基于小波变换的飞行数据清洗

毛红保[1]， 张凤鸣[1]， 冯 卉[2]
(1. 空军工程大学 工程学院，陕西 西安 710051；2. 空军工程大学 导弹学院，陕西 三原 713800)

摘 要：飞行数据因为野点和噪声的存在给其进一步处理和利用造成了困难。提出了一种基于小波变换残差直方图分析的野点识别方法，能在时间域内精确定位野点，并具有识别少量成片野点的能力。根据飞行数据噪声的特点及去噪要求，在去噪的过程中引入边缘检测，提出了分二进小波尺度乘积和小波阈值收缩两个步骤进行去噪的方法，从而在去噪的同时很好地保留了序列极值点的特性。实验结果表明本文所提方法对飞行数据中存在的质量问题具有较好的清洗效果，野点识别准确，去噪效果良好，并且对类似其它数据的处理也有一定的应用参考价值。

关键词：飞行数据；野点；去噪；小波变换；二进小波变换
中图分类号：TP391；TN911.7 **文献标识码**：A **文章编号**：1009-3516(2008)03-0011-05

军用飞机上的飞行数据记录系统，实时记录了发动机、飞机运动及航行姿态等诸多重要参数，这些数据对飞行训练质量评估、视情维修和事故分析具有极其重要的作用[1]。但是由于飞行数据是飞机在空中处于复杂的背景环境下记录的采样参数，有用信号和各种各样的干扰、误差叠加在一起，严重地影响了数据的质量，具体表现在：①数据中存在野点。野点又称离群点(Outliers)、跳点(Jump Points)或奇异点(Singularities)，是明显偏离被测信号变化规律的数据点，它不是被测对象本身正常跳变的记录，而是由传感器、变换器及无线电传输中的干扰等造成的异常跳变点。实际情况表明，野点是个别的，但它们对数据分析结果的影响却是严重的。故在进行数据分析前必须通过一定的方法对其加以判别和修正(或剔除)。②数据短期波动频繁。由于环境条件的变化或随机干扰、传感器误差等造成的影响，使得真实的参数值被大量的背景噪声所掩盖，表现为数据在短期内频繁抖动，形成锯齿状，且这些噪声不能简单地视为白噪声在一次滤波中去掉。由于飞行数据中存在上面两方面因素的影响，给其进一步处理和利用造成了困难。

关于野点的界定及识别算法受到研究人员的广泛关注[2-4]。现有的基于小波分析的野点(奇异点)识别方法主要是基于模极大值的方法[5-7]，即通过最大尺度上的模极大值沿模极大值线找到最小尺度上的奇异点小波系数。但是在小波域内很难精确定位奇异点在原信号中的位置(因为各层小波系数的长度与信号的长度不等，且存在一定的平移)，从而给奇异点的自动修正或剔除造成了困难。

1 基于残差直方图分析的野点识别

1.1 奇异信号在小波变换下的特性

通常用Lipschitz指数来描述函数的局部奇异性，对于函数$f(x)$，如果存在常数K_1及$\alpha(0\leqslant\alpha\leqslant1)$，对$x_0$邻域内的所有$x$有

$$|f(x)-f(x_0)|\leqslant K_1|x-x_0|^{\alpha} \tag{1}$$

满足式(1)的最大α即为$f(x)$在x_0处的Lipschitz指数[8]。正常的连续信号具有正的Lipschitz指数，阶跃信

收稿日期：2007-08-31
基金项目：空军工程大学工程学院创新基金资助项目(200519)
作者简介：毛红保(1979-)，男，湖北蕲春人，博士生，主要从事信息处理、时间序列分析、智能决策等研究；
E-mail：maohbao@126.com
张凤鸣(1963-)，男，重庆梁平人，教授，博士生导师，主要从事信息系统工程与智能决策等研究.

号在跃变点处的 Lipschitz 指数为 0,对于更坏的情况(野点属于这种情况),可将 α 扩展为负值[9]。

函数的 Lipschitz 指数与尺度 2^j 下的离散二进小波变换之间的关系可表示为(K_2 为常数)

$$|W_jf(x)| \leqslant K_2(2^j)^{\alpha} \tag{2}$$

从式(2)中可以看出,小波变换确实能用来估计函数的局部奇异性[6]。对于 Lipschitz 指数大于零的点,随着尺度的增加其小波变换后的幅值将呈幂增加趋势;而对于 Lipschitz 指数小于零的点,小波变换后的幅值随着尺度的增加而减小。根据这一特性,可以通过对序列的小波分解的分析来识别序列中的野点。

