有序聚类分析法的改进及其在水文序列突变点识别中的应用

袁 满,王文圣,叶濒璘

(四川大学水利水电学院,四川 成都 610065)

摘 要:有序聚类分析法是水文学中识别突变点的有效方法,但该法只考虑了同类之间的离差较小原则,忽略了类与类之间的离差较大原则。基于此,提出了改进的有序聚类分析法,改进法同时考虑了同类之间的离差较小和类类间的离差较大原则。将改进的有序聚类分析法应用于年平均流量序列突变点识别中,并与传统有序聚类分析法进行对比分析,研究结果表明,改进的有序聚类分析法原理明确,识别突变点更加有效。

关键词:突变点识别;改进的有序聚类分析法;水文序列;显著性检验

中图分类号:P333 文献标识码: A 文章编号:1000-0852(2017)05-0008-04

1 引言

水文序列是在一定自然和气候条件下形成的随机变化过程。在流域气候条件和下垫面条件相对稳定的情况下,水文序列具有良好的一致性。当气候条件急剧变化或人类活动发生显著变化时, 水文序列会产生趋势、跳跃、突变等形式的非周期成分,导致水文序列的一致性遭到破坏,其统计特性也会发生变化。利用这种不一致序列预估的未来水文情势可能会被大大歪曲。因此,识别并提取水文序列中的趋势、跳跃、突变成分,是水文序列变化特性分析的重要内容[1]。

目前,突变点识别方法大体有 4 种,一是成因分析法, 二是时序累计值相关曲线法, 三是有序聚类分析法,四是 Man-Kendall 法。有序聚类分析法简易直观,计算结果较精确,在突变点识别中应用广泛。王璨等运用有序聚类分析法识别窟野河洪水序列变异点[2]。王国庆等利用有序聚类分析法推算出了人类活动对无定河水文序列的显著干扰点[3]。刘茜等运用有序聚类分析法计算了长江大通站和宜昌站水沙变化的突变性,同时引入了二级突变点的概念[4]。陈远中等提出了用拟合的趋势线代替序列平均值的改进有序聚类分析法并进行了成功应用[5]。

有序聚类分析法识别突变点简单有效,但现行做法只考虑了同类之间的离差较小原则,忽略了类与类之间的离差较大原则,因此,在水文序列突变点识别过程中可能产生遗失或偏差。为此,本文提出了改进的有序聚类分析法, 改进法同时考虑了同类之间的离差较小及类类间的离差较大的原则。将改进的有序聚类分析法应用于年平均流量序列的突变点识别中,并与现行有序聚类分析法进行对比分析,研究结果表明,改进的有序聚类分析法识别突变点更加科学和有效。

2 有序聚类分析法及其改进

2.1 有序聚类分析法

有序聚类分析法推估突变点的实质是寻求最优分割点,其基本原则是使同类之间的离差平方和较小。

收稿日期:2016-09-23
基金项目:国家自然科学基金项目(51679155)
作者简介:袁满 (1973-),女,湖北黄冈人,硕士研究生,主要研究方向为水文水资源水环境系统分析。E-mail: 386011534@qq.com

设有水文序列 $x_1,x_2,\dots,x_n$,假设可能的突变点为 τ $(2\leqslant\tau\leqslant n-1)$,则突变点前后的离差平方和分别为

$$V_\tau=\sum_{i=1}^{\tau}(x_i-\bar{x}_\tau)^2 \tag{1}$$

$$V_{n-\tau}=\sum_{i=\tau+1}^{n}(x_i-\bar{x}_{n-\tau})^2 \tag{2}$$

式中:$\bar{x}_\tau$ 和 $\bar{x}_{n-\tau}$ 分别为 τ 前后两个序列的均值,即前后两个序列的聚类中心。

现行有序聚类分析法常用的目标函数为

$$S=\min_{2\leq\tau\leq n-1}S_0(\tau)=\min_{2\leq\tau\leq n-1}(V_\tau+V_{n-\tau}) \tag{3}$$

式中:min 表示取极小值。当式(3)中 S 取极小值时,对应的 τ 为最优二分割点,可推断为突变点。

从式(3)可以看出,目标函数仅体现了同类之间的离差平方和较小原则, 没有反映类与类之间的离差较大的原则。因此,有序聚类分析法现有的目标函数不太科学,据此推断突变点可能产生遗失或偏差。

