

时间序列异常点及突变点的检测算法

苏卫星^{1,2} 朱云龙¹ 刘芳³ 胡琨元¹

¹(中国科学院沈阳自动化研究所 沈阳 110016)

²(中国科学院大学 北京 100049)

³(华晨汽车工程研究院 沈阳 110027)

(suweixing@sia.cn)

Outliers and Change-Points Detection Algorithm for Time Series

Su Weixing^{1,2}, Zhu Yunlong¹, Liu Fang³, and Hu Kunyuan¹

¹(Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016)

²(University of Chinese Academy of Sciences, Beijing 100049)

³(Brilliance Automobile Engineering Research Institute, Shenyang 110027)

Abstract Because the conventional change-points detection method exists the shortages on time delay and inapplicability for the time series mingled with outliers in the practical applications, an outlier and change-point detection algorithm for time series, which is based on the wavelet transform of the efficient score vector, is proposed in this paper. The algorithm introduces the efficient score vector to solve the problem of the conventional detection method that statistics often increase infinitely with the number of data added during the process of detection, and proposes a strategy of analyzing the statistics by using wavelet in order to avoid the serious time delay. In order to distinguish the outlier and change-point during the detection process, we propose a detecting framework based on the relationship between Lipschitz exponent and the wavelet coefficients, by which both outlier and change-point can be detected out meanwhile. The advantage of this method is that the detection effect is not subject to the influence of the outlier. It means that the method can deal with the time series containing both outliers and change-points under actual operating conditions and it is more suitable for the real application. Eventually, the effectiveness and practicality of the proposed detection method have been proved through simulation results.

Key words outlier; change-point; wavelet transform; Lipschitz exponent; time series

摘要 针对传统突变点检测算法具有大延时的问题以及实际数据中同时含有突变点、异常点的实际情况,提出一种基于小波变换有效分数向量的异常点、突变点检测算法.该方法通过引入有效分数向量作为检测统计量,有效避免了传统检测统计量随着数据增多而无限增大的缺点;提出利用小波分析统计量的办法,有效地克服了传统突变点检测算法中存在大延时的缺陷;利用李氏指数及小波变换的关系,实现了在一个检测框架内同时在线检测异常点以及突变点,使得该检测算法更符合突变点及异常点同时存在的实际情况.仿真实验和性能比较结果证明了提出的异常点、突变点检测算法具有一定的有效性和实用性.

关键词 异常点;突变点;小波变换;Lipschitz指数;时间序列

中图法分类号 TP311.11

收稿日期:2012-07-02;修回日期:2013-05-02

基金项目:国家科技支撑计划基金项目(2012BAF10B11,2014BAF07B01);辽宁省自然科学基金项目(201202226)

时间序列在过程工业、金融业以及通信业等各个领域普遍存在,因此目前针对时间序列的分析研究受到很多学者的广泛关注.在众多研究课题中,时间序列异常值检测因其直接关系时间序列的质量而成为所有研究中的基础,因此具有重要的科研价值.另外,在过程控制^[1]或网络监控^[2-3]领域中,一般采用监控数据变化趋势的方式,达到监控系统或网络运行状况的目的,即时间序列突变点(change point)检测.由于突变点与异常点有一定的相似之处,极容易在突变点发生的短时间内被误认为是异常点.因此对于时间序列而言,在短时间内检测并区分异常点和突变点是非常必要且重要的.但目前对于异常点以及突变点检测的研究却几乎均是分别独立进行的,而在实际数据中这两种数据却是同时存在的,因此在算法研究中将其统一考虑是必要的.

异常点检测是一个相对成熟的研究领域,到目前为止已经形成了诸多较为成熟且实用的方法,例如最早的基于统计的检测算法^[4]、基于距离的检测算法^[5]、基于密度的检测算法^[6]以及后来发展的神经网络的方法^[7]、支持向量的方法^[8]以及聚类分析的方法^[9]等.小波分析由于能够在时域和频域都具有表征信号局部特征的能力,因此也被用于进行时间序列分析^[10].最有代表性的是 Mallat 等人在 1992 年提出的基于小波变换模极大值原理的时间序列异常点检测方法^[11].该方法适用于从平稳信号中提取非稳态变化,而对于非稳态过程信号,则无法区分信号的非稳态变化(即突变点)和异常变化(即异常点).

在突变点检测方面,早期方法主要基于统计的思想. Gustafsson 在 1996 年提出边缘似然率检验的方法^[12],该方法为一种批处理方法,将全部数据从中间不同的地方进行分割,通过寻找分割后前后两部分似然率最大值的方式,确定突变点的位置.该方法计算量大,不适合时间序列的突变点检测问题. Guralnik 等人在 1999 年采用了同 Gustafsson 类似的确定突变点位置的思想,提出一种迭代算法^[13],并将该算法扩展为能够适应增长型时间序列的突变点检测.但该方法的计算复杂度是时间相关的,即在长时间没有出现突变点的情况下该方法的计算量将变得异常庞大,因此不适合实时检测的应用. Sharifzadeh 等人在 2005 年提出一种基于小波足迹法的突变点检测方法^[14].该方法虽然能够适用于大规模数据集,且具有很好的检测精度和性能,但也是一种批处理方法. Alarcon-aquino 等人在 2009 年提出两窗口结构检测方法^[15],该方法通过比较参考窗

口与滑动窗口内数据所服从分布的方差是否相等来检测突变点,实现了在线检测的可能.但是通过分析,该方法存在 2 个缺点:一是随着参考窗口内的数据不断增加,其统计量也随之增加,意味着若一直没有突变点出现,则参考窗口内的统计量将无限制增加,因此,该算法中的统计量不是一个理想的统计量;其二,检测方法的准确性由滑动窗口的长度决定,即窗口越长检测越准确,检测延迟也越大.而检测延时是我们不希望见到的.