1.2 基于残差直方图分析的野点识别方法

基于序列中的奇异点在小波变换下所具有的数学特性,我们构造一种基于残差直方图分析的野点识别算法,通过对小波系数重构下的高频信号的分析来识别野点。设 S 为原始信号, $d(S,j)(j=1,2,\cdots,J)$ 为其 J 层小波分解后的第 j 层高频重构信号, $a(S,J)$ 为其 J 层小波分解后的低频重构信号,则称 $R(S,J)=S-a(S,J)=\sum_{j=1}^{J}d(S,j)$ 为原始信号在 J 层重构下的残差。通过残差可以发现野点与正常点有着别显的差异,图 1(a)示意了一个长度为 3 000 的飞行数据中某个参数的曲线,序列中含有 3 个野点(通常真实数据中的野点数量较多,这里为便于观察高频重构信号及残差信号在野点处的特征只保留了较少的野点),它在 3 层 db8 小波分解下的残差曲线如图 1(e)所示。图 1(b)-(d)为各层高频重构信号。

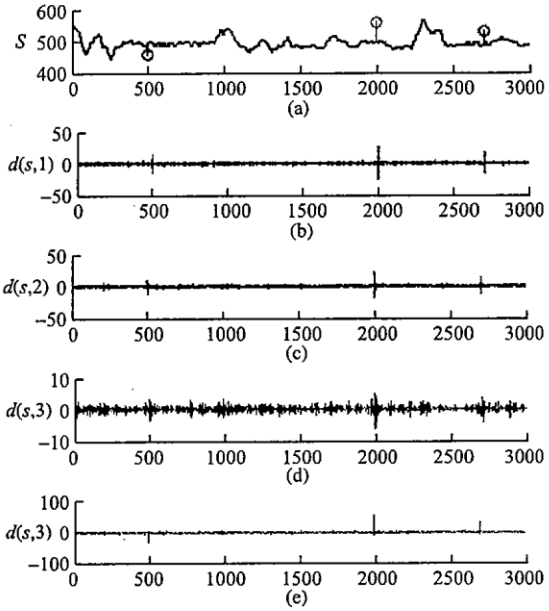

图 1 含有野点的飞行数据在小波分解下的高频重构序列及残差序列

Fig. 1 The high frequency reconstructed series and residuals of DWT to flight data with outliers

从图 1(e)中可以看出,残差对野点非常敏感,它在野点处的幅值明显比正常点突出,且野点周围点的残差并不受影响。虽然前两个尺度的高频重构信号(图 1(b)、(c))在野点处的幅值也非常突出,但野点邻近点也受到了影响(幅值也很突出,这是由于小波分解时的移位造成的),从而给野点的精确定位造成了困难。若前第三个尺度的高频重构信号(图 1(d))叠加起来,根据 1.1 节中 Lipschitz 指数与小波变换之间的关系,由于正常的信号点在小波变换后的幅值随着尺度的增加迅速增长,从而弥补了野点周边的点在前两个尺度上受到的影响。根据残差的这一特性,可以通过设置阈值的方法从残差序列中过滤出野点并对其修正。但是如何选择合理、有效的阈值作为过滤条件是一个值得研究的问题。本文利用残差序列的直方图分析来确定野点处残差的阈值,该方法非常简单,并且具有阈值选择自适应的特点,经仿真实验表明也是非常有效的。

图 2 所示为残差序列 $R(S,J)$ (简记为 R)的直方图,横坐标表示残差的取值,将残差的最小值 $\min(R)$ 与最大值 $\max(R)$ 之间的区间分成 L 个相等的小区间 $[\min(R)+\frac{\max(R)-\min(R)}{L}(i-1),\min(R)+\frac{\max(R)-\min(R)}{L}i]$, $(i=1,2,\cdots,L)$,纵坐标为每个小区间上序列点的个数。如果选择合适的 L 数目(具体确定方法后文介绍),可使正常点的残差区间是连续的,且正常点的残差区间与野点的残差区间之间是不连续的,如图 2 所示,从序列点个数最多的区间开始分别向两边检查,直到找到序列点个数为 0 的区间,将该区间对应的幅值作为阈值 λ^+ 和 λ^-,残差落在区间 $[\lambda^-,\lambda^+]$ 之外的点被识别为野点。