2.2　改进的有序聚类分析法

有序聚类分析法宜同时考虑同类之间的离差较小及类与类之间的离差较大的原则。为此,需对目标函数式(3)进行改进。

类与类之间的离差可以表示为

$$d=|\bar{x}_\tau-\bar{x}_{n-\tau}| \tag{4}$$

式中:$\bar{x}_\tau$、$\bar{x}_{n-\tau}$ 意义同前。

为了保持与(4)式同一量纲和量级,同类间的离差用均方差 $\bar{\sigma}_\tau$、$\bar{\sigma}_{n-\tau}$ 表示,即

$$\sigma_\tau=\sqrt{\sum_{i=1}^{\tau}(x_i-\bar{x}_\tau)^2/\tau} \tag{5}$$

$$\sigma_{n-\tau}=\sqrt{\sum_{i=\tau+1}^{n}(x_i-\bar{x}_{n-\tau})^2/(n-\tau)} \tag{6}$$

基于同类之间的离差较小及类与类之间的离差较大原则,新的目标函数构造为

$$S'=\min_{2\leq\tau\leq n-1}S_1(\tau)=\min_{2\leq\tau\leq n-1}(\sigma_\tau+\sigma_{n-\tau}-d) \tag{7}$$

式中:min 意义同前。当式(7)中 S' 取极小值时,对应的 τ 为最优二分割点,可推断为突变点。

从式(7)可以看出,改进的有序聚类分析法同时考虑了同类之间的离差较小及类类间的离差较大的原则。

3　实例分析

3.1　在三皇庙站年平均流量序列突变点识别中的应用

收集了沱江流域三皇庙站 1941~2008 年平均流量序列,载于图 1 中。采用改进的有序聚类分析法对三皇庙站年平均流量序列突变点进行识别,由式(7)计算 $S_1(\tau)$,点绘 $S_1(\tau)$ 随 τ 的变化过程(图 2)。计算表明, 三皇庙站年平均流量序列突变点为 1961 年和 2001 年。

为对比分析,用传统的有序聚类分析法对三皇庙站年平均流量序列突变点进行识别,由式(3)计算出 $S_0(\tau)$,将其随 τ 的变化过程点绘于图 3 中。从图 3 可知,年平均流量序列突变点为 1961 年。可见,传统有序聚类分析法无法识别出 2001 年突变点。

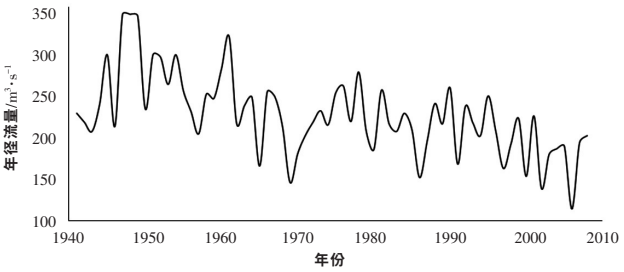

图 1　三皇庙站 1941~2008 年平均流量序列
Fig.1　The average annual flow seriesatthe Sanhuangmiao station during 1941–2008

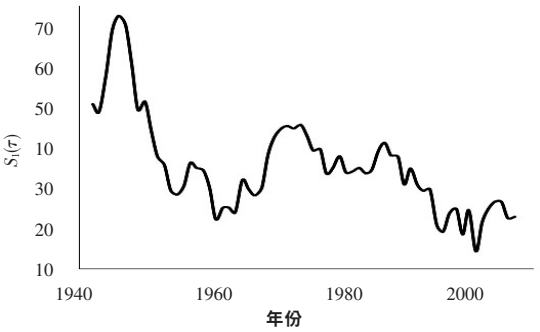

图 2 基于改进有序聚类分析法的三皇庙站年平均流量序列 $S_1(\tau)$变化过程
Fig.2 $S_1(\tau)$ change process based on the improved sequential clustering method for the annual flow series at the Sanhuangmiao stationmethod

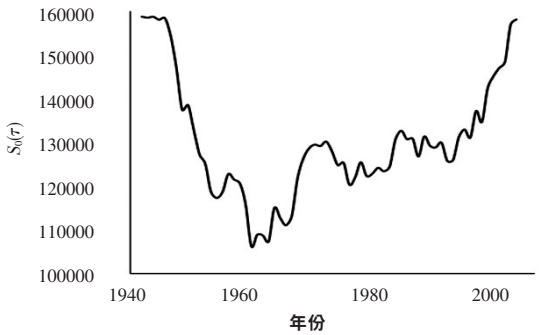

图 3 基于传统有序聚类法的三皇庙站年平均流量序列 $S_0(\tau)$变化过程
Fig.3 $S_0(\tau)$ change process based on the traditional sequential clustering for the annual flow series at the Sanhuangmiao station

