406	1183	828	547.67	1 358.75	257.4
20100407	1 282	828	547.67	1 358.75	257.4

注：*UCL* = 均值 + 1.2 标准差；*LCL* = 均值 − 1.2 标准差

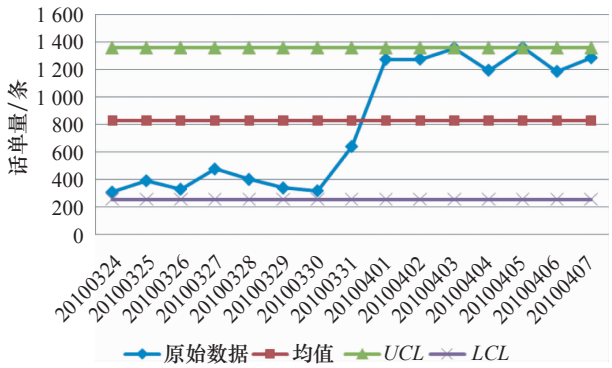

图 2 休哈特控制图

3.2 本文使用的变点检测方法

上节提到的变点检测方法对时间序列有一定的要求，或者是只对存在一个变点的时间序列有效。

文献[3]用中用另外一种 CUSUM 法来检测是否存在变点，该方法的适用范围更广，并且对存在多个变点的时间序列同样有效。

检测算法如下：

（1）读入时间序列 $x = (x_1, x_2, \cdots, x_n)$；

（2）求均值 $\bar{x}$；

（3）求 s_i 其中 $s_i = s_{i-1} + (x_i - \bar{x})$；

(4) $s_{diff} = s_i(\max) - s_i(\min)$;

(5) 把 x 重排列,取排列中的若干个,对每一个排列,执行步骤 3 和 4;

if(当前的 $s_{diff} < s_{diff}$) $count$ ++;

(6) $confidence\_level = count/n$;

(7) if ($confidence\_level > 90\%$),有变点,转(8)否则无变点,结束;

(8) 计算变点 m, 其中 m 满足 $|s_m| = \max|s_i|$ (这里的 s 是针对原序列的);

(9) 返回 m。

说明:算法中 $confidence\_level$ 为置信水平,该值越大,说明有变点的可信度越大;90% 是控制参数,可根据需要自行设置其大小,该参数越小,意味着算法的灵敏度越高。

对表 1 中的数据,在 R[5] 中写程序,用变点检测方法检测时,发现 9 为变点,可信度为 99.83%,即 2010.4.1 起数据的均值开始发送显著变化。

用上面的方法一旦检测到一个变点,就可以把原始数据分成两部分,对每一部分重复上面的方法,每次检测到变点后,就对数据进行一次分割,直到没有变点存在。这就能够发现时间序列中存在的多个变点。

3.3　变点检测方法缺陷之一

在实际应用[3]中的变点检测算法的过程中,发现有些重要的变化不会被发现。对于下面这组数据 $x = (1,2,3,4,101,102,103,104)$。在 R[5] 里运行结果显示,x 全排列的个数共有 40 319 个,s_{diff} 值小于原始序列的 s_{diff} 的序列共有 35 712 个,置信水平 $confidence\_level$ = 88.57%。显然如果参数 $confidence\_level$ 设为 90%,将不会发现数据的变化。

通过降低参数 $confidence\_level$ 的控制值,可以检测到更多的数据异常,但同时一些我们不关心的很细小的变化也会被检测到。从众多异常中挑选我们真正关心的较大的数据变化,费时费力。