针对目前突变点检测算法中的两点不足以及同时检测异常点和突变点的必要性,本文提出一种能够在线运行、及时检测异常点和突变点的方法——基于小波分析有效分数向量的异常点、突变点检测算法.该方法克服了传统突变点检测大延时以及检测统计量随数据增大而无限增大不足,将异常点检测和突变点检测统一起来,实现在线检测异常点以及突变点的可能,完全适合时间序列数据量大、实时性强以及要求在线检测的要求.通过仿真实验证明,本文提出的检测算法具有一定的有效性和实用性.

1 传统基于有效分数向量的突变点检测算法

传统的基于有效分数向量(efficient score vector, ESV)的突变点检测算法是 Gombay 等人提出的^[16],属于一种假设检验的方法.该方法通过检测时间序列所服从分布中的某些参数的变化来判断该序列的变化,进而找出突变点的位置.

1.1 有效分数向量

设 x_1, x_2, \dots 是一组独立同分布待检测数据,其密度函数为 $f(x; \theta, \eta)$, 其中 $\theta \in \Omega_1 \subset \mathbb{R}^d, d \geq 1$ 表示算法中“感兴趣”的参数变量; $\eta \in \Omega_2 \subset \mathbb{R}^p, p \geq 0$ 表示算法中“不感兴趣”的参数变量,又叫冗余参数; $\Omega = \Omega_1 \times \Omega_2$ 为参数变量所在的空间.据此给出传统突变点检测算法中的两种假设^[16],如式(1)所示:

$$\begin{aligned} H_0: & \theta = \theta_0, \text{ 对于所有观测值 } \eta \text{ 未知;} \\ H_A: & \text{对于 } x_1, \dots, x_{\tau-1}, \text{ 满足 } f(x; \theta_0, \eta), \eta \text{ 未知;} \\ & \text{对于 } x_{\tau}, x_{\tau+1}, \dots, \text{ 满足 } f(x; \theta_A, \eta), \eta, \theta_A \text{ 未知;} \end{aligned} \quad (1)$$

其中, θ_0 为突变点发生前序列服从分布 f 中的参数, θ_A 为突变点发生后序列服从分布 f 中的参数, τ 为突变点所在时刻.从假设可以看出冗余参数为未知参数.

所以 ESV 表达式定义为^[17]

$$V_k(\theta, \eta) = \sum_{i=1}^k \nabla_{\xi} \log f(x_i; \theta, \eta), \quad \xi = (\theta, \eta), \quad (2)$$

其中冗余参数 $\boldsymbol{\eta}$ 的估计值 $\hat{\boldsymbol{\eta}}_k$ 由式(3)给出:

$$\sum_{i=1}^k \nabla_{\boldsymbol{\eta}} \log f(x_i; \boldsymbol{\theta}_0, \boldsymbol{\eta}) = 0.$$

(3)

依据式(2)、式(3),ESV 表达式变为

$$V_k(\boldsymbol{\theta}_0, \hat{\boldsymbol{\eta}}_k) = \sum_{i=1}^k \nabla_{\xi} \log f(x_i; \boldsymbol{\theta}_0, \hat{\boldsymbol{\eta}}_k) = \sum_{i=1}^k \nabla_{\boldsymbol{\theta}} \log f(x_i; \boldsymbol{\theta}_0, \hat{\boldsymbol{\eta}}_k).$$

(4)

从式(4)中的 ESV 表达式 $\{V_k, k > 1\}$ 可以看出,当假设 H_0 为真时,ESV 值序列 $V_k, k = 1, 2 \cdots$, 具有零均值,从而避免了 Alarcon-aquino 的两窗口检测算法中,当没有突变点发生时,统计量会无限制增加的缺点;而当假设 H_A 为真时,ESV 值序列 $V_k, k = 1, 2 \cdots$ 的值将会逐渐变大,而且其变化程度随着突变点后数据的增多而呈线性比例地增大。

1.2 布朗过程

为了方便讨论,这里假设 f 属于幂指数函数族,因此可以将其转换成:

$$\log f(x; \boldsymbol{\theta}, \boldsymbol{\eta}) = T_1 \boldsymbol{\theta}' + T_2(x) \boldsymbol{\eta}' + S(x) - A(\boldsymbol{\theta}, \boldsymbol{\eta}),$$

(5)

其中 T_1, T_2, S, A 均为已知函数。

在介绍布朗运动前,首先给出以下 3 个条件:

- 1) 向量 $\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}, \boldsymbol{\eta}), \nabla_{\boldsymbol{\eta}} A(\boldsymbol{\theta}, \boldsymbol{\eta})$ 存在,且有唯一的逆;
- 2) 矩阵 $\nabla_{\xi}^2 A(\boldsymbol{\theta}, \boldsymbol{\eta})$ 存在且正定,其中各个变量的 Lipschitz 指数大于零;
- 3) 对于 $T(x)$ 中各个向量组分,有: $E(T_i)^{2+\delta} < \infty, \delta > 0$ 。

