

DOI:10.3876/j.issn.1000-1980.2020.04.003

基于柯西分布的水文序列异常值检测方法

高熠飞^{1,2}, 王建平², 李林峰²

(1. 河海大学水文水资源学院, 江苏 南京 210098;
2. 南瑞集团(国网电力科学研究院)有限公司, 江苏 南京 211000)

摘要: 针对基于正态分布的检测算法在窗口中出现偏差较大的极端异常值时,检测模型出现不稳定的现象,提出一种基于柯西分布的异常检测算法。通过试验发现,用滑动窗口内的中位数以及中位数绝对偏差来代替均值以及标准差,可以较好地消除滑动窗口内极端异常值对检测算法的影响。探讨了算法中滑动窗口以及置信度的设置,并通过实例对算法进行验证。结果表明,所提算法的检出率在窗口以及置信度设置恰当的情况下保持较高水平。与其他算法对比表明,所提算法对于处理局部波动剧烈的流量时间序列展现出较好的适用性以及鲁棒性。

关键词: 水文时间序列; 数据异常检测; 滑动窗口; 中位数; 中位数绝对偏差; 柯西分布

中图分类号: P33 **文献标志码:** A **文章编号:** 1000-1980(2020)04-0307-07

Outlier detection method of hydrological time series based on Cauchy distribution

GAO Yifei^{1,2}, WANG Jianping², LI Linfeng²

(1. College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China;
2. State Grid Electric Power Research Institute, NARI Group Corporation, Nanjing 211000, China)

Abstract: When the algorithm based on the normal distribution is applied in the sliding window, the detection model will yield an extreme outlier and exhibit unstable phenomenon. To address this issue, this study proposed an algorithm based on the Cauchy distribution for detecting outliers in hydrological time series. Test results indicate that using the median and median absolute deviation in the sliding window instead of the mean and standard deviation can well eliminate the influence of extreme outliers. The setup of the sliding window and the confidence in the algorithm were analyzed. Examples were used to verify this method and the results indicate that the detection rate of the proposed algorithm maintains a high level when the window and confidence settings are appropriate. The comparison with other algorithms shows that the proposed algorithm exhibits good applicability and robustness for dealing with time series with severe local fluctuations.

Key words: hydrological time series; outlier detection; sliding window; median value; median absolute deviation; Cauchy distribution

时间序列泛指随时间或空间有序变化的数据集合。水文行业中的流量、水位数据都是典型的时间序列数据。在水文时间序列中,存在 2 种情况:(a) 数据受到新机制的作用,如突降大暴雨或是闸门启闭造成数据点的跳变;(b) 人为差错或测量仪器故障导致错数或漏数,以至于数据测量收集产生误差。经常会产生与其他数据在某种相似度量上存在预定义或显著差异的数据^[1],一般将这种数据称为水文序列中的异常值。

异常值对时间序列的模型识别、参数估计、诊断检验乃至预测都有重要的影响^[2],当前水情测报与水库

基金项目: 国家电网公司总部科技项目(SGFJDK00DYJS1600091)

作者简介: 高熠飞(1995—),男,硕士研究生,主要从事水情测报与水库调度自动化研究。E-mail:1092308082@qq.com

引用本文: 高熠飞,王建平,李林峰. 基于柯西分布的水文序列异常值检测方法[J]. 河海大学学报(自然科学版),2020,48(4):307-313.

GAO Yifei, WANG Jianping, LI Linfeng. Outlier detection method of hydrological time series based on Cauchy distribution[J]. Journal of Hohai University(Natural Sciences),2020,48(4):307-313.

调度自动化系统在工作过程中需要大量调用并分析历史时间序列,异常值的存在直接影响系统相关决策的正确性。此外,水文序列中的异常值可能潜藏着更有价值和意义的水文信息和知识^[3]。因此,正确检测出时间序列中的异常值具有十分重要的现实意义,可以减少异常值对数据分析的干扰,提高预报等活动的准确性,另外,还可以帮助分析异常值背后蕴藏的信息。