3.4　变点检测方法缺陷之二

有时,我们对较细微的并且已经恢复正常的变化(如图 3),并不关心,也不愿再追究,那么不希望产生此类告警。否则,会大大提高日常工作量。

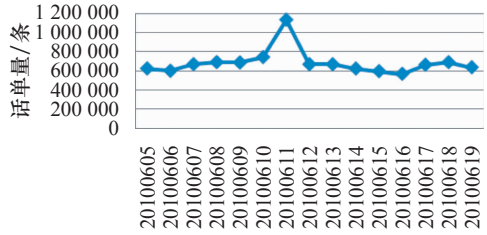

图 3　含有变点的时间序列

3.5　变点检测方法缺陷之三

图 4 中,数据的变化我们也不关心,因为近期已经恢复,且追查前两天的变化意义不大,于是希望过滤此类告警。

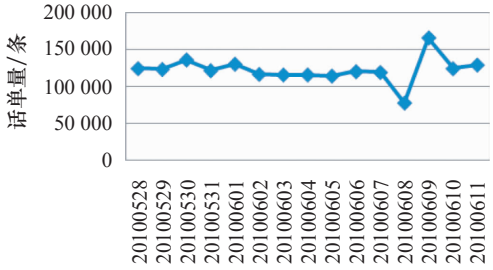

图 4　含有变点的时间序列

3.6　改进的变点检测方法

针对实际应用中算法存在的不足,增加参数 $change\_rate$,记录变点前后数据的变化率,以及最后一个数据较变点前一个数据的变化率。改进后整个算法如下:

(1) 读入时间序列 $x = (x_1, x_2, \cdots, x_n)$;

(2) 求均值 $\bar{x}$;

(3) 求 s_i, 其中 $s_i = s_{i-1} + (x_i - \bar{x})$;

(4) $s_{diff} = s_i(\max) - s_i(\min)$;

(5) 把 x 重排列,取排列中的若干个,对每一个排列,执行步骤 3 和 4;

if(当前的 $s_{diff} < s_{diff}$)

$count$ ++;($count$ 初始值为 0);

(6) $rate = count/n$;

(7) 计算 m, 其中 m 满足 $|s_m| = \max|s_i|$;

(8) 返回 m;

(9) $change\_rate1 = |(x_{m+1} - x_{m-1})/x_{m-1}|$;

$change\_rate2 = |(x_{m+2} - x_{m-2})/x_{m-2}|$;

$change\_rate3 = |(x_n - x_{m-2})/x_{m-2}|$。

(10) *confidence _ level*, *change _ rate*1, *change _ rate*2, *change_rate*3 四个控制参数结合,共同为发现变点服务。

3.7 实例分析与应用

抽取个人点对点短信业务的一组数据为 x = (423 306,445 448,457 406,513 551,458 296,463 453,421 082,522 801,714 927,814 304,963 035,575 062,498 308,489 325,421 098),改进前后变点检测方法的结果如下表 2。

表 2 改进前后变点检测结果比较

	参数	参数值	参数的控制值	变点
变点检测	①*confidence_level*	0.985	大于 95%	8
改进的变点检测	①*confidence_level*	0.985	①>0.9&& ②>0.5&&	无
	②*change_rate*1	0.558	③>0.3&& ④>0.3 OR	
	③*change_rate*2	1.287	①>0.98&& ②>0.3&&	
	④*change_rate*3	0.000	③>0.3&& ④>0.3	

从实验结果发现,通过适当的设置参数 *confidence_level*, *change_rate*1, *change_rate*2, *change_rate*3 的控制值,可使误报率大大降低。而从参数的控制值及控制方法,不难发现,不会遗漏那些非恢复性的数据变化。

4 控制图和变点检测方法的比较

控制图在发现孤立点方面有较优的性能,简单且易实现。变点检测主要用于发现趋势异常。很多情况下,二者可以结合起来使用。

变点检测用来发现被控制图忽略的较细微的变化,可以更好的描述业务数据的变化趋势,并且通过置信等级来描述趋势变化的可信度。改进后的变点分析过程相当灵活,无论是特征数据(服从一定分布规律的数据)还是没有任何分布规律的数据,无论待分析数据是否含孤立点,也无论待分析数据的数据量多少,都可以用变点检测的方法找到变点。

