文章编号: 1001-0920(2005)05-0521-04

基于小波包的时间序列变点探测算法

李订芳[1,2], 胡文超[1], 章 文[1]
(1. 武汉大学 数学与统计学院, 湖北 武汉 430072; 2. 武汉大学 计算机学院, 湖北 武汉 430072)

摘 要: 基于小波能对信号的各个频率段进行分离的原理, 设计了基于小波包分析的变点探测算法, 并以此来研究时间序列的突变. 通过对长江宜昌站的年径流水文时间序列进行实验分析发现, 在过去120 a, 长江宜昌站的年最小流量序列和年平均流量序列的均值都极有可能存在着突变, 而且其变化趋势都是均值明显减小. 该方法不需要对时间序列作任何概率分布和相依性的假定, 便能方便地通过调节小波包变点探测算法的参数应用于其他领域.

关键词: 小波包分析; 时间序列; 变点分析; 变点探测算法
中图分类号: TP18 **文献标识码**: A

Wavelet packet based time series change-points detecting algorithm

LI Ding-fang[1,2], *HU Wen-chao*[1], *ZHANG Wen*[1]
(1. School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China; 2. School of Computer, Wuhan University, Wuhan 430072, China. Correspondent: LI Ding-fang, E-mail: whuldf@gsmail.whu.edu.cn)

Abstract: Based on the principle that different frequency signal can be separated from each other by wavelet, a change-points detection algorithm on the basis of wavelet packet analysis is designed to study the abrupt change of the mean value of time series. It is applied to detect change-points of the annual discharge series of Yangtze river at Yichang hydrological station. The results show that, during the past 120 years, the mean values of both the annual minimum discharge series and the annual average discharge series are very likely to have changed abruptly, and the change tendency for both series is that the mean value has significantly reduced. This algorithm does not need any probability distribution and interdependent hypothesis to the time series, and can be conveniently applied to other fields by regulating its parameters.

Key words: wavelet packet analysis; time series; change-point analysis; change-points detection algorithms

1 引 言

近年来, 变点检测已引起人们的广泛兴趣, 具有重要的理论意义和应用背景, 如质量控制、故障检测、气候突变检测、金融数据中突变点检测以及图像处理中的边缘提取等. 时间序列中变点探测的中心问题是确定变点的个数和位置. 熊立华等应用贝叶斯方法研究了水文时间序列的变点检测问题[1]; 朱颖元等应用Mann-Kendall方法、最小方差法、有序聚类分割法对降水量序列的统计特性进行了分析[2]; 丁晶等详细介绍了(加权)最小二乘法、极大似然法在水文时间序列变点检测中的应用[3]; Lavielle等基于MCMC的贝叶斯方法研究了多变点的探测问题[5]. 但由于大多数变点探测算法是参数化的, 需要对问题作某些假定, 有着很大的局限性.

小波分析能将由不同频率成分组成的复杂时间序列分解成频率不相同的子序列. 本文基于这一思想, 将时间序列分解成不同尺度下的小波系数和尺度系数, 由分解所得到的系数得到不同尺度下的能

收稿日期: 2004-07-14; **修回日期**: 2004-09-21.
基金项目: 国家重大基础研究前期专项项目(2003CCA00200); 国家重点实验室开放基金项目(2004B009); 湖北省自然科学基金项目(201130485).
作者简介: 李订芳(1966—), 男, 湖南平江人, 副教授, 博士后, 从事数据库、数据挖掘与水利信息化的研究; 胡文超(1982—), 男, 湖北仙桃人, 硕士生, 从事数据挖掘与知识发现的研究.

量及由这些能量作为分量构成的特征向量; 然后比较变点前后两个特征向量之间的距离度量, 进而确定可能存在的变点. 基于上述思想, 本文建立了一个具有非参数化特征的小波包变点探测算法, 并将其应用于判断长江宜昌站的年径流资料系列中可能存在的变点. 所得结果与文献[1]以及实际相符合. 另外, 当小波函数和尺度函数或滤波器确定后, 分解和重构过程不需估算参数, 也不必进行前期分析和任何假定. 该方法具有一定的适应性和非参数化特征, 可应用于具有复杂结构的时间序列分析.