引理 1^[18]. 在 $(\boldsymbol{\theta}_0, \boldsymbol{\eta})$ 的一个邻域内,如果条件 1)~3) 在假设 H_0 下满足,则存在一个过程 $\boldsymbol{W}(k)$ 满足:

$$\|\boldsymbol{W}_k - \boldsymbol{W}(k)\| = o(t^{1/(2+\delta)}),$$

(6)

该过程 $\boldsymbol{W}(k) = (\boldsymbol{W}^{(1)}(k), \cdots, \boldsymbol{W}^{(d)}(k))$, 而 $\boldsymbol{W}^{(i)}, i = 1, \cdots, d$ 为独立 Wiener 过程,又叫 Brownian(布朗)过程。

式(6)中 \boldsymbol{W}_k 为一个包含 ESV 的表达式:

$$\boldsymbol{W}_k = \Gamma^{-1/2}(\boldsymbol{\theta}_0, \boldsymbol{\eta}) V_k(\boldsymbol{\theta}_0, \hat{\boldsymbol{\eta}}_k),$$

(7)

其中 $\Gamma(\boldsymbol{\theta}, \boldsymbol{\eta})$ 的表达式为

$$\Gamma(\boldsymbol{\theta}, \boldsymbol{\eta}) = \boldsymbol{I}_{11} - \boldsymbol{I}_{12} \boldsymbol{I}_{22}^{-1} \boldsymbol{I}_{21},$$

(8)

而 $\boldsymbol{I} = \begin{pmatrix} \boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_{21} & \boldsymbol{I}_{22} \end{pmatrix}$ 为信息矩阵,其计算式为

$$\boldsymbol{I} = \left(\frac{\partial^2 A(b)}{\partial \xi_i \partial \xi_j} \right)_{(d+p) \times (d+p)}.$$

(9)

该引理说明在假设 H_0 为真的情况下,统计量

\boldsymbol{W}_k 近似于布朗过程 $\boldsymbol{W}(k)$. 因此当序列发生变化后,此近似将不存在. 传统的基于 ESV 突变点检测算法即以此原理检测突变点。

1.3 具体检测算法

传统的算法中考虑了两种情况:1) 已知全部数据的情况下,对其进行检测;2) 针对无穷多数据的情况进行检测. 针对以上两种情况简单介绍算法如下^[19]:

1.3.1 有限数据下的检测

设总数据量为 n , 首先给出两个函数如下:

$$a(x) = (2 \log x)^{1/2},$$
$$b(x) = 2 \log x + \frac{1}{2} \log \log x - \frac{1}{2} \log \pi.$$

(10)

引理 2. 在引理 1 的条件下有:

$$\lim_{n \rightarrow \infty} P(a(\log n) \max_{1 \leq k \leq n} k^{-1/2} \boldsymbol{W}_k \leq t + b(\log n)) = \exp(\exp(t)), \quad -\infty < T < \infty,$$

(11)

$$\lim_{n \rightarrow \infty} P(a(\log n) \max_{1 \leq k \leq n} k^{-1/2} |\boldsymbol{W}_k| \leq t + b(\log n)) = \exp(-2 \exp(-t)), \quad -\infty < t < \infty.$$

(12)

依据引理 2 中的式(11)以及式(12),得到单边检测原则以及双边检测原则。

1) 单边检测. 随着数据个数 k 的不断增加,如果出现

$$k^{-1/2} \boldsymbol{W}_k \geq \boldsymbol{C}_1(\boldsymbol{\alpha}),$$

(13)

则说明序列在 k 附近出现突变点,其中 $\boldsymbol{\alpha} = [\alpha^1, \cdots, \alpha^i, \cdots, \alpha^d]$ 为置信度向量,一般取 0.05. 而 $\boldsymbol{C}_1(\boldsymbol{\alpha}) = [C_1^{(1)}(\alpha^1), \cdots, C_1^{(i)}(\alpha^i), \cdots, C_1^{(d)}(\alpha^d)]$ 中的分量可以根据式(11)计算得出:

$$C_1^{(i)}(\alpha^i) = (2 \log \log n)^{-\frac{1}{2}} \left[-\log(-\log(1 - \alpha^i)) + 2 \log \log n + \frac{1}{2} \log \log n - \frac{1}{2} \log \pi \right],$$

(14)

其中 $i = 1, \cdots, d$.

2) 双边检测. 随着数据个数 k 的不断增加,如果出现:

$$k^{-1/2} |\boldsymbol{W}_k| \geq \boldsymbol{C}_1^*(\boldsymbol{\alpha}),$$

(15)

则说明数据在 k 附近出现突变点. 同样, $\boldsymbol{\alpha}$ 为置信度向量. 而 $\boldsymbol{C}_1^*(\boldsymbol{\alpha})$ 中的各个分量可以根据式(12)计算得出:

$$\boldsymbol{C}_1^*(\boldsymbol{\alpha}) = (2 \log \log n)^{-\frac{1}{2}} \left[-\log\left(-\frac{1}{2} \log(1 - \alpha^i)\right) + 2 \log \log n + \frac{1}{2} \log \log n - \frac{1}{2} \log \pi \right],$$

(16)

其中 $i = 1, \cdots, d$.

