

一种水文时间序列异常模式检测方法研究

李云霞¹, 姚建国², 万定生¹, 赵 群¹

(1. 河海大学 计算机与信息学院, 江苏 南京 211100;

2. 淮河水利委员会水文局, 安徽 蚌埠 233001)

摘 要:时间序列数据是一类常见的多维复杂类型数据,它客观记录了观测系统随时间次序而变化的、在各观测时刻点的重要信息。时间序列数据具有海量性、高维性、复杂性等特点,直接对原始水文时间序列进行异常检测需要花费大量的时间,因此提出一种基于两阶段的水文时间序列异常检测方法。该方法通过分段线性表示方法对原始时间序列进行表示,提取子序列的斜率,极值差和均值三个特征值来表示原始时间序列。第一阶段在每个子序列为一个三元组的基础上用层次聚类算法对数据进行聚类,得到聚类结果。第二阶段基于聚类结果计算每一类的异常因子,根据异常因子判定异常模式。为验证该方法的有效性,采用龙门站的实测数据和人工合成数据进行实验检测,取得了较好的效果。

关键词:时间序列;分段线性表示;层次聚类;异常因子;异常模式

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2019)07-0159-05

doi:10.3969/j.issn.1673-629X.2019.07.032

An Anomaly Pattern Detection Method for Hydrological Time Series

LI Yun-xia¹, YAO Jian-guo², WAN Ding-sheng¹, ZHAO Qun¹

(1. School of Computer and Information, Hohai University, Nanjing 211100, China;

2. Bureau of Hydrology of Huaihe River Commission, Bengbu 233001, China)

Abstract: Time series data is a kind of common multi-dimensional complex data, which objectively records the important information of the observation system changing with time order at each observation point. Time series data is characterized by massiveness, high dimensionality, complexity and so on, and it takes a lot of time to conduct anomaly detection in the original hydrological time series. Therefore, we present an anomaly detection method for hydrological time series based on the two-stage. The original time series is represented by piecewise linear representation method, the slope, the extreme difference and the mean of the subsequence are extracted to express the original sequence. In the first stage, the hierarchical clustering algorithm is used to cluster the data on the basis subsequence, and the clustering result is obtained. In the second stage, based on clustering results the outlier factors of each type are calculated to detect anomaly patterns. In order to demonstrate the effectiveness of this method, the actual dataset of Longmen Station as well as artificial dataset are used for testing, and better results are obtained.

Key words: time series; piecewise linear representation; hierarchical clustering; outlier factor; anomaly pattern

0 引 言

时间序列数据是一种序列值按照时间发生先后顺序排列的特殊数据,在时间序列大量的数据中,有些极个别的子序列与其他子序列有着明显的不同,这些子序列称为异常模式。近年来,异常检测占据着非常重要的地位,受到了越来越多的关注。然而直接在原始时间序列上进行异常检测效率低、可靠性差,因此时间序列表示是必要的。常用的时间序列表示法主要有符

号化表示法、频域表示法^[1]、奇异值表示法^[2]、分段线性表示法^[3]。Obuchowski 在文献[1]中通过傅里叶变换,将时间序列从时域映射到频域,但傅里叶变换会平滑掉有重要特征的点,对非平稳的时间序列不适用。奇异值表示法将时间序列看成一个矩阵,利用 Karhune-Loeve 技术实现高维时间序列向低维数据的转化,缺点是奇异值表示法的时空复杂度高。分段线性表示法通过首尾相连的线段将时间序列分割成多个

收稿日期:2018-08-26

修回日期:2018-12-27

网络出版时间:2019-03-21

基金项目:国家重点研发计划(2018YFC0407900);公益性行业科研专项(201501022)

作者简介:李云霞(1995-),女,硕士研究生,研究方向为信息处理与信息系统;姚建国,高级工程师,从事水文水资源方面的研究与应用工作;万定生,教授,CCF 会员(08015S),研究方向为信息处理与信息系统。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20190321.0917.058.html>

子序列,具有良好的数据压缩和噪声过滤的作用,是使用较多的时间序列数据表示方法之一。

异常检测方法大致分为四类:基于统计的方法^[4]、基于距离的方法^[5]、基于密度的方法^[6-7]和基于聚类的方法^[8]等。

1 水文时间序列异常检测方法

根据水文时间序列具有的高维性、海量性等特征,提出一种两阶段的水文时间序列异常检测方法,该方法主要包括三个步骤:时间序列分段线性表示、层次聚类、异常模式检测。文献[9]也采用了两阶段异常检测方法,但其方法与文中方法相差很大,文中方法在异常检测前对时间序列进行分段线性表示,且两阶段检测方法中的聚类方法以及判定异常的方法都不同。