3.2 在宜昌站年平均流量序列突变点分析中的应用

收集了宜昌水文站1890~2010年平均流量序列(图4)。采用改进的有序聚类分析法对宜昌站年平均流量序列突变点进行识别,计算$S_1(\tau)$并点绘其变化过程(图5)。由图5可以看出,突变点为1894年、1966年和1996年。同样,用传统有序聚类分析法对宜昌站年平均流量序列突变点进行识别,将$S_0(\tau)$变化过程绘制于图6中,突变点为1966年和1996年。由此看出,传统有序聚类分析法无法识别出1894年突变点。

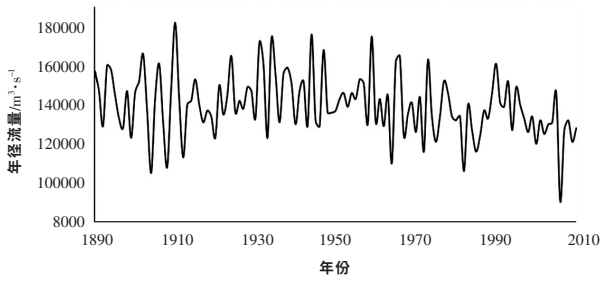

图4 宜昌站1890~2010年平均流量序列

Fig.4 The average annual flow series at theYichang station during 1890–2010

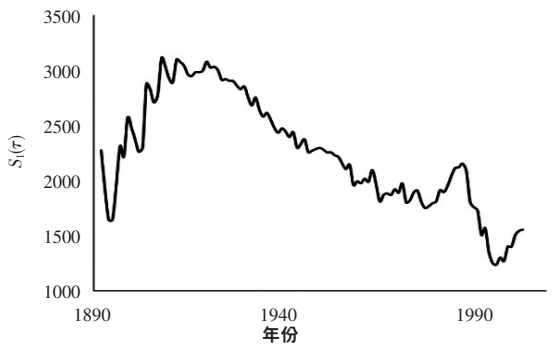

图5 基于改进有序聚类分析法的宜昌站年平均流量序列$S_1(\tau)$变化过程

Fig.5 $S_1(\tau)$ change process based on the improved sequential clustering method for the annual flow series at Yichang

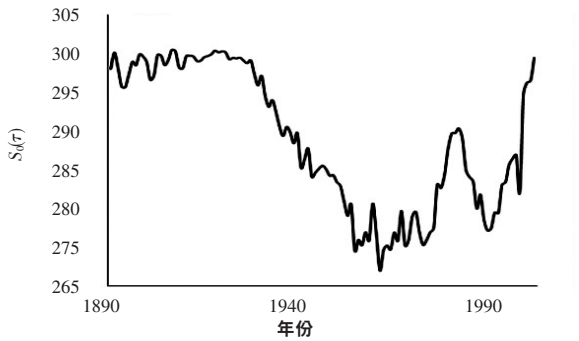

图6 基于传统有序聚类法的宜昌站年平均流量序列$S_0(\tau)$变化过程

Fig.6 $S_0(\tau)$change process based on the traditional sequential clustering station method for the annual flow series at the Yichang station

3.3 突变点检验

推断出突变点τ后,需要检验突变点是否显著。本文采取游程检验法对突变点进行显著性检验。

当n_1、n_2>20时,k趋于正态分布,则统计量

$$U=\frac{k-\left(1+\frac{2n_1n_2}{n}\right)}{\sqrt{\frac{2n_1n_2(2n_1n_2-n)}{n^2(n-1)}}}\sim N(0,1) \tag{8}$$

式中:n为样本系列的总数;n_1为τ之前系列的样本容量;n_2为τ之后系列的样本容量[1]。给定显著性水平α后,查出$U_{\alpha/2}$。当$|U|<U_{\alpha/2}$时,说明识别出的突变点不显著;反之,说明突变点前后来自不同的总体,突变成分显著。

检验结果见表1。从表1可以看出,改进的和传统的有序聚类分析法识别出的突变点1961年和1966年都是显著的。

表1 显著性检验结果表

Table1 The results of the significance test

水文站	突变点	k	U	$U_{\alpha/2}$	显著性分析
三皇庙站	1961年	19	-3.16	1.96	突变点显著
宜昌站	1966年	10	-9.28	1.96	突变点显著

n_1、n_2不满足同时大于20时的边缘突变点显著性检验可通过比较突变点前后序列离差大小进行判断。分别计算了各站各突变点前后序列的离差,结果列于表2中。从表2中可知,三皇庙站显著突变点1961年前后序列的离差为60.18,而突变点2001年前后序列的离差为61.98,故可推估突变点2001年是显著的;宜昌站显著突变点1966年前后序列的离差为1 080.16,而突变点1894年前后序列的离差为1 083.53,可见,突变点1894年也是显著的。

