

DOI :10.3876/j.issn.1000-1980.2008.05.015

融合小波分析检测大坝数据仓库中数据孤立点

李婷婷

(江苏省体育彩票管理中心,江苏 南京 210029)

摘要:为充分利用大坝迅速增长的海量监测数据,剔除大坝数据仓库中有害的错误数据,为大坝安全监控提供有效的决策支持,探讨了以数据仓库作为基础分析平台,利用小波分析的时频局域化等特点检测大坝数据仓库中数据孤立点的方法,有效地提高了大坝数据仓库中数据信息的质量。

关键词:数据仓库 小波分析 数据孤立点 大坝

中图分类号:TV698.1 **文献标识码:**A **文章编号:**1000-1980(2008)05-0650-04

通常,大坝安全监测信息来源于多个监测数据采集系统和多种文件。监测数据中存在众多噪声数据、空缺数据和不一致数据。产生这些数据的因素很多,如,不同监测项目采用不同类型的监测仪器、数据采集设备故障、人工录入资料 and 操作的错误、数据存储系统的问题等^[1]。此外,不同的源数据库中表的命名规则不同,也会产生数据的不一致问题。

大坝数据仓库的设计是靠数据驱动的。数据仓库着眼于有效地抽取、综合、集成和挖掘已有源数据库的数据资源^[2],而数据质量的好坏直接影响到构建数据仓库的成功与否,进而影响到大坝决策支持及趋势分析预测等。同时,数据仓库在进行数据建模分析时,受污染的噪声数据将对建模分析结果产生很大的不利影响,甚至导致错误的结果或系统模型。因此,要使噪声数据在建模分析、规则精确性等方面影响最小,数据仓库必须对数据孤立点进行预处理,从大量的数据中剔除有害的错误数据并清洗受污染的数据。

现有的数据仓库在进行数据孤立点检测时还存在一些局限性。如,数据仓库通常是依据已知噪声数据分布的统计技术来处理噪声的,但是在数据集中各数据项之间关系未知的情况下,检测噪声数据比较困难。在大坝安全监控领域,采集的监测数据信息既具有时变特点,同时又经常受到大坝运行中的工程加固、改造等许多不可预测因素的影响。因此,仅仅依靠数据仓库进行数据孤立点检测还难以达到很好的效果^[3-6]。

近年来不断走向成熟的小波分析方法因其优良的时频局域化等特点为信号去噪提供了一种新的解决途径。小波分析的主要优点是在时间(或空间)域和频率域中同时具有良好的局部化性质,从而非常适合具有时变信号的大坝监测数据信息的分析和处理。为此,本文充分利用了小波分析的时频局域化等特点,提出了融合小波分析检测大坝数据仓库中数据孤立点的方法。

1 小波分析的特点

小波分析是 20 世纪 80 年代发展起来的一种新的数学方法,是对傅里叶分析的重大突破。小波即小区域的波,是一种特殊的长度有限、平均值为 0 的波形。小波变换的定义是把某一被称为基本小波(也叫母小波 mother wavelet)的函数 $\varphi(t)$ 位移 τ 后,再在不同尺度 a 下与待分析的信号 $x(t)$ 做内积:

$$T_x(a,\tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \varphi^* \left(\frac{t-\tau}{a} \right) dt \quad (a > 0) \tag{1}$$

式中: φ^* —— $\varphi(t)$ 的共轭函数; $\varphi(t)$ ——小波函数; $x(t)$ ——待分析函数; a ——尺度因子; τ ——平移因子; $T_x(a,\tau)$ ——小波系数。

其等效的频域表示为

$$T_x(a, \tau) = \frac{\sqrt{a}}{2\pi} \int_{-\infty}^{\infty} X(\omega) \phi^*(a\omega) e^{+j\omega\tau} d\omega \quad (2)$$