1.1 时间序列分段线性表示

给定长度为 n 的时间序列 $t = (t_1, t_2, \dots, t_n)$, 对时间序列进行等长分割, 然后用子序列的极值差、斜率、均值表示这段子序列^[10], 其中子序列的极值差是子序列中各特征值的最大值和最小值的差值, 利用极值差可以表示子序列的起伏程度大小, 可以判断这段序列是否平稳; 斜率是子序列的实际斜率, 可以很好地表现分段子序列的变化趋势; 均值是每段子序列的均值, 能代表子序列值的一般水平, 描述子序列值的集中程度。所以用这三个值能准确地表示子序列的特征。异常检测将用极值差、斜率、均值表示的子序列进行计算, 由于三个特征值的值域差别很大, 因此要对三者的值域进行规范化。

设 $t_1 = (t_{11}, t_{12}, \dots, t_{1n})$ 为一个子序列, 则利用式 1 将该组特征值规范到 0 至 1 之间。

$$\text{norm}(t_{1i}) = \frac{t_{1i} - t_{\min}}{t_{\max} - t_{\min}} \quad (1)$$

其中, t_{\max} 和 t_{\min} 分别表示子序列每个特征值的最大值和最小值。

算法 1: 时间序列分段线性表示算法。

输入: 时间序列 (t_1, t_2, \dots, t_n) , 子序列长度 l ;

输出: 规范化的子序列。

Step1: 将时间序列按长度 l 进行分段, 设 $m = n/l$ 。

$$t = \{t_1(t_1, t_2, \dots, t_l), t_2(t_{l+1}, t_{l+2}, \dots, t_{2l}), \dots, t_m(t_{n-l}, t_{n-l+1}, \dots, t_n)\}$$

Step2: 计算子序列的极值差、斜率、均值。

for($i = 1$ to m) {

$\text{avg}_i = \text{sum}(t_i) / \text{len}(t_i)$

$\text{xlvi} = (t_{in} - t_{i(n-1)}) / \text{len}(t_i)$

$\text{edi} = t_{\max} - t_{\min}$ }

Step3: 输出规范化的子序列。

$$t = \{(\text{ed}_1, \text{xlvi}_1, \text{avg}_1), (\text{ed}_2, \text{xlvi}_2, \text{avg}_2), \dots, (\text{ed}_m, \text{xlvi}_m, \text{avg}_m)\}$$

1.2 层次聚类

在对时间序列进行分段线性表示之后进行层次聚类, 层次聚类方法包括凝聚的层次聚类和分裂的层次聚类^[11-12]。凝聚的层次聚类是自底向上形成的, 这种方法最初将每个对象作为一个簇, 然后这些簇寻找距离最近的簇进行合并, 不断重复, 直至达到定义的簇的数目。

分裂的层次聚类是自顶向下形成的, 这种方法首先将所有对象看成一个簇, 然后逐步分为一个个小簇, 直到达到某个结束条件。层次聚类具有可以发现类的层次关系, 不需要预先制定聚类数, 距离和规则的相似度容易定义, 限制少等优点。文中采用凝聚的层次聚类方法进行实验。

流程如下:

输入: $t = \{(\text{ed}_1, \text{xlvi}_1, \text{avg}_1), (\text{ed}_2, \text{xlvi}_2, \text{avg}_2), \dots, (\text{ed}_m, \text{xlvi}_m, \text{avg}_m)\}$;

输出: 类簇 C_1, C_2, \dots, C_i 。

Step1: 将每个对象看作一个初始簇;

Repeat:

Step2: 计算两两类之间的距离, 找到距离最小的两个类簇 C_1 和 C_2 ;

Step3: 合并 C_1 和 C_2 为一个新类;

Until: 达到定义的簇的数目。

1.3 异常模式检测

异常检测是在类与类之间距离的基础上算出异常因子, 通过判断异常因子是否超过给定的阈值来判断模式是否异常, 因为异常因子 $\text{OF}(C_j)$ 度量了 C_j 与其他类之间的距离, 距离越大, 说明该类与整个数据集的差异程度越大, 从而说明该模式异常。其中, 在较短时间内某站流量数据突然暴涨或突然下降的事件认定为异常事件, 可用于验证检测结果是否准确。

算法 2: 异常模式检测。

输入: C_1, C_2, \dots, C_i, d ;

输出: 异常模式。

Step1: 计算类与类之间的距离^[13], 设 C_i 的质心为 $p = (x_p, y_p, z_p)$, C_j 的质心为 $q = (x_q, y_q, z_q)$ 。

$$\text{dist}(p, q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2 + (z_p - z_q)^2} \quad (2)$$

Step2: 计算异常因子。

$$\text{OF}(C_j) = \sum_{i=1, i \neq j}^n \frac{\|C_i\|}{\|D\|} \cdot \text{dist}(C_j, C_i) \quad (3)$$

Step3: 判定异常模式。

if($\text{OF}(C_j) \geq d$) {

输出 C_j }

2 实验及结果分析

2.1 实验结果

选取山西省龙门站 2002 年 1 月 1 日至 2016 年 12 月 31 日共 15 年的汛期小时流量数据作为实验数据,

共 7 892 条。原始时间序列如图 1 所示。

分别用文中提出的异常检测方法和相关性分析的检测方法^[14-15]对龙门站的数据进行检测,通过多次实验发现,相关性分析的方法当 $e_1=0.6$, $e_2=0.03$ 时检测结果最好,检测结果如表 1 所示。