2 变点分析

变点分析是指利用统计学和小波分析等方法, 判断和检验时间序列中变点的存在、位置、个数及其跳跃度, 在测量数据处理中有着广泛的应用, 如GFS跳周识别、变形数据分析以及粗差探测[6]等. 变点的识别可分为两个方面: 一是信号的断点和尖点的识别; 另一个是分布变点的识别[7], 如均值变点、概率变点和模型变点. 下面仅给出本文将讨论的均值变点的离散模型.

均值变点离散模型的提法是: 对于时间序列

$$X_i = \mu_i + e_i, \quad i = 1, 2, \cdots, n, \tag{1}$$

如果存在 $\mu_1 = \mu_2 = \cdots = \mu_{m_1-1} = b_1, \mu_{m_1} = \mu_{m_1+1} = \cdots = \mu_{m_2-1} = b_2, \cdots, \mu_{m_q} = \mu_{m_q+1} = \cdots = \mu_n = b_{q+1}$, $1 < m_1 < m_2 < \cdots < m_q \leq n$, 随机误差项 $e_i (1 \leq i \leq n)$ 的方差和期望都为0. 如果 $b_j \neq b_{j+1}$, 则 m_j 是一个变点. 需要指出的是, 这里假定模型均值发生变化, 而方差不变.

3 小波包方法

在正交小波分解过程中, 一般的方法是将低频系数分解成两部分, 分开后获得一个新的低频系数和一个高频系数向量, 两个连续低频系数中间丢失的信息被高频系数获得; 然后, 将新的低频系数向量继续分解成两部分, 而高频系数不会被分解. 在小波包分解中, 每一个高频系数向量也象低频部分一样, 被分解成两部分. 因此, 它提供了更加丰富的分析方法. 在一维情况下, 产生一个完整的二叉树; 二维情况下, 产生一个完整的四叉树[8, 9].

1992 年Mallat 对小波分析在变点检测中的理论与应用进行了较为完整的讨论[10], 随后有关小波分析应用于变点检测的文章不断出现[11, 12], 引起了人们广泛的注意. 而小波包方法推广了小波分析理论, 实现了更加精细的结构, 更符合信号非平稳的变频带结构特征, 因而在检测信号奇异性等方面具有广泛的应用价值.

4 基于小波包分析的变点探测算法

给定时间序列$\{x_1, x_2, \cdots, x_n\}$, 假设变点发生的任何可能位置为 k, $1 < k < n$. 变点将时间序列分割成两部分, 这两部分的某些统计特征, 比如均值、方差等会有明显的不同.

当时间序列出现变点时, 变点之前的时间序列和变点之后的时间序列的幅频特性和相频特性会有明显改变. 因此, 变点之前的信号 S_A 与变点之后的信号 S_B 相比, 相同频带内信号的能量会存在较大的差别. 基于这一点, 本文设计了利用能量的变化来判断存在变点的变点探测算法. 基于小波包分析的变点探测算法不需要系统的模型结构, 直接利用各频率成分能量的变化来检测变点. 其基本思想是建立能量变化到变点的映射关系, 得到表征信号的特征向量; 然后采用一种度量来表示信号 S_A 与信号 S_B 的特征向量之间的距离. 对距离超过阈值的可能变点进行符号秩和检验[13], 给定显著性水平, 剔除错误变点, 最后得到变点最可能发生的位置.

算法 1 一种基于小波包分析的变点探测算法步骤如下:

Step 1: 给出初步估计的变点位置, 得到变点之前的信号 S_A 和变点之后的信号 S_B.

Step 2: 对信号 S_A 进行二层小波包分解, 分别提取第 2 层从低频到高频 4 个频率成分的信号特征. 其中 $(0, 0)$ 节点代表原始信号 S_A, (i, j) 节点代表小波包分解的第 i 层第 j 个节点的系数, $i = 1, 2$, $j = 1, 2, 3$.