式中 $X(\omega)$ 和 $\phi^*(\omega)$ 分别是 $x(t)$ 和 $\varphi(t)$ 的傅里叶变换。

可以打个比喻形象地理解式(2)的意义:用镜头观察目标 $x(t)$ (即待分析信号), $\varphi(t)$ 代表镜头所起的作用(如滤波或卷积), τ 相当于使镜头相对于目标平行移动, a 的作用相当于镜头向目标推进或远离。由此可见小波变换具有以下 3 个特点^[7-9]:

a. 具有多分辨率(multi-resolution),也叫多尺度(multi-scale)的特点。采用小尺度 a 的小波变换对应的分析区间短,能够获取 $x(t)$ 的变换细节;反之采用大尺度 a 的小波变换能够获取 $x(t)$ 变化的概貌(即变化的总体趋势)。因而小波分析可以由粗及细地逐步观察信号。

b. 可以看成用基本频率特点为 $\phi^*(\omega)$ 的带通滤波器在不同尺度 a 下对信号做滤波。由傅里叶变换的尺度特点可知,这组滤波器具有品质因数恒定,即相对带宽(带宽与中心频率之比)恒定的特点。

c. 适当选择基本小波,使 $\varphi(t)$ 在时域上为有限支撑, $\phi^*(\omega)$ 在频域上较集中,就可以使 $T_x(a, \tau)$ 在时域、频域都具有表征信号局部特征的能力,因此有利于检测信号的瞬态或奇异点。

正是由于上述特点,小波变换被誉为分析信号的数学显微镜。

2 数据孤立点

在数据集中经常存在一些数据对象,它们不符合数据的一般模型,与数据的其他部分不同或不一致,这样的数据被称为数据孤立点。数据孤立点可能是数据仓库对源数据进行度量转换时或执行错误所导致的,也可能是大坝源数据库中固有的数据变异,因此,并非所有的孤立点都是错误数据。对于大坝数据仓库而言,在检测出数据孤立点之后还需要结合大坝安全监控领域专业知识和数据仓库中所存储的元数据信息进行综合分析,最终找出真正的错误数据。

由于大坝安全监测数据信息可以看作是不同频率成分组成的数字信号,数据孤立点可以看作是信号的奇异点,为此,可以利用小波分析的信号奇异性理论进行数据孤立点的检测。

3 小波分析的信号奇异点

信号的奇异点就是指信号中的突变点。通常情况下,信号奇异点分为 2 种类型:第 1 种类型奇异点是信号在某一个时刻内,其幅值或频率的突变处;第 2 种类型奇异点是信号外观上很平滑,幅值没有突变,但是信号的一阶微分有突变产生,且一阶微分是不连续的^[10]。

Lip 指数是表征信号局部奇异点特征的一种量度,定义如下:设有正整数 $n, n \leq \alpha \leq n+1$, 如果存在正整数 $A > 0$ 及 n 次多项式 $P_n(x)$, 使得

$$|f(x) - P_n(x - x_0)| \leq A |x - x_0|^\alpha \quad (3)$$

对于 $x \in (x_0 - \delta, x_0 + \delta)$ 成立,则称 $f(x)$ 在点 x_0 为 $\text{Lip}\alpha$ 的。

函数在某一点的 Lip 指数表征了该点奇异性的大小。 α 越大,该点的光滑度越高; α 越小,该点的奇异性越大。函数在某一点可导,它的 $\alpha \geq 1$; 函数在某点不连续但其值有限,则 $0 \leq \alpha \leq 1$ 。对一般信号而言, $\alpha \geq 0$ 。脉冲信号的 $\alpha = -1$; 白噪声的 $\alpha < 0$ ($\alpha = -\frac{1}{2} - \varepsilon, \varepsilon > 0$)。

在尺度 s 下,若 $\forall x \in \delta x_0$, 有

$$|Wf(s, x)| \leq |Wf(s, x_0)| \quad (4)$$

则称 x_0 为小波变换在尺度 s 下的局部模极大值点。信号 $f(x)$ 的 Lip 指数与小波变换模极大值满足

$$\log_2 |W_2^j f(t)| \leq \log_2 k + j\alpha \quad (5)$$

式中 k 为与函数 $\varphi(t)$ 相关的常数,尺度 $j = 1, 2, 3, \dots$ 。

由(5)式可知,对于一般信号,由于 $\alpha \geq 0$,小波变换的模极大值将随着 j 的增大而增大;而对于白噪声,由于 $\alpha < 0$,其模极大值随着 j 的增大而减小。因此,观察不同尺度间小波变换模极大值变化的规律,去除幅度随尺度的增加而减小的点(对应噪声的极值点),保留幅度随尺度增加而增大的点(对应于有用信号的极值点),然后再由保留的模极大值点用交替投影法进行重建,即可以达到去除数据孤立点的目的。