目前国内外对于异常检测的研究可以分为以下几大类:

a. 基于预测模型的检测方法。基于预测的方法通过构建模型计算待检测点的预测值,基于预测值与实测值的偏离程度进行异常值判别。基于线性模型如自回归模型、滑动平均模型^[3-5]等的算法计算快捷,但线性模型适用于平稳数据,对于非平稳的水文数据预测准确率不高。非线性模型方法例如基于神经网络模型^[6]或是支持向量机模型^[7]的检测算法,对水文数据的预测较准确,但在实际工程应用中会面临阈值设置以及随着检测进行,阈值可能发生变化等问题,且非线性模型计算量大,耗费资源。

b. 基于距离的方法。通过设定距离函数,计算数据空间中数据点之间的距离,当某一个数据与其他数据距离较大时,认为该点数据异常^[8]。付培国等^[9]提出一种基于密度偏倚抽样的异常检测方法,首先利用基于密度偏倚的概率抽样对待测数据进行概率抽样,之后对抽样数据计算局部异常系数,局部异常系数越大的数据越有可能异常。该方法改善了传统基于距离的算法时间复杂度高的情况,处理大数据集时效果明显,但前期需要设置大量参数,相当依赖使用者的经验,且该方法适合对历史序列进行异常值挖掘,不适合作为实时异常值检测方法使用。

c. 基于聚类的方法。将数据集分成若干类,不属于任何类的或某个类中数据量特别少的点被判定为异常点^[10]。此类方法面临初始类采用随机选取带来的局部最优问题。左进等^[11]提出一种改进 k 均值聚类的异常检测算法,通过计算所有数据点的紧密性,在数据紧密的地方均匀选择 k 个初始中心,避免了陷入局部最优,对于异常点的检测效率具有明显的提升,但基于聚类的方法依赖初始类个数的设置以及数据集中存在异常点的假定,且基于聚类的方法本质在于分类,用来进行异常点的诊断不够高效。

本文的异常值检测算法是针对水情测报与水库调度自动化系统实际运行中对水文流量数据的质量需求而研究的,因此算法不能占用主系统太多的资源。复杂的算法例如神经网络以及支持向量机,对单个点的准确性可能更高,但随着数据量的增大,算法的时间复杂度较高,对资源的消耗较大,不适合水情测报与水库调度自动化系统实时调用。

实际采集得到的流量数据由于风浪、船舶、发电负荷调整、泄洪闸门启闭以及其他人类活动等影响,产生了较多的噪音数据,因此还需要算法具有较强的抗干扰能力。基于线性模型预测的算法计算简便,但流量数据局部波动以及水文数据非平稳的特性使得模型预测结果的准确性难以保证。

综上,本文考虑利用假设检验的统计学方法进行异常检测,对于数据的局部波动从统计学层面进行判断。统计学中常用的基于高斯分布的判断方法,其参数均值以及标准差在序列中存在极端异常值时会产生较大偏差,不够稳定^[12],因此考虑选用鲁棒性更好的中位数^[13]、中位数绝对偏差来替代均值和方差,利用待测点前向滑动窗口中的中位数以及中位数绝对偏差作为参数回归柯西分布,计算待测点的置信区间,根据待测点是否落在置信区间来判断待测点是否异常。该方法利用待测点历史数据的信息,可以自适应调整判断区间,并且算法受窗口中可能存在异常值的影响小,具有较好的鲁棒性,计算量小,适用于对大量水文数据的实时自动化检测。

1 算法介绍

1.1 相关定义

定义1 水文时间序列。由观测值及记录时间组成的有序集合 $X=\{(x_1,t_1),(x_2,t_2),\cdots,(x_n,t_n)\}$,令 $d_i=(x_i,t_i)$ 表示包含时刻以及观测值的水文对象,其中, t 表示当前水文对象的时间, x 表示对应时间的值,比如水位值。

定义2 滑动窗口。滑动窗口常分为单边以及双边。双边窗口更适合对待检测节点前驱和后继窗口数据都已知的历史数据进行挖掘,而单边窗口则是忽略待测点以后的数据,只利用前驱窗口中的数据信息进行分析,而本文的期望是在工程应用中实时判断数据异常与否,因此,单边窗口明显更适合。因此对于待检测点 d_i ,假定窗口宽度为 w 个时间点,滑动窗口即是待测点 d_i 的前 k 时刻观测值构成的集合,即

$$\eta_i = \{x_{i-k}, \cdots, x_{i-2}, x_{i-1}\}$$

(1)