1.3.2 无穷数据下的检测

当数据无穷多时,1.3.1 节中所提方法仅适用于有限数据检测. 这里介绍一种无穷数据下的检测

方法. 仅给出双边检测如下:

首先有式(17):

$$P\left\{\sup_{t>0} |W(t)| \geq [(t+1)(a^2 + \log(t+1))]^{1/2}\right\} = \exp\left(-\frac{1}{2}a^2\right). \quad (17)$$

在双边检测中,随着数据个数 t 的不断增大,如果出现:

$$|W(t)| \geq C_2^*(\alpha), \quad (18)$$

则说明数据在 t 附近出现突变点. 同样, α 为置信度向量. 而 $C_2^*(\alpha)$ 中的各个分量可以根据式(17)计算得出:

$$C_2^{*i}(\alpha) = [-2 \log \alpha^i + \log(t+1)]^{1/2} [t+1]^{1/2}, \quad (19)$$

其中, $i=1, \dots, d$.

2 改进突变点检测算法

传统基于 ESV 检测算法的优点在于当数据没有发生异常时,其统计量 V_k 保持在零附近,不会随着数据的增多而无限地增大;其缺点为同样存在检测延时问题:序列突变后的幅度越小延迟越严重. 此现象可以从仿真部分看到. 为了改进 ESV 算法的这一不足,这里引入小波变换理论,介绍如下.

2.1 问题描述

当引用 ESV 算法时,存在一个问题:ESV 算法中假设待检测序列服从分布 f ,而现实生活中,对于大多数时序而言,其分布却是未知的. 为了克服这一问题,本文引入基于模型的思想对其进行处理. 即首先采用时间序列鲁棒建模算法对时间序列进行在线建模. 得到模型如下:

$$x_t = g(x_{t-o}, \dots, x_{t-1}) + e_t, \quad (20)$$

其中,下标中的 o 表示模型阶次; $e_t, t=1, 2, \dots$ 表示拟合残差,服从高斯分布,即 $N(\mu, \sigma^2)$.

当时间序列 x_1, x_2, \dots 中没有出现突变点时,建立的数据模型为 $g(\cdot)$,若某时刻 τ 出现突变点,即时刻 τ 以后的数据不再符合模型 $g(\cdot)$,如果依然采用该模型对 τ 以后的数据进行拟合,将会出现较大的拟合残差,此种情况可以看成 $e_t, t \geq \tau$ 服从高斯分布的方差发生了变化. 依据此分析以及式(1),重新给出假设如下:

$$\begin{aligned} H_0: & \sigma^2 = \sigma_0^2, \text{ 对于所有观测值 } \mu \text{ 未知;} \\ H_A: & \text{对于 } e_1, \dots, e_{\tau-1}, \text{ 满足 } N(e; \mu, \sigma_0^2), \mu \text{ 未知;} \\ & \text{对于 } e_\tau, e_{\tau+1}, \dots, \text{ 满足 } N(e; \mu, \sigma_A^2), \mu, \sigma_A^2 \text{ 未知;} \end{aligned} \quad (21)$$

从式(21)可以看出,由于我们只对高斯分布中的方差 σ^2 感兴趣,因此统计量 W_k 在此之后仅是一个参数方差的统计量,即标量.

2.2 小波分析算法

如前所述,为了尽可能减少突变点检测的延迟问题,本文引入小波变换方法分析统计量 W_k 值. 其理由为突变点之前统计量 W_k 均值为零,而当突变点出现后, W_k 随着突变点数量的增多正比例增大. 利用一组实际数据的图像说明如图 1 所示:

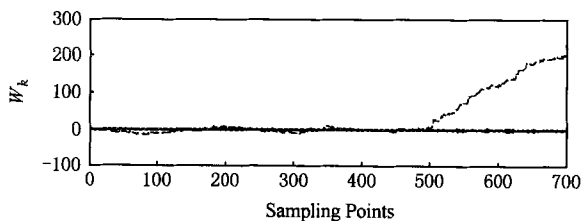


Fig. 1 The chart of W_k for change-point.

图 1 突变点时的 W_k 统计量图像

在图 1 中,横坐标表示样本数,纵坐标表示 W_k 值,纵坐标始终为零的一条直线为零基准线. 数据在 500 步时出现突变点,可以看出突变点前 W_k 值几乎为零,将其分解得到的小波系数也应在零附近;而从突变点处起, W_k 值变成斜坡函数形式,变化处(突变点所在处)的小波系数将出现模极大值. 据此,可以根据函数的 Lipschitz 指数和小波变换模极大值之间的关系^[11],通过采用小波分解 W_k 值曲线的方式确定出突变点的位置.

考虑到时间序列数据量大,需要在线突变点检测的要求,这里采用文献[20]提出的在线递推小波分解方法对 W_k 值进行在线小波分析. 由于篇幅原因,这里只给出递推小波的母小波函数如式(22)所示,以及最后推导出的递推小波分解公式如式(23)所示.

$$\psi(t) = \left(1 + \beta|t| + \frac{\beta^2}{2}t^2\right) \exp(-\beta|t|) \exp(i\omega_0 t), \quad (22)$$

其中, $\beta = 2\pi/\sqrt{3}$; $\omega_0 = 2\pi$, 此时 $\psi(0) = 0$, 保证基本小波满足容许性条件.

$$\begin{aligned} W_{x,\psi}(kT, f) = & \sqrt{f} T \{ \delta_1 x[(k-1)T, f] + \\ & \delta_2 x[(k-2)T, f] + \delta_3 x[(k-3)T, f] + \\ & \delta_4 x[(k-4)T, f] + \delta_5 x[(k-5)T, f] \} - \\ & \lambda_1 W_{x,\psi}[(k-1)T, f] - \lambda_2 W_{x,\psi}[(k-2)T, f] - \\ & \lambda_3 W_{x,\psi}[(k-3)T, f] - \lambda_4 W_{x,\psi}[(k-4)T, f] - \\ & \lambda_5 W_{x,\psi}[(k-5)T, f] - \lambda_6 W_{x,\psi}[(k-6)T, f], \end{aligned} \quad (23)$$