Step 3: 重构小波包分解系数, 提取各频带范围的信号. 以 S_{20} 表示 X_{20} 的重构信号, S_{21} 表示 X_{21} 的重构信号, 其他依此类推. 这里只对第 2 层的所有节点进行分析, 则总信号 S_A 可表示为

$$S_A = S_{20} + S_{21} + S_{22} + S_{23}. \tag{2}$$

Step 4: 求各个频带信号的总能量. 设 $S_{2j} (j = 0, 1, 2, 3)$ 对应的能量为 $E_{2j} (j = 0, 1, 2, 3)$, 则有

$$E_{2j} = \int |S_{2j}(t)|^2 \mathrm{d}t = \sum_{k=1}^{n} |x_{jk}|^2, \tag{3}$$

$$j = 0, 1, 2, 3, \quad k = 0, 1, 2, \cdots, n.$$

其中 x_{jk} 表示重构信号 S_{2j} 的离散点幅值.

Step 5: 构造特征向量. 时间序列出现变点时, 会对各频带内信号的能量有较大的影响, 因此以能量为元素可以构造一个特征向量. 特征向量 T 构造如下:

$$T = [E_{20}, E_{21}, E_{22}, E_{23}].$$

能量较大时, $E_{2j} (j = 0, 1, 2, 3)$ 通常是一个较大的数值, 在数据分析上会带来一些不便. 为此, 可对特征向量 T 进行改进, 即对向量进行归一化处理, 令

$$E = \left(\sum_{j=0}^{3} |E_{2j}|^2 \right)^{1/2}, \quad (4)$$

$$T' = [E_{20}/E, E_{21}/E, E_{22}/E, E_{23}/E], \quad (5)$$

向量 T' 即为归一化后的向量

Step 6: 由 Step 4 和 Step 5 可得到 T_A; 对 S_B 重复 Step 2～ Step 5 便可得到 T_B.

Step 7: 定义距离度量函数

$$D = \| T_A - T_B \| = \left(\sum_{i=1}^{4} |T_{Ai} - T_{Bi}|^2 \right)^{1/2}. \quad (6)$$

Step 8: 给定一阈值 h, $D > h$ 时可认为该点就是可能存在的变点; 然后给定显著性水平 α, 对它们逐一进行符号秩和检验, 剔除错误变点, 最后得到变点最可能发生的位置

5　实验分析

在时间序列数据分析中, 变点检测具有特殊的意义, 因为从直观上难以把握数据中是否存在跳跃点, 而且变点通常意味着一些包含着许多有用信息和知识的事件: 气候变化和人类活动对水文循环系统的影响而导致的水文变异问题, 是时间序列变点探测的一个重要研究课题. 本文将上述算法应用于水文时间序列, 选用长江宜昌站 120 a (1882～ 2001 年) 的年径流时间序列数据, 实验目的是找到均值变点发生的时间位置. 实验中假设变点个数 $q = 1$, 得到了如下结果:

1) 年最小流量序列

变点可能发生位置的前半信号与后半信号特征向量之间的距离度量如图 2 所示. 给定阈值后, 对可能存在的变点逐一进行符号秩和检验, 在 5% 显著性水平下, 1921 年和 1934 年前后年最小流量序列的均值变化是显著的, 即长江宜昌站的年最小流量序列的均值最有可能在 1921 年或 1934 年发生变化. 事实上, 以 1921 年或 1934 年为变点, 长江宜昌站的年最小流量序列的均值分别下降了 297. 375 0 m^3/s 和 298. 622 9 m^3/s, 降幅分别为 7. 95 % 和 8. 06 %. 需要指出的是, 这里的 k 为 1921 年在文献[1] 中并没有被探测到