4 小波分析检测大坝数据仓库的数据孤立点

利用小波分析检测大坝数据仓库中数据孤立点的基本方法是 对监测数据(信号)进行多尺度分析,在信号出现突变时,其小波变换后的系数具有模量极大值,因而可以通过对模量极大值点的检测来快速检测大坝监测数据中的数据孤立点.

以古田溪一级大坝的引张线 2120M00EX11-1 监测点自 1999 年 2 月 26 日至 2004 年 9 月 3 日的监测数据序列为例,经多种小波检测并比较发现,只需检测该监测数据序列中的数据孤立点时,选用 dmey 小波进行一维离散 5 层分解效果比较好,分解的效果如图 1 所示.其中 s 为含有数据孤立点的原始信号. d_1, d_2, d_3, d_4, d_5 分别对应着第 1~5 层的高频部分, a_5 对应着第 5 层的低频部分.

在图 1 中,第 1 层(d_1)的高频部分在 $t = 1622$ (t 表示时间,图中对应的监测日期为 2004 年 2 月 9 日)处有 1 个模量极大值点,很显然为信号的奇异点.另外,在第 2 层(d_2)上有几处模量值较大的点,说明可能存在数据孤立点,还需要结合大坝安全监控领域专业知识和数据仓库中所存储的元数据信息进行综合分析来确定.