定义 3 中位数绝对偏差。描述序列各点与序列中位数的偏离程度。假设式(1)中 η_i 序列的中位数为 m_i ,则计算出序列中每个点与中位数 m_i 的残差,生成残差序列 $\{|x_{i-k}-m_i|, \cdots, |x_{i-2}-m_i|, |x_{i-1}-m_i|\}$,计算该残差序列的中位数即原序列的中位数绝对偏差。

定义 4 水文时间序列异常。根据待定点 d_i 所对应的滑动窗口 η_i 中数据计算出窗口内数据的中位数 m_i 以及中位数绝对偏差 δ_i ,代入式(2),计算 t_i 时刻对应的累积分布函数值,记为 $C(x_i)$,若 $C(x_i)$ 在置信区间之外,则判定 x_i 为异常点。其中 $C(x_i)$ 计算公式如下:

$$C(x_i) = \frac{\arctan(x_i - m_i)}{\pi} + \frac{1}{2}$$

(2)

定义 5 基于中位数与中位数绝对偏差的柯西分布。柯西分布,又称柯西-洛伦兹分布,其概率密度函数为

$$f(x;x_0,\gamma) = \frac{1}{\pi} \left[\frac{\gamma}{(x - x_0)^2 + \gamma^2} \right]$$

(3)

式中: x_0 ——分布峰值位置的位置参数,在本文中用中位数进行估计; γ ——最大值 1/2 处 1/2 宽度的尺度参数,本文用中位数绝对偏差进行估计。

1.2 方法对比及选择

为了直观地看出基于高斯分布的算法在遇到极端异常值干扰时的情况,选取一段存在 6 个异常值的 2 h 流量数据进行试验。

图 1 是一段包含异常点的水文流量时间序列集 $X = \{d_1, d_2, \cdots, d_n\}$ 的流量-时间关系曲线,可以看到,流量数据中有较明显的 6 个偏差很大的异常点。

设定滑动窗口 η 宽度为 10 个时间点。对于每一时刻的待测点 d_i 都有其对应的滑动窗口 η_i 。对于每个待测点对应的滑动窗口内的数据进行分析,分别绘制其 Q-Q 图 (quantile-quantile plot),其横轴为水文流量值的分位值,纵轴为标准正态分布的分位值。若样本数据近似于正态分布,Q-Q 图上的点会排列在直线 $y = \sigma x + \mu$ 附近,其中 σ 表示该组数据的方差, μ 表示均值。点的散布与直线越接近,就越接近正态分布^[14]。

虽然根据随机选取部分滑动窗口内数据的 Q-Q 图实例,可以看出,窗口宽度较小时,窗口内数据可基本看成符合正态分布。

分别计算窗口内数据的平均值、中位数、标准差以及中位数绝对偏差,得到序列 $\{x_{\text{mean}}\}$ 、 $\{x_{\text{median}}\}$ 、 $\{x_{\text{SD}}\}$ 、 $\{x_{\text{MAD}}\}$ 。其中,均值序列与中位数序列的对比见图 2,标准差序列与中位数绝对偏差序列的对比见图 3。