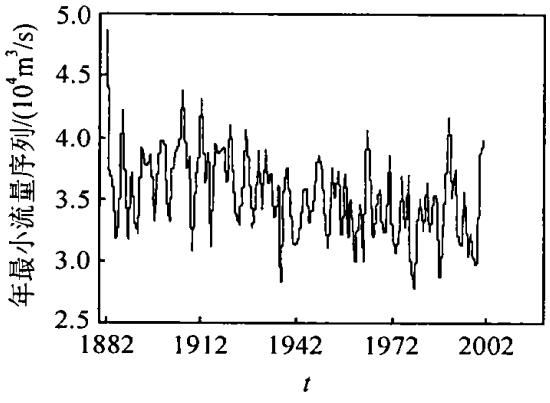

图 1　宜昌站年最小流量序列

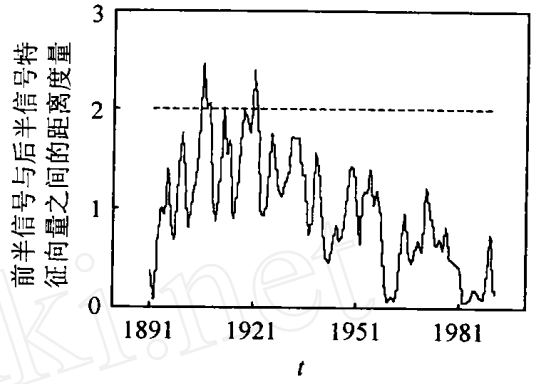

图 2　宜昌站年最小流量序列变点可能发生的年份和阈值

2) 年平均流量系列

变点可能发生位置的前半信号与后半信号特征向量之间的距离度量如图 4 所示. 给定阈值后, 对可能存在的变点逐一进行符号秩和检验, 在 5 % 显著性水平下, 1968 年前后年最小流量序列的均值变化是显著的, 即长江宜昌站的年最平均流量序列的均值最有可能在 1968 年发生变化. 事实上, 以 1968 年为变点, 长江宜昌站的年平均流量序列的均值下降了 297. 375 0 m^3/s, 降幅为 6. 14 %.

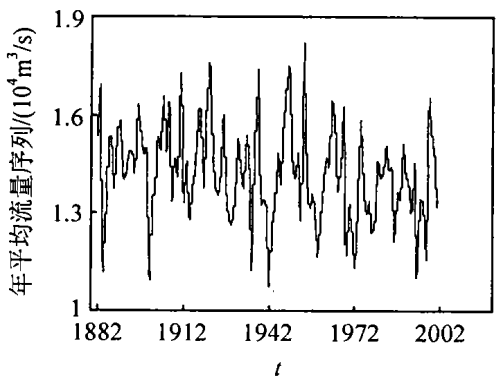

图 3　宜昌站年平均流量序列

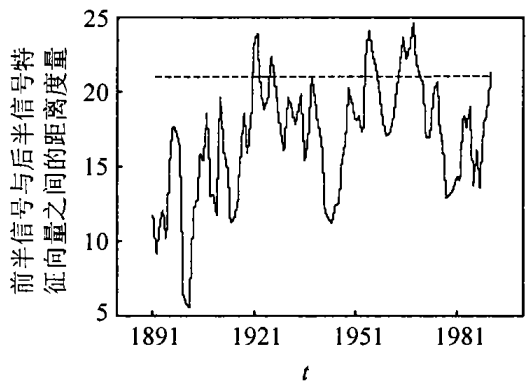

图 4　宜昌站年平均流量序列变点可能发生的年份和阈值

应用小波包变点探测算法研究气候变化和人类活动影响而发生的水文变异现象, 探索水文时间序列在时空分布规律上发生的变异, 对水文风险评价、水文统计计算具有理论意义, 对洪水预报调度、防洪减灾决策制定具有实践价值

6 结 语

本文建立了用于时间序列分析的小波包变点探测算法. 通过计算变点可能发生位置的前半信号与后半信号特征向量之间的距离度量, 判断出变点最可能发生的时间位置, 将该算法应用于长江宜昌站的年径流水文时间序列, 得到的结果与文献[1]等得到的结果是一致的. 另外, 本文建立的时间序列分析的小波包变点探测算法避免了文献[1]对时间序列所作的服从正态分布的假设, 只要求满足某种分布, 并不需知道具体的分布函数, 可根据具体的应用方便地调节小波包变点探测算法的参数. 算法具有一定的适应性, 可很好地应用于时间序列变点的探测. 但必须指出, 时间序列变点探测分析是一个涉及面广而复杂的领域, 对其全面分析尚有待进一步探讨.