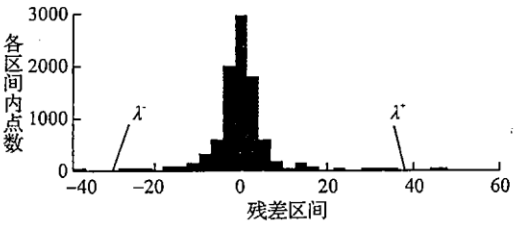

图 2 残差序列的直方图

Fig. 2 The histogram of residuals

区间个数 L 的确定很关键,过小的 L 值会将野点的残差值与正常点的残差值划分到同一个区间中,而过大 L 值会使区间个数过多,导致正常的残差值构成的区间不连续。经反复实验发现,利用下面的公式确定区间划分数目比较合适:

$$L=(\max(R)-\min(R))/\mathrm{std}(R) \tag{3}$$

其中 std(R) 表示序列 R 的标准差,即序列的标准差越大则每个小区间的划分宽度也越大,因为较大的标准差说明原始序列 S 的波动性较强,通过较宽的区间划分来防止将序列本身的极值点误判为野点。

因为不同干扰因素的影响,野点偏离正常值的程度也可能有很大差异(有些野点偏离正常值非常多,也有些野点非显著地偏离正常值),因此上面的野点识别算法在实际运用中需多次使用,每次识别出野点后按一定的方法对其进行修正(如插值法),然后再对修正后的序列重复该算法,直到没有识别出野点为止(所有残差区间上的序列点个数都不为0)。

1.3 成片野点的处理

在实际的飞行数据中,除了会出现单个的野点,还可能出现少量数值相等(或大致相等)的野点连成一片的情况。针对少量成片野点的识别问题,只需对上面的算法作一些改进。因为成片野点在进行小波变换的时候会影响与其相邻的正常点的小波系数,并且这些边缘点的残差也具有较大的幅值。但野点与边缘点的影响情况恰好相反,他们的残差符号是相异的。根据这个特性,在每一轮野点识别完后,如果检测到成片的野点出现,需对两侧的野点进行重新校验,若它与相邻的野点残差值是异号的,说明它是伪野点,取消其野点标记。通过对边缘处的野点进行校验处理,从而解决了少量成片野点的识别问题。

2 边缘检测与小波收缩相结合的飞行数据去噪

目前广泛使用的 Donoho 和 Johnstone[10-11]提出的小波阈值收缩去噪方法基本上都是在噪声满足高斯分布的前提下推导出来的,而飞行数据因为传感器采样精度、环境因素等造成的影响,其噪声通常不能按高斯分布对待(见图3,它是图1(a)中抽取一段放大后的曲线),因此通过一次滤波处理很难达到较好的去噪效果(大量的实验也证明了这一点)。为了达到既能去除噪声,具有较好的视觉效果,又忠实原始数据(保留信号在关键点的特征)的要求,本文分两次对飞行数据进行滤波处理,首先通过离散二进小波变换(简称二进小波变换)检测信号的边缘点并去除一部分非边缘点的噪声,第二次再对信号进行全局小波阈值收缩,实验证明取得了较好的去噪效果。

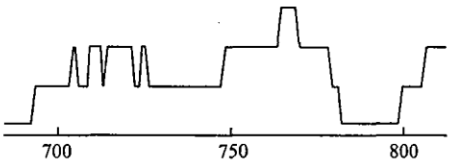

图3　飞行数据噪声的特点

Fig. 3　The noising characteristic of flight data

2.1 基于二进小波尺度乘积的第一次去噪

为实现去噪的同时尽量保留信号在关键点的特征,在去噪的同时引入边缘检测的思想是一个很好的选择,其中利用二进小波尺度乘积是一种较好的将边缘检测与去噪相结合的方法[12-13]。二进小波变换的滤波器可以通过离散小波变换滤波器的变换获得,且两者具有相同的算法结构,所不同的是二进小波变换在每次小波分解时并没有向下采样,因此对信号的每次小波分解都产生与原始信号长度相同的小波系数和尺度系数,从而能实现相邻尺度下小波系数的乘积。并且,二进小波变换是一种非正交变换,信号的表达存在冗余,部分小波系数的的扰动不会带来信号的严重失真,从而减小了信号重构效果对单个小波系数的依赖性。

设 $\boldsymbol{W}_j$ 为第 j 层二进小波分解时的小波系数向量,二进小波变换相邻尺度的小波系数乘积定义为

$$\boldsymbol{P}_j = \boldsymbol{W}_j \cdot \boldsymbol{W}_{j+1} \qquad (j = 1, 2, \cdots, J-1) \tag{4}$$