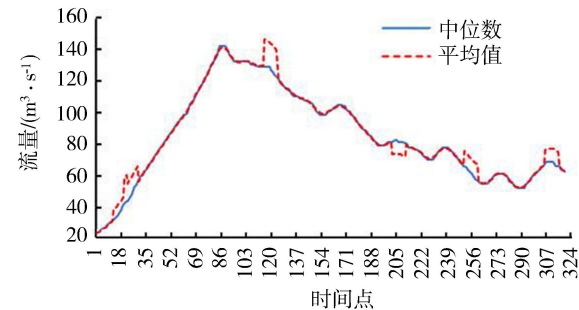


图 2 滑动窗口内数据的均值与中位数对比
Fig.2 Comparison of mean and median data in sliding windows

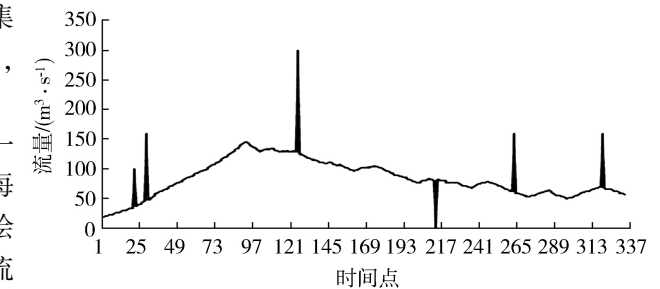


图 1 包含异常点的流量数据
Fig.1 Hydrological flow data containing outliers

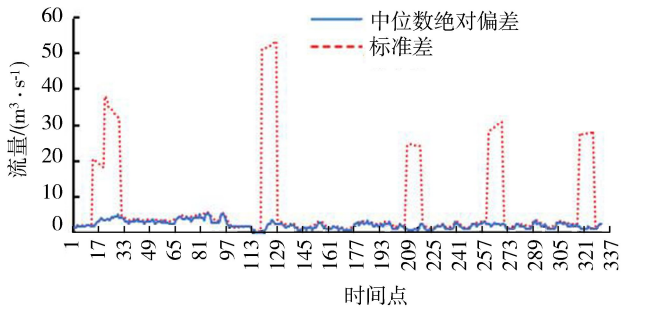


图 3 滑动窗口内数据的标准差与中位数绝对偏差对比
Fig.3 Comparison of standard deviation and median absolute deviation of data in sliding windows

由图2、图3可以明显看出,当滑动窗口内包含异常值时,均值与标准差均会发生窗口宽度的跳变,然而同时刻的中位数以及中位数绝对偏差却几乎不受窗口中异常值的影响,显得很稳定。

基于正态分布的检测算法在窗口中出现异常值,尤其是偏差较大的异常值时,其参数均值、标准差会受到很大的影响,极易造成误判,鲁棒性差。所以每次检测到异常值后都需要及时修正来减少异常值的影响,然后利用修正值进行下一个时间点的检测。修正一般分为人工修正或是自动修正^[3]。基于人工的修正对操作人员的专业知识要求较高,且实时性难以保证。自动修正一般是基于预测,例如通过前向 k -近邻窗口内的数据得出预测值^[3],以此代替异常值,但以水情测报与水库调度自动化系统中最常用的水库出库流量数据为例,数据局部波动很大,这种情况下预测模型的准确性难以保证,因此对后续判别的准确性也会带来影响。因此,期望能够在检测到异常值的情况下,即使异常值偏差很大,也可以在不对异常值做任何操作的情况下继续准确地进行下一个点的检测。

观察到中位数以及中位数绝对偏差受到窗口中极端异常值的影响几乎可以忽略,有很好的主体数据(或称大部分数据)的特性^[15]。

利用中位数判别异常值的方法早有一定的研究^[16],主要是直接考虑中位数与待测数据之间的差异来判断是否异常。但此种方法忽略了窗口中数据的局部波动情况,会使一些在合理波动内的数值被误判成异常值。因此,考虑在原方法的基础上加入反映数据变化尺度的参数来进行综合判断。本文在中位数的基础上加入中位数绝对偏差来进行综合判断。

鉴于柯西分布与正态分布的相似性,窗口内数据也可近似看作符合柯西分布。有研究证明,用样本的中位数来对柯西分布进行参数估计是一种相当优良的估计^[17]。有学者使用中位数以及中位数绝对偏差作为参数回归柯西分布并进行了网络流量数据的异常检测,取得了很好的效果。因此本文利用滑动窗口内数据的中位数以及中位数绝对偏差来回归柯西分布进行水文序列的异常值检测,这样既能考虑实际水文流量数据的局部波动情况,还大幅度减小了检测出的异常数据对后续检测的影响。

1.3 算法

1.3.1 算法步骤

本文提出的基于柯西分布的流量时间序列异常值检测方法的核心思想是:读取水文流量时间序列;设置滑动窗口的宽度为 N ;将 d_i 的前 N 时刻的数据的观测值 $\{x_{i-N}, x_{i-N+1}, \dots, x_{i-1}\}$ 放入窗口中;计算出前向窗口内数据的中位数、中位数绝对偏差;在预设置信度 p 的前提下将待检测点观测值 x_i 代入基于柯西分布的异常值判断函数 $E(x)$ 进行判断,对数据状态进行标记(正常或是异常);将窗口向后滑动一个时间点并进行下一个数据的判断。