需注意的是,变点检测并不能代替控制图法,它与控制图法互相补充,更好地为发现异常数据服务。

5 结论

数据业务监控的思路是:以较小的代价实现对业务数据异动的有效监控,主动暴露可能存在的问题;立足现有的大量业务数据,从中抽取隐含特征和相互关系,以数据的分布、形态作分析和判断;基于测量的业务数据建模,对实际业务数据进行收集和分析,检测、识别并且量化其中的一些显著特征。控制图和变点检测在数据业务监控和分析方面发挥着重要作用。

参 考 文 献

1 王乙红,徐 艳,杨 波,等. 休哈特控制图原理在医疗安全信息报告中的应用探讨. 中国卫生资源,2009;9(3):41—43

2 陈 健,王 军. 统计过程控制在产品质量管理中的应用研究. 淮阴工学院学报,2008; 17(6):56—58

3 Taylor W A. Change-pointaAnalysis:a powerful new tool for detecting changes. Baxter Healthcare Corporation, Round Lake, IL 60073

4 盛 骤,谢式千,潘承毅. 概率论与数理统计(第四版). 北京:高等教育出版社,2008

5 汤银才. R 语言与统计分析. 北京:高等教育出版社,2008

6 王 玲,同小娟. 基于变点分析的地形起伏度研究. 地理与地理信息科学,2007;11:23(6):65—67

7 缪柏其,魏登云. 关于刻度参数变点的非参数估计. 中国科学技术大学学报,1994;24(3):263—270

Change Point and Control Charts in Application of Data Monitoring

JI Chun-fang, ZHUO Xin-jian
(College of Science, Beijing University of Post and Telecommunications, Beijing 100876, P. R. China)

[**Abstract**] For the purpose of monitoring the business operations support system, the control charts and change point detection are applied to detect abnormal data. But practice has shown that change point detection existed shortness in dealing with some types of time series. Through increasing three parameters, the original algorithm can be improved. Practice proves that after the improvement, misinformation can be greatly reduced and nearly all the abnormal data can be detected at the same time.

[**Key words**] control charts　change point detection　data exceptional analysis

(上接第 7479 页)

2 Jhon M V Jr, Van Ryzin J. Convergence rates in empirical Bayes two-action problems II. Continuous Case Ann Math Statist, 1973; 43: 934—947
3 洪　坚,韦来生. 指数分布定数截尾样本下经验 Bayes 双侧检验问题. 中国科学技术大学学报,2006;36(12):1289—1293
4 康会光,许　勇. 非对称损失下单边截断分布族参数的经验 Bayes 检验问题. 工程数学学报,2005;22(3):493—498
5 陈家清,刘次华. 线性指数分布族参数的经验 Bayes 检验问题. 系统科学与数学,2008;28(5):616—626
6 Soliman A A. Estimation of parameter of life from progressively censored data using Burr XII model. IEEE Transions on Reliability, 2005; 54 (1): 34—42
7 王婷婷,师义民. Burr-XII 部件可靠性指标的贝叶斯估计. 系统工程,2009;27(5):113—115
8 韦程东,陈志强,韦　师,等. 两参数 Burr XII 分布的经验 Bayes 检验问题. 工程数学学报,2010;27(2):331—341
9 刘　英,师义民,王婷婷. 含有屏蔽数据的串联系统 Burr Ⅻ部件可靠性指标的 Bayes 估计. 系统工程理论与实践,2010;4:689—694

Empirical Bayes Test for the Shape Parameter of Burr Type XII Distribution under Weighted Linear Loss Function

XING Jian-ping
(Department of Business, Hunan Radio and Television University, Changsha 410004, P. R. China)

[**Abstract**] By using the kernel-type density estimation and empirical distribution function in the case of independent and indentically distributed random variables, the empirical Bayes one-sided test rules: for the parameter of Burr Type XII distribution are constructed under weighted linear loss function , and the asymptotically optimal property is obtained . It is shown that the convergence rates of the proposed EB test rules can arbitrarily close to $O(n^{-1/2})$ under suitable conditions.

[**Key words**] empirical Bayes test　asymptotic optimality　convergence rates　weighted linear loss function