其中, $W_{x,\psi}(kT, f)$ 为时刻 kT 、频率 f 下的小波系数, T 为采样周期, k 为整数标记采样点.

$$\begin{aligned} h &= \exp(-fT(\beta - i\omega_0)); \\ \delta_1 &= \left[\frac{(\beta f T)^3}{3} - \frac{(\beta f T)^4}{6} + \frac{(\beta f T)^5}{15} \right] h; \\ \delta_2 &= \left[\frac{2(\beta f T)^3}{3} - \frac{5(\beta f T)^4}{3} + \frac{26(\beta f T)^5}{15} \right] h^2; \\ \delta_3 &= \left[\frac{-6(\beta f T)^3}{3} + \frac{22(\beta f T)^5}{5} \right] h^3; \\ \delta_4 &= \left[\frac{2(\beta f T)^3}{3} + \frac{5(\beta f T)^4}{3} + \frac{26(\beta f T)^5}{15} \right] h^4; \\ \delta_5 &= \left[\frac{(\beta f T)^3}{3} + \frac{(\beta f T)^4}{6} + \frac{(\beta f T)^5}{15} \right] h^5; \\ \lambda_1 &= -6h; \lambda_2 = 15h^2; \lambda_3 = -20h^3; \\ \lambda_4 &= 15h^4; \lambda_5 = -6h^5; \lambda_6 = h^6. \end{aligned}$$

从式(23)可知, 只需计算出初始的 6 个小波系数 $W_{x,\psi}$, 就可以利用前 5 个时刻的信号 x 和前 6 个时刻的小波系数计算出当前的小波系数, 实现在线小波分解, 以满足在线检测的要求. 由于式(22)为紧支撑小波, 因此初始化小波系数只需要支撑范围内的数据, 无需全部数据.

3 突变点、异常点统一的检测算法

3.1 有效分数向量的异常点表征

首先以一个例子说明突变点、异常点的 ESV 统计量 W_k 值的不同表现形式: 取一组零均值白噪声数据模拟 $e_t, t=1, 2, \dots$ 值, 在 200 步处加入异常值, 在 500 步时出现突变点, 其 W_k 值曲线如图 2 所示:

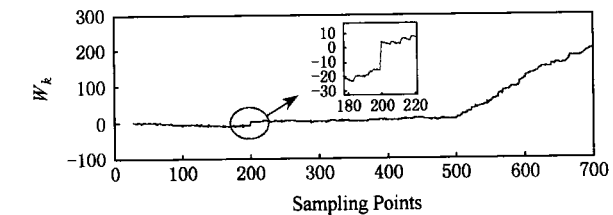


Fig. 2 The values of W_k for outlier and change-point.

图 2 异常点和突变点的 W_k 值表现

从图 2 可知, 异常点和突变点的 W_k 表现不同, 异常点处的 W_k 曲线为阶跃函数形式, 而突变点及其后数据的 W_k 曲线表现为斜坡函数形式.

3.2 小波分析算法

本文采用小波分解 W_k 值的方法检测并区分异常点和突变点. 在数学上, 利用 Lipschitz 指数(同 ν 表示)表述函数的光滑程度^[21]: 函数越光滑 ν 越大. 因此阶跃函数的 ν 值为 0, 由于斜坡函数较阶跃函

数更光滑连续, 其 ν 值为 1. Mallat 等人^[11] 在 1992 年建立了 Lipschitz 指数与小波系数的关系, 并以此提出小波变换模极大值原理, 其中小波变换模极大值与 Lipschitz 指数关系如下:

- $\nu > 0$ 时, 小波系数随小波尺度的增大而增大;
- $\nu = 0$ 时, 小波系数与尺度无关.

利用此关系可以检测并区分异常值和突变点. 具体算法如下:

- 步骤 1. 在两个小波尺度下对拟合残差 $e_t, t=1, 2, \dots$ 进行在线小波分解.
- 步骤 2. 计算两尺度下小波分解系数的模, 并计算差值得到 E_k .
- 步骤 3. 异常点、突变点检测:
 - 1) 步骤 1 中未出现模极大值点处, 并且步骤 2 中 E_k 没有突变, 说明此处 W_k 值曲线始终维持在零附近, 没有发生变化, 说明此处既没有异常点也没有突变点;
 - 2) 步骤 1 中出现模极大值, 而步骤 2 中 E_k 没有模极大值点, 说明此处两尺度下小波系数相同, 应为异常点所在处;
 - 3) 步骤 1 和步骤 2 中均存在模极大值点, 说明两尺度下小波系数不同, 应为突变点所在处.

4 仿真实验

4.1 验证

为了验证本文提出的异常点、突变点检测算法的有效性, 利用一组零均值白噪声数据模拟由数据模型得到的拟合残差值 e_t , 让其发生不同程度的突变, 并加入异常点, 形成两组待检测数据如下:

- 1) 数据的方差在 500 步处从 1 突变成 3.5, 并在 200 步处加入异常点, 形成第 1 组数据;
- 2) 数据的方差在 500 步处从 1 突变成 24.5, 并在 200 步处加入异常点, 形成第 2 组数据.

以上两组数据如图 3 所示, 对其进行异常点、突变点检测结果分别如图 4、图 5 所示.