1.3.2 标准化函数

柯西分布的自变量范围是 $(-\infty, +\infty)$,然而流量等水文数据一般都是正值,这样即使当前点相对于窗口中的数据发生陡降,只要不是负值,那么计算出的累积分布函数值所能达到的理论最小也只能是 $C(0)$,这显然会造成一定的漏检,因此考虑将待测点映射到 $(-\infty, +\infty)$ 范围内再进行判断。将数据变换到无穷区间一般利用三角函数 $\tan \frac{\pi}{2}$ 来进行处理。因此定义标准化函数 $N(x)$:

$$N(x) = \begin{cases} x & x \geq m \\ \frac{2m \tan\left(\frac{x-k}{k} \frac{\pi}{2}\right)}{\pi} + m & x < m \end{cases} \tag{4}$$

式中: m ——待测点对应滑动窗口内数据的中位数。

1.3.3 异常值判断函数

假定当前数据 d_i 的观测值 x_i 、对应的滑动窗口 η_i 及其中位数 $x_{\text{median}i}$ 、中位数绝对偏差 $x_{\text{MAD}i}$ 均已得到。 d_i 与滑动窗口 η_i 内的数据可近似认为同分布。

根据式(4)首先对待测数据用标准化函数 $N(x)$ 进行处理,设置信度为 p ,则异常值判断函数 $E(x)$ 定义如下(其中 $E(x)=1$ 代表通过检测, $E(x)=0$ 代表数据存在异常):

$$E(x)=\begin{cases} 0 & C(N(x))<\frac{1-p}{2} \text{ 或 } C(N(x))>1-\frac{1-p}{2} \\ 1 & \frac{1-p}{2}\leq C(N(x))\leq 1-\frac{1-p}{2} \end{cases} \tag{5}$$

1.3.4 参数选择

在读取数据后,对 d_i 进行异常检测的第一步就是确定滑动窗口 η_i 的宽度 N 。若窗口宽度过宽,则窗口内的数据不一定符合柯西分布的假定。又因为基于柯西分布的方法依赖窗口中数据的中位数以及中位数绝对偏差,若窗口宽度太窄,则窗口内数据的代表性不够,会导致判断出现误差。

判断函数 $E(x)$ 中,置信度 p 的影响很大。 p 太大,会出现漏检现象;反之 p 若是一味求小,则会出现误检测,对于水情测报与水库调度自动化系统的正常工作会造成很大的影响。

定义检出率 α_{DR} 、误检率 α_{FR} 。分别表示正确检测出的异常值占有所有异常值的比例、被误检测为异常值的正常值占有所有正常值的比例。

对于待检测的水文序列,检测前要确定 2 个参数,即选择滑动窗口的宽度 N 以及置信度 p 。窗口宽度太小,窗口内数据缺乏代表性。但窗口宽度太大,窗口内数据就无法服从柯西分布的假设。 p 越大,判断函数允许范围越小,但可能会造成误检; p 越小,发生误检的概率越低,但发生漏检的概率也随之提升。因此需要从检出率以及误检率来综合考虑。首先选取一段人工标记过异常值的历史序列,对算法的参数选择实际是选择使得检出率与误检率的比值 $\alpha_{\text{DR}}/\alpha_{\text{FR}}$ 最大的一组 N 、 p 组合。因此从 $(N, p) = (1, 75\%)$ 开始,首先固定 p ,考虑 N ,设置增量为 1,计算每个 $(N, 75\%)$ 对应的模型在历史数据集上的 $\alpha_{\text{DR}}/\alpha_{\text{FR}}$ 值,当比值连续出现降低时停止增加 N 。接着设置 p 的增量为 1%,计算每一个 p 对应每个 N 模型的 $\alpha_{\text{DR}}/\alpha_{\text{FR}}$ 值直到 $p = 100\%$,取其中比值最大的一组 N 、 p 作为后续检测的参数。由于算法检测到异常值后无需进行修改数据的操作即可继续进行检测,单次计算量小,且滑动窗口内的数据较多时,窗口内数据不再服从柯西分布,检出率开始下降,因此参数选择使用穷举法能在实现参数自动化设置的同时不会带来很大的算法时间复杂度。