参考文献(References)

[1] Xiong L H, Guo S L. Trend test and change-point detection of the annual discharge series of the Yangtze river at the Yichang hydrological station [J]. *Hydrological Sciences J*, 2004, 49(1): 99-112.

[2] Lavielle M, Lebarbier E. An application of MCMC methods for the multiple change-points problem [J]. *Signal Processing*, 2001, 81(1): 39-53.

[3] 朱颖元, 石凝. 福州市一百年来(1900～1999年)年降水量序列统计特性分析[J]. *水文*, 2002, 22(3): 22-25.
(Zhu Y Y, Shi N. Analysis on the statistical characteristics of Fuzhou annual precipitation time series during the past one hundred years [J]. *Hydrology*, 2002, 22(3): 22-25.)

[4] 丁晶, 邓育仁. *随机水文学*[M]. 成都: 成都科技大学出版社, 1988.

[5] Basseville M, Nikiforov I V. *Detection of abrupt changes: Theory and applications* [M]. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[6] 李朝奎, 徐望国, 邹峥嵘. 均值变点分析理论及其在桥梁健康监测中的应用[J]. *中国公路学报*, 2001, 14(4): 52-54.
(Li C K, Xu W G, Zou Z R. Mean value change-point theory and its application for bridge monitoring [J]. *China J of Highway and Transport*, 2001, 14(4): 52-54.)

[7] 李元. *时间序列中变点的小波分析及非线性小波估计* [M]. 北京: 中国统计出版社, 2001: 11-49.

[8] Meyer Y. *Wavelets: Algorithms and applications* [M]. Philadelphia: Society for Industrial and Applied Mathematics, 1993: 13-31, 101-105.

[9] Kaiser G. *A friendly guide to wavelets* [M]. Boston: Verlag, Springer, 1994: 44-45.

[10] Mallat S, Hwang W L. Singularity detection and processing with wavelets [J]. *IEEE Trans Information Theory*, 1992, 38(2): 617-643.

[11] Loschi R H, Cruz F R B. Applying the product partition model to the identification of multiple change points [J]. *Advances in Complex Systems*, 2002, 5(4): 371-387.

[12] Anestis Antoniadis, Irene Gijbels. Detecting abrupt changes by wavelet methods [J]. *Nonparametric Statistics*, 2002, 14(1-2): 7-29.

[13] 刘则毅. *科学计算技术与*Matlab [M]. 北京: 科学出版社, 2001: 172-178.

(上接第520页)

参考文献(References)

[1] Zlot R, Stentz A, Dias M B, et al. Multi-robot exploration controlled by a market economy [A]. *Proc of the IEEE Int Conf on Robotics and Automation (ICRA)* [C]. Washington: IEEE Press, 2002: 3016-3023.

[2] Burgard W, Moors M, Fox D, et al. Collaborative multi-robot exploration [A]. *IEEE Int Conf on Robotics and Automation (ICRA)* [C]. San Francisco: IEEE Press, 2000: 476-481.

[3] Mataric' M J, Sukhatme G S, Φstergaard E. Multi-robot task allocation in uncertain environments [J]. *Autonomous Robots*, 2003, 14(2): 255-263.

[4] Simmons R, Apfelbaum D, Burgard W, et al. Coordination for multi-robot exploration and mapping [A]. *Proc AAAI National Conf on Artificial Intelligence* [C]. Austin, 2000: 852-858.

[5] Yamauchi B. Frontier-based exploration using multiple robots [A]. *Proc of the Ind Int Conf on Autonomous Agents* [C]. Paul, 1998: 47-53.

[6] Burgard W, Fox D, Jans H, et al. Sonar-based mapping with mobile robots using EM [A]. *Proc of the Int Conf on Machine Learning* [C]. Bled, 1999: 67-76.

[7] Thrun S. Probabilistic algorithms in robotics [J]. *AI Magazine*, 2000, 21(4): 93-109.

[8] Thrun S. Learning occupancy grids with forward models [A]. *Proc of the Conf on Intelligent Robots and Systems (IROS' 2001)* [C]. Hawaii, 2001: 1676-1681.