在图 4(a) 和图 5(a) 中, 为 W_k 统计量曲线, 其中椭圆标注处为异常点所在处, 由于阶跃不明显, 将其放大后依然显示在图 4(a) 和图 5(a) 中, 可以看出异常点处 W_k 呈阶跃曲线形式, 而 500 步突变点后, W_k 呈斜坡曲线形式, 斜率大小与突变点突变程度有关. 图 4(b) 和图 5(b) 为对 W_k 统计量进行两尺度下的小波分解图像. 其中虚线为 $f=13$ 下的小波系数, 实线为 $f=15$ 下的小波系数. 可以看出无论是

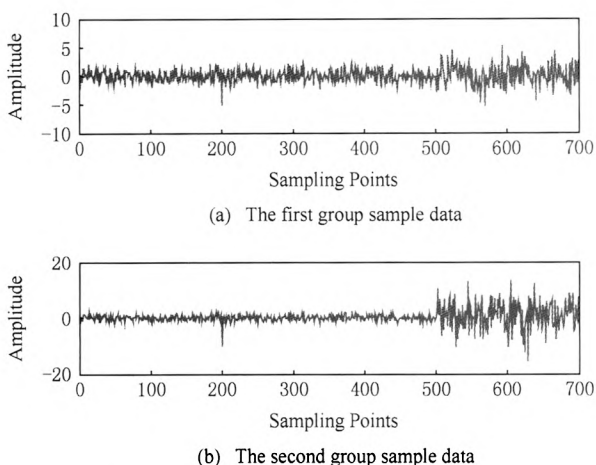


Fig. 3 Two group of data for detection.

图3 两组待检测数据

异常点还是突变点处,小波系数均出现模极大值现象.图4(c)和图5(c)为对两尺度小波系数取模后作差后的图像.可以看出异常点处差值几乎为零,说明此处小波模极大值与小波尺度无关;而突变点处差值依然很大.以此可以区分并检测出异常点和突变点.因此通过对以上两组数据的仿真可以证明,本文提出的异常点、突变点检测算法具有一定的有效性.

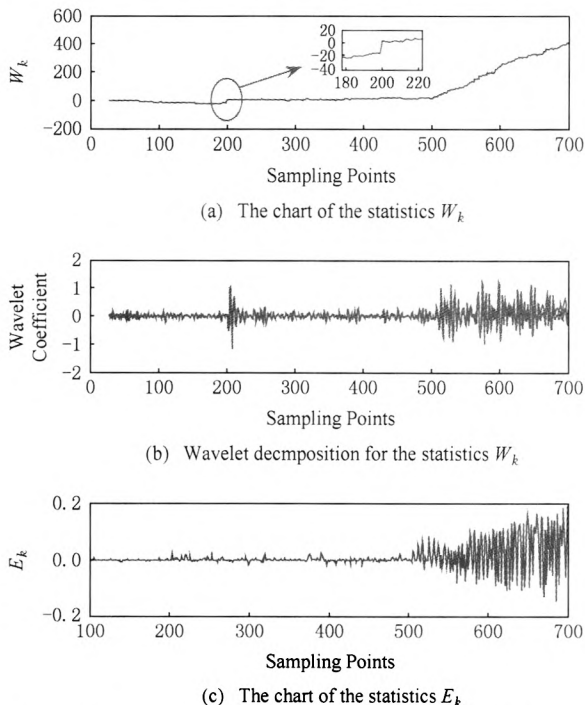


Fig. 4 Detection results for the first group of data.

图4 第1组数据的检测结果

4.2 比较

为了进一步说明本文提出的检测算法较传统基于ESV值的突变点检测算法要更优越,这里对以上两组数据采用第2节介绍的传统ESV算法进行检测,

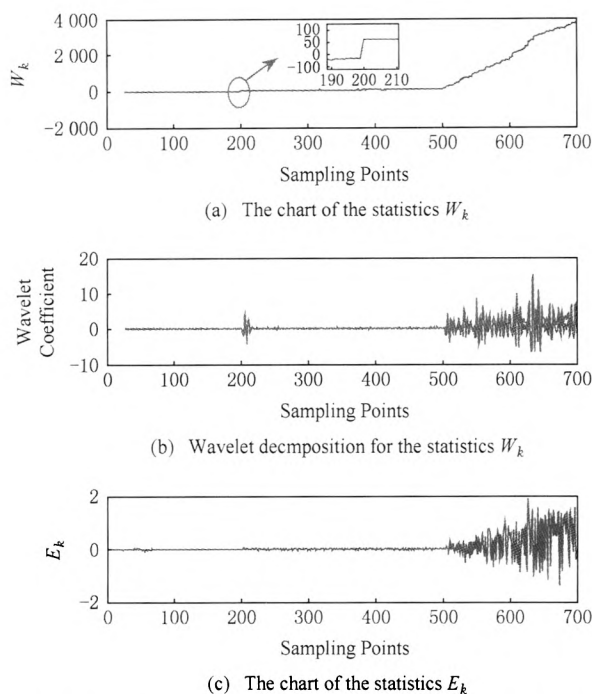


Fig. 5 Detection results for the second group of data.