2 实例分析

2.1 数据准备

选择三峡出库流量数据来进行验证。采用 2012-05-18 至 2012-07-23 的实测小时流量数据。为了更直观地显示,将具体时间(日期)用时间点表示。

从图 4 可以看出,流量数据的局部波动剧烈,异常值信息可能隐藏在波动中。

2.2 结果分析

为了验证算法的有效性以及准确性,将试验结果分为 4 类^[18]:(a) S_{TP} :异常值被定为异常;(b) S_{FN} :异常值被定为正常;(c) S_{FP} :正常值被定为异常;(d) S_{TN} :正常值被定为正常。可以看出,最理想的情况就是所有数据都被归为 S_{TP} 以及 S_{TN} ,

即没有发生误检、漏检。根据定义,检出率 $\alpha_{\text{DR}} = \frac{S_{\text{TP}}}{S_{\text{TP}} + S_{\text{FN}}} \times 100\%$,误检率 $\alpha_{\text{FR}} = \frac{S_{\text{FP}}}{S_{\text{FP}} + S_{\text{TN}}} \times 100\%$, α_{DR} 越大、 α_{FR} 越小则证明算法检测效果越好。

不同参数下算法的检测结果如表 1 所示。
评估结果表明,本文的算法能够对流量数据中的异常点进行较准确的判别,尽管出库流量数据局部波动剧烈,但在窗口宽度以及置信度选取合理的情况下,检出率以及误检率均比较理想,说明算法具有可行性。并且发现,当检测到偏差较大的异常值时,算法仍然能

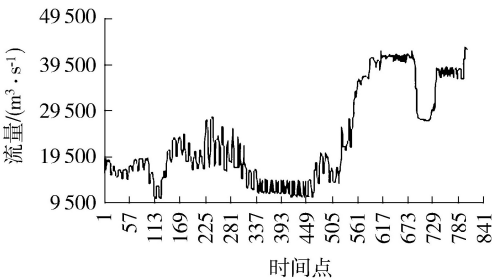


图 4 三峡原始小时出库流量数据集
Fig.4 Originally and hourly outbound flow data set of Three Gorges

Table 1 Test results under different parameters						
(N, p)	S _{TP}	S _{TN}	S _{FP}	S _{FN}	α _{DR}	α _{FR}
(15, 95%)	26	755	14	4	86.7	1.8
(14, 95%)	25	753	16	5	83.3	2.1
(15, 96%)	20	760	9	11	66.7	1.2

对接下来的数据进行正确检测,证明本算法具有较好的鲁棒性,可以保证实际工程应用中检测系统的平稳运行。同时本算法计算量小,随着数据量的增大,算法时间复杂度线性增加,面对海量数据也能进行快速检测。因此,本算法很适合水情测报与水库调度自动化系统中实时快速检测的应用场景需求。

图5是窗口宽度为15个时间点、置信度为95%时本算法的异常检测结果,大部分点被判定为正常,也有部分点在置信区间以外(根据本算法判断函数,即判定为异常点)。

本算法以前向窗口中的中位数以及中位数绝对偏差作为判断参数,无法准确描述窗口中数据的趋势特征,虽然使用滑动窗口一定程度上减弱了趋势项的影响,但趋势的存在还是造成了一些漏检、误检。

2.3 算法对比与分析

为了分析本文所提算法相较于传统的基于正态分布的算法在水文数据检测下的优势,将本文算法与常用的基于自回归模型的算法^[4]以及基于滑动窗口预测^[3]的算法进行对比,在本文数据集上的检测效果如表2所示。

基于自回归模型的算法在拟合回归时采用相邻几个时间点的流量数据,由于流量数据局部波动较大,模型拟合的准确性难以保证,在本文数据集上的试验也证实了这一现象,且该算法基于高斯分布判断异常,异常点的出现会使临近时间点的均值方差发生偏差,影响到对异常值后续时刻的数据检测,因此该算法在本文数据集检出率较低且出现了较多的误检。基于滑动窗口预测的算法检测到异常值后要将异常值进行修正,并且利用修正值继续进行异常值检测,根据试验发现,线性模型对这种剧烈波动的数据预测适用性较差,导致检测出现了误判,发生误判后,模型又使用预测值错误地修改了误判点的数据,导致后续的检测连续出现误检,在本文数据集的检测效果很不理想。

最后绘制这3种算法在本文数据集上的操作特性曲线^[19]。判断某个算法的优劣主要是看曲线下方面积的大小,面积越大检测效果越好。本文算法和其他算法在本文数据集上检测效果的操作特性曲线对比如图6所示。