表2 离差及突变点显著性检验表

Table2 The deviates and significance test of the jump points

水文站	突变点	突变点前后序列间离差	显著性强度
三皇庙站	1961年	60.18	突变点显著
	2001年	61.98	突变点显著
宜昌站	1966年	1080.16	突变点显著
	1894年	1083.53	突变点显著
	1996年	1470.69	突变点显著

3.4 对比分析

改进有序聚类分析法和传统有序聚类分析法识别的各站突变点列于表 3 中。由表 3 知,改进有序聚类分析法识别的突变点更全面、准确,能识别出传统有序聚类分析法无法识别的边缘突变点。可见改进有序聚类分析法是科学、有效的。

表3 两种方法突变点识别结果表

Table3 The identificationresults ofthe jump points by two methods

方法 \ 结果	三皇庙站	宜昌站
改进有序聚类法	1961、2001 年	1894、1966、1996 年
传统有序聚类法	1961 年	1966、1996 年

4 结语

(1) 改进的有序聚类分析法是在传统有序聚类分析法的基础上,增加了类与类之间离差较大的原则,构建了新的目标函数,新方法是科学的。

(2) 三皇庙站和宜昌站年平均流量序列突变点识别分析表明, 改进的有序聚类分析法识别出的突变点更为精确,且可以识别出水文序列的边缘突变点,改进方法是适用的。

(3) 提出的改进有序聚类分析法目标函数只是一种方式,更多、更好的函数值得进一步探讨和尝试。

参考文献:

[1] 王文圣,金菊良,丁晶.随机水文学(第三版)[M].北京:中国水利水电出版社,2016.(WANG Wensheng, JIN Juliang, DING Jing. Stochastic Hydrology (The Third Edition)[M].Beijing: China WaterPower Press, 2016. (in Chinese))

[2] 王璨,周秀平,王文圣.窟野河洪水序列变异点综合诊断[J].水电能源科学,2012,30 (7):50-53.(WANG Can, ZHOU Xiuping,WANG Wensheng.Comprehensive diagnosis of change point of flood series in Kuye River[J].Water Resources and Power,2012,30(7):50-53.(in Chinese))

[3] 王国庆,贾西安,陈江南,等. 人类活动对水文序列的显著影响干扰点分析—以黄海中游无定河流域为例 [J]. 西北水资源与水工程,2001,12(3):14-16. (WANG Guoqing, JIA Xian, CHEN Jiangnan, et al.Analysis on the transition point of hydrological series impacted by human activities-a case study of Wudinghe basin in the middle reach of the Yellow River[J].Northwest Water Resources & Water Engineering,2001,12(3):14-16.(in Chinese))

[4] 刘茜,王延贵.江河水沙变化突变性与周期性分析方法及比较[J].水利水电科技进展,2015,35(2):17-23.(LIU Xi Qian, WANG Yangui. Comparison of analytical methods of runoff and sediment load mutation and periodical variation [J]. Advances in Science and Technology of Water Resources,2015,35(2):17-23.(in Chinese))

[5] 陈远中,陆宝宏,张育德,等.改进的有序聚类分析法提取时间序列转折点[J].水文,2011,31 (1):41-44. (CHEN Yuanzhong, LU Baohong, ZHANG Yude, et al. Improvement of sequential cluster analysis method for extracting turning point of time series[J]. Journal of China Hydrology, 2011,31(1):41-44. (in Chinese))

Improvement of Sequential Clustering Method and Its Application to Diagnose Jump Points of Hydrological Series

YUAN Man, WANG Wensheng, YE Binlin

(*College of Water Resource and Hydropower, Sichuan University, Chengdu* 610065, *China*)

Abstract: The sequential clustering method is an effective way to diagnosejump points in hydrologic research. However, the method only considers smaller deviations between the same classes, but ignores the larger deviations between different classes. For that reason, this paper proposedthe improved sequential clusteringmethod considering both smaller deviations between same classes and larger deviations between different classes. The improved sequential clustering methodhas been applied to identify jump points ofsome average annual flow seriesand comparedwith traditional sequential clustering method. The results show that theimproved sequential clusteringmethod is clear and effective to identify jump points.

Key words: diagnosis of jump points; improved sequential clustering method; hydrological series; significance test