图5 第2组数据的检测结果

测,由于传统的ESV算法仅可以检测突变点,因此将以上两组数据中的异常点去掉,仅比较突变点检测结果,检测结果如图6所示:

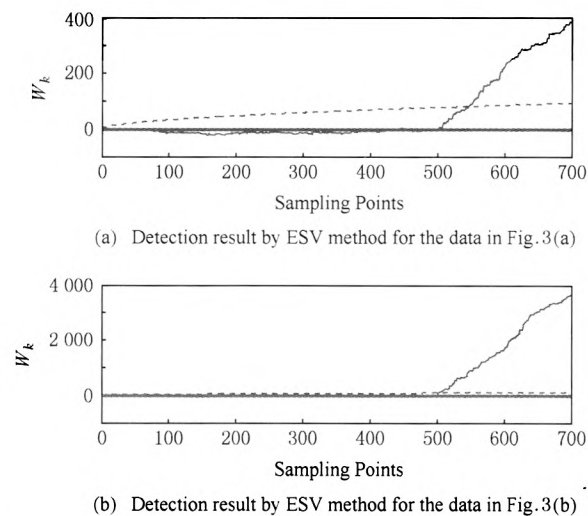


Fig. 6 Detection results by conventional ESV method.

图6 传统ESV算法检测结果

图6(a)为针对第1组数据进行突变点检测的结果;图6(b)为针对第2组数据的检测结果.图6(a)和图6(b)中,500步后出现斜坡的曲线为 W_k 统计量曲线;虚线为由式(19)计算得到的检测阈值曲线(更一般地,这里采用无限数据量检测方法);另外一条实线为零基准线,用以说明 W_k 统计量在突变点前的数值几乎为零.从图6可知,当突变

点突变幅度较小时,检测延时很大(如图 6(a)所示);而当突变点突变幅度增大时,其突变点检测延迟减小很多.对比采用本文提出的检测算法得到的检测结果(如图 4、图 5 所示)可以看出,本文提出的检测算法无论突变幅度大小,检测延时均较传统检测方法小很多.从而说明本文提出的检测方法在检测突变点方面较传统 ESV 算法更具优势.

5 结 论

本文针对短时间内难以区分异常点和突变点这一问题以及时间序列的特性,提出一种适合于时间序列的在线检测区分异常点和突变点的方法.该方法采用小波分析 ESV 统计量的方法,弥补了传统突变点检测算法中延时大、检测滞后的缺点.针对突变点以及异常点 ESV 值表现的差别以及小波模极大值原理和 Lipschitz 指数之间的关系,提出了利用小波分解 ESV 曲线的方法,实现了同时检测并区分异常值和突变点的可能,极大地减小了突变点检测的延迟问题.通过仿真实验和比较说明了本文提出的异常点、突变点检测算法具有一定的有效性和实用性.

参 考 文 献

- [1] Shao Jidong, Rong Gang, Lee Jongmin. Learning a data-dependent kernel function for KPCA-based nonlinear process monitoring [J]. Chemical Engineering Research & Design, 2009, 87(11A): 1471-1480
- [2] Zou Boxian, Liu Qiang. ARMA-based traffic prediction and overload detection of network [J]. Journal of Computer Research and Development, 2002, 39(12): 1645-1652 (in Chinese)
(邹柏贤, 刘强. 基于 ARMA 模型的网络流量预测[J]. 计算机研究与发展, 2002, 39(12): 1645-1652)
- [3] Zou X, Deng Z, Ge M, et al. GPS data processing of networks with mixed single-and dual-frequency receivers for deformation monitoring [J]. Advances in Space Research, 2010, 46(2): 130-135
- [4] Barnett V, Lewis T. Outlier in Statistical Data [M]. New York: John Wiley & Sons, 1994
- [5] Knorr E M, Ng R T. Finding intentional knowledge of distance-based outliers [C] //Proc of the 25th Int Conf on Very Large Data Bases. San Francisco: Morgan Kaufmann, 1999: 211-222
- [6] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets [C] //Proc of the ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2000: 427-438
- [7] Markou M, Singh S. Novelty detection: A review—part 2: neural network based approaches [J]. Signal Processing, 2003, 83(12): 2499-2521
- [8] Mourao-Miranda J, Hardoon D R, Hahn T, et al. Patient classification as an outlier detection problem: An application of the one-class support vector machine [J]. Neuroimage, 2011, 58(3): 793-804
- [9] Wang J S, Chiang J C. A cluster validity measure with outlier detection for support vector clustering [J]. IEEE Trans on Systems Man and Cybernetics, Part B-Cybernetics, 2008, 38(1): 78-89
- [10] Percival D B, Walden A T. Wavelet Methods for Time Series Analysis [M]. Cambridge: Cambridge University Press, 2006
- [11] Mallat S, Hwang W L. Singularity detection and processing with wavelets [J]. IEEE Trans on Information Theory, 1992, 38(2): 617-642
- [12] Gustafsson F. The marginalized likelihood ratio test for detecting abrupt changes [J]. IEEE Trans on Automatic Control, 1996, 41(1): 66-78
- [13] Guralnik V, Srivastava J. Event detection from time series data [C] //Proc of the 5th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 1999: 33-42
- [14] Sharifzadeh M, Azmoodeh F, Shahabi C. Change detection in time series data using wavelet footprints [C] //Proc of the 9th int Conf on Advances in Spatial and Temporal Databases. Berlin: Springer, 2005: 127-144
- [15] Alarcon-aquino V, Barria J A. Change detection in time series using the maximal overlap discrete wavelet transform [J]. Latin American Applied Research, 2009, 39(2): 145-152
- [16] Gombay E, Serban D. Monitoring parameter change in AR(p) time series models [J]. Journal of Multivariate Analysis, 2009, 100(4): 715-725
- [17] Gombay E. Parametric sequential tests in the presence of nuisance parameters [J]. Theory Stochastic. Processes, 2002, 8(24): 106-118
- [18] Gombay E. Change detection in autoregressive time series [J]. Journal of Multivariate Analysis, 2008, 99(3): 451-464
- [19] Gombay E. Sequential change-point detection and estimation [J]. Sequential Analysis, 2003, 22(3): 203-222
- [20] Chaari O, Meunier M, Brouaye F. Wavelets: A new tool for the resonant grounded power distribution systems relaying [J]. IEEE Trans on Power Delivery, 1996, 11(3): 1301-1308
- [21] Pittner S, Kamarthi S V. Feature extraction from wavelet coefficients for pattern recognition tasks [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1999, 21(1): 83-88