由图6可以看出,本文所提算法,在局部波动较强烈的流量数据集上的检测效果明显优于另外2种算法,且更加稳定,表明该算法可以应对这种场景下水情测报与水库调度自动化系统对异常值检测的要求。

3 结 语

针对实测水文序列中数据存在异常跳变点的问题,基于滑动窗口的异常检测算法,提出了基于柯西分布的方法来检测流量时间序列中的点异常。采用中位数以及中位数绝对偏差来代替均值以及标准差进行回归。即使数据局部波动剧烈或是滑动窗口中出现极端异常值,模型都保持稳定,且检测到异常值后无需对异常值进行处理即可继续对下一个值进行检测。采用三峡水库小时出库流量数据进行试验,并通过不同算法的检测结果进行对比,验证了本文算法在这种数据条件下的可行性以及鲁棒性。未来需要考虑对滑动窗口内的数据进行趋势影响消除,降低水文数据的趋势对算法准确性的影响。

参考文献:

[1] 严宏,杨波,杨红雨. 基于异方差高斯过程的时间序列数据离群点检测[J]. 计算机应用,2018,38(5):1346-1352. (YAN

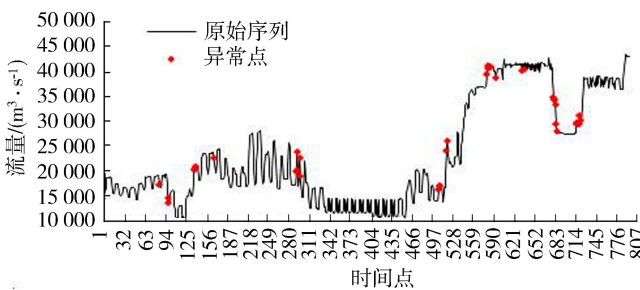


图5 (N, p) = (15, 95%) 条件下本算法的检测结果
Fig.5 Results of the algorithm when (N, p) = (15, 95%)

表2 不同算法在本文数据集上的检测结果对比
Table 2 Comparison of detection results of different algorithms in the data set of this study

算法	S_{TP}	S_{TN}	S_{FP}	S_{FN}	α_{DR}	α_{FR}
本文算法	26	755	14	4	86.7	1.8
自回归模型	20	725	44	10	66.7	5.7
滑动窗口预测	15	717	52	15	50.0	6.8

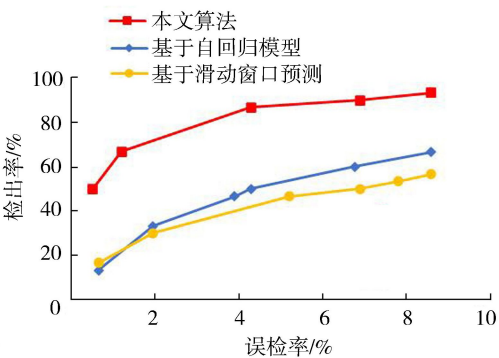


图6 不同方法在本文数据集下的检测结果 ROC 对比
Fig.6 Comparison of ROC results of different algorithms in the data set of this study

Hong, YANG Bo, YANG Hongyu. Outlier detection in time series data based on heteroscedastic Gaussian Processes[J]. Journal of Computer Applications, 2018,38(5):1346-1352. (in Chinese))

[2] REINSEL G E P B. Time series analysis: forecasting and control[M]. New Jersey: Wiley, 2008.

[3] 余宇峰,朱跃龙,万定生,等. 基于滑动窗口预测的水文时间序列异常检测[J]. 计算机应用, 2014,34(8): 2217-2220. (YU Yufeng, ZHU Yuelong, WAN Dingsheng, et al. Time series outlier detection based on sliding window prediction[J]. Journal of Computer Applications, 2014,34(8): 2217-2220. (in Chinese))

[4] 张玲,刘波. 基于残差统计的时间序列加性离群点检测算法研究[J]. 电子技术应用, 2015,41(9): 85-87. (ZHANG Ling, LIU Bo. Residuals statistics-based additive outlier detection algorithm for time series[J]. Application of Electronic Technique, 2015,41(9): 85-87. (in Chinese))