其中"·"表示两个向量的点积运算,J 为最大分解层数。文献[13]给出的从乘积序列中提取边缘处小波系数的方法比较复杂,且需人为指定一个常数,实现上不方便且不直观。为此,本文将乘积序列 P_j 中幅值较大的前 $\alpha\%$ 个系数作为序列边缘处的小波系数予以保留($\alpha\%$ 值可根据需要指定,一般取10%左右),其它系数置零。将经过上述处理之后的乘积序列 $P_j \sim P_{J-1}$ 作为前 $J-1$ 层小波系数进行二进小波重构,即得到第一次滤波处理之后的信号。

2.2 基于小波阈值收缩的第二次去噪

在第一次去噪结果的基础上,我们采用目前使用较多的小波阈值收缩方法进行第二次去噪,具体分如下3个步骤进行:①对信号进行离散小波变换,得出各尺度小波分解系数;②应用阈值函数处理各尺度小波系数的估计值;③在各尺度小波系数估计值的基础上应用离散小波逆变换,即得到去噪后的信号。

上面第②步阈值的确定至关重要,它是区分信号和噪声的分水岭,阈值太高会引起信号失真,太低则又去噪不完全。我们采用硬阈值函数,可描述为

$$d_{j,i}=\begin{cases}d_{j,i}-\lambda, d_{j,i}\geqslant\lambda\\ 0, |d_{j,i}|<\lambda\\ d_{j,i}+\lambda, d_{j,i}\leqslant-\lambda\end{cases} \tag{5}$$

式中：$d_{j,i}$为第 j 尺度下的第 i 个小波系数，阈值 $\lambda=\tilde{\sigma}\sqrt{2\log_2 N}$。$N$ 为信号的长度，$\tilde{\sigma}$ 为信号的标准偏差估计。$\tilde{\sigma}$ 的估计值可取为 $\tilde{\sigma}=\mathrm{Med}(|\cdot|)/0.6745$，其中 $\mathrm{Med}(\cdot)$ 表示该尺度下小波系数绝对值的中值。

另外，在使用阈值去噪方法时小波系的选取和分解层数的确定对去噪效果也有重要影响，本文采用 Db 系列小波来进行飞行数据去噪(具有正交性、紧支撑性和一定的消失矩)；分解层数也不宜过多，因为在第一次二进尺度乘积的去噪中已经滤去了大部分噪声，通常进行 3 层左右的分解即可。

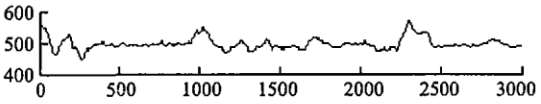

图 4 基于残差直方图分析识别并修正野点后的序列

Fig. 4 The series after recognizing and correcting outliers based on residuals histogram analysis

3 实验结果

为了验证本文提出的算法，我们以图 1(a)所示的飞行数据进行实验(从一个飞行架次中截取的一段长度为 3 000 的某个参数的飞行数据)。图 4 为运用基于残差直方图分析的野点识别算法识别并修正野点之后得到的序列。在此基础上，我们再进行飞行数据去噪，图 5 为基于二进小波尺度乘积的第一次去噪后得到的序列(取 $\alpha=10\%$)，从图中可以看出，去噪后的序列在局部极值点处保持了原有的特性，但在其它地方还存在一些明显的噪声(此时的噪声可以视为白噪声)。再对其进行基于小波阈值收缩的第二次去噪(采用硬阈值)，去噪结果如图 6 所示。本文的实验在 Matlab7.1 下进行，采用的小波是 db8 小波，分解尺度数为 3。

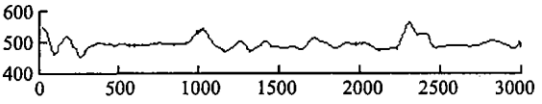

图 5 基于二进小波尺度乘积第一次去噪后的序列

Fig. 5 The series after the first time denoising with dyadic wavelet coefficients product

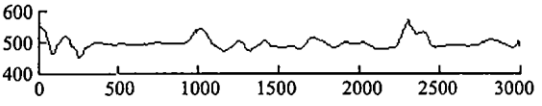

图 6 基于小波阈值收缩第二次去噪后的序列

Fig. 6 The series after the second time denoising with wavelet shrinkage

4 结论

飞行数据因其野点的存在及由于干扰和误差引起的频繁抖动，给其进一步的处理和利用造成了困难。本文提出了一种基于残差直方图分析的野点识别方法，在无需任何数据先验物理知识的前提下便能准确识别野点，并具有识别少量成片野点的能力。根据飞行数据的噪声不满足高斯分布的特点，提出了基于边缘检测和小波阈值收缩两个步骤进行去噪的思想，从而在保留数据关键特征的基础上具有较好的去噪效果。实验证明本文提出的飞行数据清洗方法能较好地解决飞行数据中存在的主要质量问题，为其后续的处理和利用提供了一个良好的数据环境。同时，本文提出的方法对其它类似数据的清洗问题也具有一定的参考价值。