Su Weixing, born in 1980. Lecturer and PhD candidate at Shenyang Institute of Automation, Chinese Academy of Sciences. His research interests include data mining, data process and detection technology.



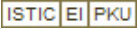
Liu Fang, born in 1983. PhD. Senior engineer in Brilliance Automobile Engineering Research Institute. Her research interests include system modeling and industrial detection technology (liufang19830311@163.com).



Zhu Yunlong, born in 1967. PhD. Professor in Shenyang Institute of Automation, Chinese Academy of Sciences. His main research interests include CIMS, distributed intelligence technology, collaborative manufacturing theory and methods, SCM/ERP/CRM systems etc (ylzhu@sia.cn).



Hu Kunyuan, born in 1972. PhD. Professor in Shenyang Institute of Automation, Chinese Academy of Sciences. His main research interests include intelligent information processing technology, intelligent optimization method and E-commerce technology (hukunyuan@sia.cn).

作者：[苏卫星](#), [朱云龙](#), [刘芳](#), [胡琨元](#), [Su Weixing](#), [Zhu Yunlong](#), [Liu Fang](#), [Hu Kunyuan](#)
作者单位：[苏卫星, Su Weixing\(中国科学院沈阳自动化研究所 沈阳110016; 中国科学院大学 北京 100049\)](#), [朱云龙, 胡琨元, Zhu Yunlong, Hu Kunyuan\(中国科学院沈阳自动化研究所 沈阳110016\)](#), [刘芳, Liu Fang\(华晨汽车工程研究院 沈阳 110027\)](#)
刊名：[计算机研究与发展](#) 
英文刊名：[Journal of Computer Research and Development](#)
年, 卷(期)：2014, 51(4)

参考文献(21条)

1. [Shao Jidong;Rong Gang;Lee Jongmin](#) [Learning a data-dependent kernel function for KPCA-based nonlinear process monitoring](#) 2009(11A)
2. [邹柏贤;刘强](#) [基于ARMA模型的网络流量预测\[期刊论文\]-计算机研究与发展](#) 2002(12)
3. [Zou X;Deng Z;Ge M](#) [GPS data processing of networks with mixed single-and dual-frequency receivers for deformation monitoring](#) 2010(02)
4. [Barnet V;Lewis T](#) [Outlier in Statistical Data](#) 1994
5. [Knorr E M;Ng R T](#) [Finding intentional knowledge of distance-based outliers](#) 1999
6. [Ramaswamy S;Rastogi R;Shim K](#) [Efficient algorithms for mining outliers from large data sets](#) 2000
7. [Markou M;Singh S](#) [Novelty detection:A review-part 2:neural network based approaches](#) 2003(12)
8. [Mourao-Miranda J;Hardoon D R;Hahn T](#) [Patient classification as an outlier detection problem:An application of the one-class support vector machine](#) 2011(03)
9. [Wang J S;Chiang J C](#) [A cluster validity measure with outlier detection for support vector clustering](#) 2008(01)
10. [Percival D B;Walden A T](#) [Wavelet Methods for Time Series Analysis](#) 2006
11. [Mallat S;Hwang W L](#) [Singularity detection and processing with wavelets](#) 1992(02)
12. [Gustafsson F](#) [The marginalized likelihood ratio test for detecting abrupt changes](#) 1996(01)
13. [Guralnik V;Srivastava J](#) [Event detection from time series data](#) 1999
14. [Sharifzadeh M;Azmoodeh F;Shahabi C](#) [Change detection in time series data using wavelet footprints](#) 2005
15. [Alarcon-aquino V;Barria J A](#) [Change detection in time series using the maximal overlap discrete wavelet transform](#) 2009(02)
16. [Gombay E;Serban D](#) [Monitoring parameter change in AR \(p\) time series models](#) 2009(04)
17. [Gombay E](#) [Parametric sequential tests in the presence of nuisance parameters](#) 2002(24)
18. [Gombay E](#) [Change detection in autoregressive time series](#) 2008(03)
19. [Gombay E](#) [Sequential change-point detection and estimation](#) 2003(03)
20. [Chaari O;Meunier M;Brouaye F](#) [Wavelets:A new tool for the resonant grounded power distribution systems relaying](#) 1996(03)
21. [Pittner S;Kamarthi S V](#) [Feature extraction from wavelet coefficients for pattern recognition tasks](#) 1999(01)

引用本文格式：[苏卫星](#). [朱云龙](#). [刘芳](#). [胡琨元](#). [Su Weixing](#). [Zhu Yunlong](#). [Liu Fang](#). [Hu Kunyuan](#) [时间序列异常点及突变点的检测算法](#) [期刊论文]-[计算机研究与发展](#) 2014(4)