[5] 迟道才,王子凰,陈涛涛,等. ARIMA 和蒙特卡洛方法在预测降水量中的应用[J]. 沈阳农业大学学报, 2015,46(2): 187-191. (CHI Daocai, WANG Zihuang, CHEN Taotao, et al. Application of Monte Carlo and ARIMA models for precipitation forecast[J]. Journal of Shenyang Agricultural University, 2015,46(2): 187-191. (in Chinese))

[6] 张轩,张行南,江唯佳,等. 秦淮河流域东电站水位预报研究[J]. 水资源保护,2020,36(2):41-46. (ZHANG Xuan,ZHANG Xingnan,JIANG Weijia,et al. Study on water level forecast of Dongshan Station in Qinhuai River Basin[J]. Water Resources Protection,2020,36(2):41-46. (in Chinese))

[7] MOUNCE S, MOUNCE R, BOXALL J. Novelty detection for time series data analysis in water distribution systems using support vector machines[J]. Journal of Hydroinformatics, 2010,13(4): 672-686.

[8] 陈运文,吴飞,吴庐山,等. 基于异常检测的时间序列研究[J]. 计算机技术与发展, 2015,25(4): 166-170, 176. (CHEN Yunwen, WU Fei, WU Lushan, et al. Research on time series based on anomaly detection[J]. Computer Technology and Development, 2015,25(4): 166-170,176. (in Chinese))

[9] 付培国,胡晓惠. 基于密度偏倚抽样的局部距离异常检测方法[J]. 软件学报, 2017,28(10): 2625-2639. (FU Peiguo, HU Xiaohui. Anomaly detection algorithm based on the local distance of density-based sampling data[J]. Journal of Software, 2017,28(10): 2625-2639. (in Chinese))

[10] JÖRG S. LOF: identifying density-based local outliers[J]. Acm Sigmod Record, 2000,29(2):93-104.

[11] 左进,陈泽茂. 基于改进 K 均值聚类的异常检测算法[J]. 计算机科学, 2016,43(8): 258-261. (ZUO Jin, CHEN Zemao. Anomaly detection algorithm based on improved K-means clustering[J]. Computer Science, 2016,43(8): 258-261. (in Chinese))

[12] 唐芳. 算术平均数、众数、中位数的合理运用[J]. 中国统计,2008,56(3): 49-50. (TANG Fang. Arithmetic mean, mode, median rational use[J]. China Statistics, 2008,56(3): 49-50. (in Chinese))

[13] 周芯玉,廖菲. 利用中位数方法对风廓线雷达数据质量控制的研究[J]. 热带气象学报, 2015,31(6): 804-810. (ZHOU Xinyu, LIAO Fei. A study on the quality control of wind profiler data by using the median method[J]. Journal of Tropical Meteorology, 2015,31(6): 804-810. (in Chinese))

[14] 谭振江,朱冰. 基于 R 语言的学生成绩分析[J]. 智能计算机与应用, 2017,7(1): 76-78. (TAN Zhenjiang, ZHU Bing. Analysis of students' performances based on R language[J]. Intelligent Computer and Applications, 2017,7(1): 76-78. (in Chinese))

[15] 李泽安,陈建平,赵为华. 高维数据挖掘中基于中位数回归的特征提取新方法[J]. 计算机应用研究, 2013,30(2): 374-376. (LI Zean, CHEN Jianping, ZHAO Weihua. New feature extraction method in high-dimensional data mining based on median regression[J]. Application Research of Computers, 2013,30(2): 374-376. (in Chinese))

[16] BASU S, MECKESHEIMER M. Automatic outlier detection for time series: an application to sensor data[J]. Knowledge and Information Systems, 2007,11(2): 137-154.

[17] 郭彦. 对柯西分布性质的进一步讨论[J]. 淮阴工学院学报, 2005,12(5): 8-9. (GUO Yan. Further study on properties of cauchy distribution[J]. Journal of Huaiyin Institute of Technology, 2005, 12(5): 8-9. (in Chinese))

[18] 孙建树,娄渊胜,陈裕俊. 基于 ARIMA-SVR 的水文时间序列异常值检测[J]. 计算机与数字工程, 2018,46(2): 225-230. (SUN Jianshu, LOU Yuansheng, CHEN Yujun. Outlier detection of hydrological time series based on ARIMA-SVR model [J]. Computer and Digital Engineering,2018,46(2): 225-230. (in Chinese))

[19] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006,27(8): 861-874.