参考文献：

[1] 张 亮，张凤鸣，惠晓滨，等. 一种基于动态模糊神经网络的飞机数据模型辨识方法[J]. 空军工程大学学报：自然科学版，2006，7(6)：16－18.
ZHANG Liang, ZHANG Fengming, HUI Xiaobin, et al. An Indentification of Flight Data Model Based on Dynamic Fuzzy Neural Network[J]. Journal of Air Force Engineering University: Natural Science Edition, 2006, 7(6): 16－18. (in Chinese)

[2] Victoria J Hodge, Jim Austin. A Survey of Outlier Detection Methodologies[J]. Artificial Intelligence Review ,2004,(22): 85－126.

[3] Qi Hongwei ,Wang Jue . A Model For mining Outliers From Complex Data Sets[C]. 2004 ACM Symposium on Applied Computing, Nicosia, Cyprus: SAC, 2004.

[4] Zbigniew R Struzik, Arno P J M Siebes. Wavelet Transform Based Multi - fractal Formalism in Outlier Detection and Localisation for Financial Ttime Sseries[J]. Physica A309,2002:388 - 402.

[5] Zhang W, Liu B, Zhang X T, et al. Application of the Wavelet based Multi - Fractal for Outlier Detection in Financial High - Frequency Time Series Data[J]. IEEE International Conference on Engineering of Intelligent Systems, 2006, 4: 1 - 6.

[6] 张小飞, 徐大专, 齐泽锋. 基于模极大值小波域的去噪算法研究[J]. 数据采集与处理, 2003, 18(3): 315 - 318.
ZHANG Xiaofei, XU Dazhuan, QI Zefeng. Denoising Algorithm Based on Modulus Maximum Wavelet Field[J]. Journal of Data Acquisition & Processing, 2003, 18(3): 315 - 318. (in Chinese)

[7] 郑　诚, 舒　坚. 多尺度时间序列的异常事件检测[J]. 计算机工程与应用, 2006, 42(31): 50 - 52.
ZHENG Cheng, SHU Jian. Multiscale Detection of Aberrant Events in Time Series[J]. Computer Engineering and Application, 2006, 42(31): 50 - 52. (in Chinese)

[8] Mallat S, Hwang W L. Singularity Detection and Processing with Wavelets[J]. IEEE Transactions on Information Theory, 1992, 38(2): 617 - 643.

[9] Zhang L, Bao P. Edge Detection By Scale Multiplication in Wavelet Domain[J]. Pattern Recognition Letters, 2002, 23(4): 1771 - 1784.

[10] Donoho D L. Denoising by Soft - Thresholding[J]. IEEE Trans on Information Theory, 1995,41(3):613 - 627.

[11] Donoho D L, Johnston I M. Ldeal Spatial Adaptation by Wavelet Shrinkage[J]. Biometrika, 1994,81(3):425 - 455.

[12] 郭显久, 王　伟. 基于尺度乘积与小波收缩相结合的去噪方法[J]. 控制与决策, 2005, 20(6): 698 - 701.
GUO Xianjiu, WANG Wei. Denoising Method Based on Combining Multi - Scale Product with Wavelet Shrinkage[J]. Control and Decision, 2005, 20(6): 698 - 701. (in Chinese)

[13] Zhang L, Bao P. Edge Detection by Scale Multiplication in Wavelet Domain[J]. Pattern Recognition Letters,2002,23(4): 1771 - 1784.

(编辑:姚树峰)

Flight Data Cleaning Based on Wavelet Transforms

MAO Hong - bao[1], ZHANG Feng - ming[1], FENG Hui[2]

(1. Engineering Institute, Air Force Engineering University, Xi'an 710038, China; 2. Missile Institute, Air Force Engineering University, Sanyuan 713800, Shaanxi, China)

Abstract: Outliers and noise will cause difficulties during processing and using flight data. This paper proposes an outlier detection method based on histogram analysis of wavelet transform residuals, which can locate outliers in time - domain precisely, and can recognize little outliers in succession. Then according to the characteristics of flight data noise and its denoising demand, edge detection is introduced and a two - step denoising method including dyadic wavelet coefficients product and wavelet shrinkage is put forward, which can keep the characteristic of extremum points very well. Finally the experiment shows that the method presented in this paper is effective on flight data cleaning, with which outliers can be recognized exactly and denoising effect is good. The method can also be used for reference in processing other similar data.

Key words: flight data; outlier; denoising; wavelet transform; dyadic wavelet transform