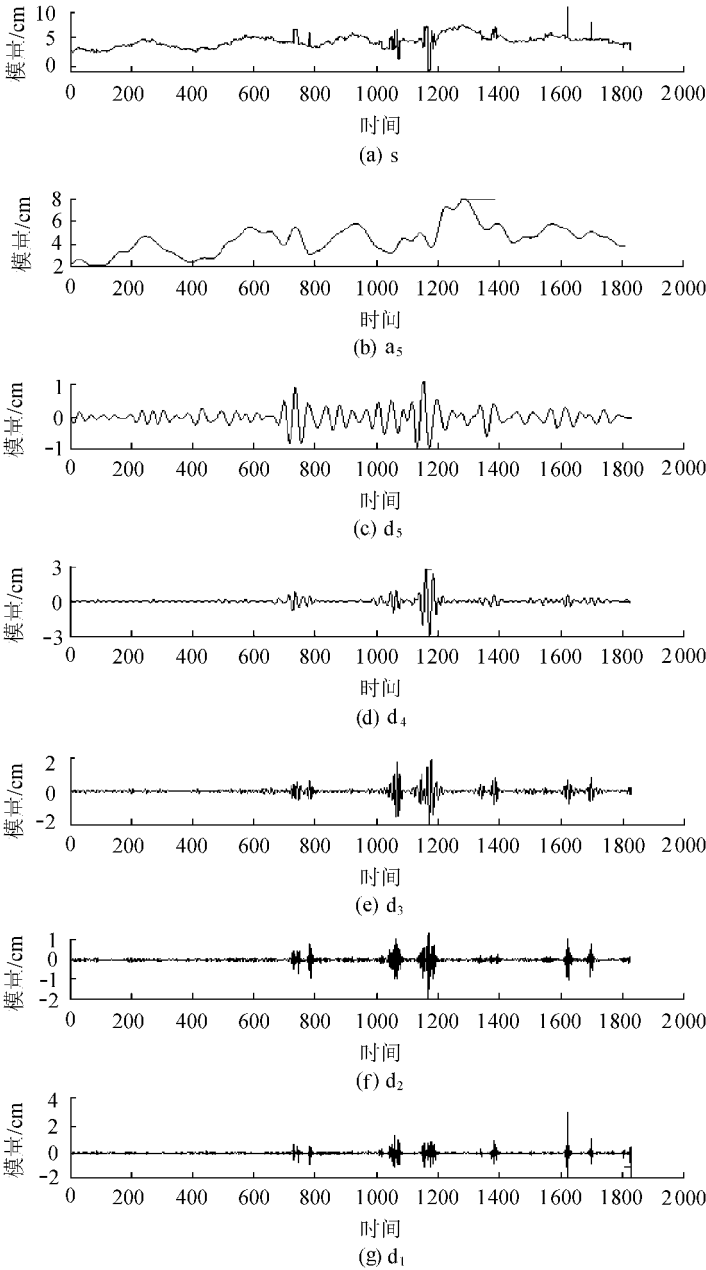


图 1 一维离散小波 5 层分解

Fig.1 Five-level 1-D wavelet decomposition

由此可见,采用小波分析的信号奇异性理论能快速精确地检测出大坝数据仓库中数字信号的孤立点. 另外,由于利用小波分析检测数据孤立点只是根据效应量序列本身之间的内在联系进行检验,未涉及效应量的成因,如环境量(水位、温度、时效)等影响因素.在实际应用中,应首先利用小波分析初步检测出大坝监测数据信息中的数据孤立点及其监测点的信息.然后依据数据仓库中对该监测点的详细描述(如监测点处坝底高程、补强加固记录等)以及相关元数据信息,进行再次分析并确定数据孤立点.最后再采用大坝安全监控的数学模型(统计模型、确定性模型、混合模型等)对已确定的数据孤立点进行综合分析检验.当实测值与拟合值之间的残差未超过 $\pm 3S$ (S 为标准差)时,说明这些孤立值属于环境、施工等因素显著变化引起的正常测值;否则,属于真正的数据孤立点.

5 结 语

本文根据大坝安全监测信息的特点,提出了融合小波分析检测大坝数据仓库中数据孤立点的方法,并分析了其局限性.实际应用表明,该方法有效地提高了大坝数据仓库中数据信息的质量,为更准确高效进行大坝健康诊断和决策分析提供了数据信息保障.

参考文献:

[1] 吴中如,顾冲时.大坝原型反分析及其应用[M].南京:江苏科学出版社,2000.
[2] 胡侃,夏绍玮.基于大型数据仓库的数据采掘[J].软件学报,1998,9(1):5-7.
[3] 宋小刚,李德仁.基于数据仓库技术的大坝资料分析与安全决策系统研究[J].河海大学学报:自然科学版,2006,34(3):280-284.
[4] 黄红女,华锡生,宋小刚.大坝安全监测数据仓库的设计与构建:以小浪底大坝为例[J].水利水电科技进展,2005,25(4):56-58.
[5] 房道伟,周国义.应用数据仓库技术创建金融信息管理系统[J].河海大学常州分校学报,2007,21(3):55-57.
[6] 李婷婷,包腾飞.大坝群自动监测报警系统[J].水力发电学报,2005,24(5):119-122.
[7] JAIDEBA C,CHAN G I.小波分析理论、算法及其应用[M].许天周,黄春光,译.北京:国防工业出版社,2007.
[8] 刘明才.小波分析及其应用[M].北京:清华大学出版社,2005.
[9] 王文圣,向红莲,黄伟军,等.基于连续小波变换的径流分维研究[J].水利学报,2005(5):10-13.
[10] 毕玲.小波聚类算法的研究及应用[D].大连:大连理工大学,2006.

Detection of mutation data combining wavelet analysis with dam data warehouse
LI Ting-ting
(Sports Lottery Management Center of Jiangsu Province, Nanjing 210029, China)

Abstract: In order to fully utilize the increasingly abundant dam monitoring data, to get rid of erroneous data in the dam data warehouse and to provide effective decision-making support for dam safety monitoring, a method of detecting mutation data in the dam data warehouse was proposed using the data warehouse as the base platform and time-frequency localization wavelet characteristic. The results of a case study show that the method effectively improves the quality of information in the dam data warehouse.

Key words: data warehouse; wavelet analysis; mutation data; dam