比较表 1 可以看出,两种方法都检测出五个异常模式,但是其中有两个异常模式是不同的,为了证明检测的正确性,分别将检测出来的异常模式单独画图,验证其是否异常。图 2 中的 (a) 至 (f) 表示两种方法检测出来的异常模式。

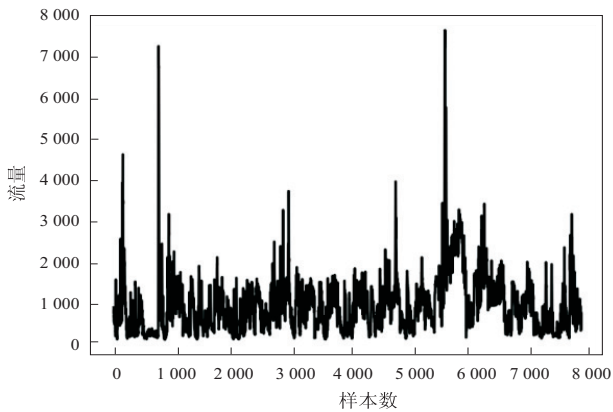


图 1 龙门站 2002–2016 年汛期流量数据时间序列

表 1 文中方法与相关性分析方法实验结果对照(实测数据)

文中方法				相关性分析方法		
类簇	OF	$\text{len}(C_i)$	异常时间段	e_1	e_2	异常时间段
$C(3)$	1.15	1	2003/07/29–2003/07/31	0.6	0.03	2003/07/29–2003/07/31
$C(6)$	1.46	1	2012/07/25–2012/07/28			2012/07/25–2012/07/28
$C(7)$	1.19	1	2012/07/28–2012/07/29			2012/07/28–2012/07/29
$C(1)$	0.60	2	2002/07/04–2002/07/05			2002/07/04–2002/07/05
			2010/09/19–2010/09/20			2016/06/03–2016/06/09

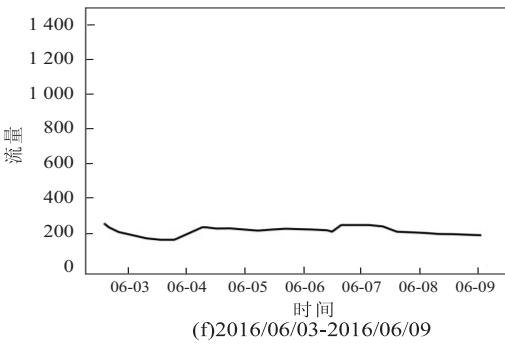
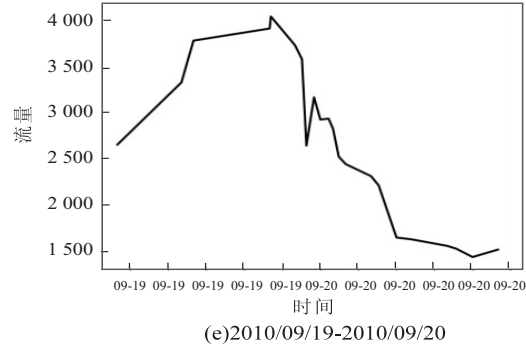
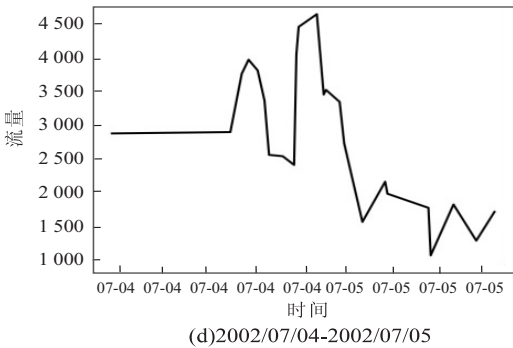
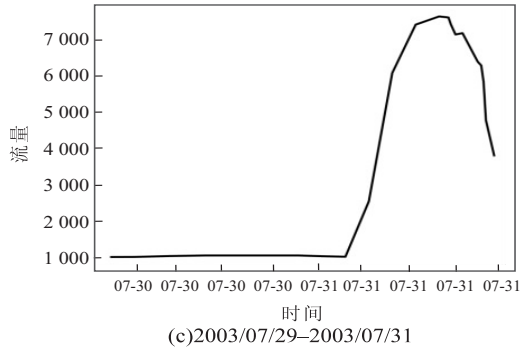
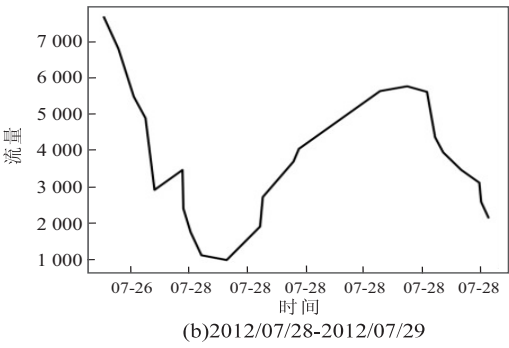
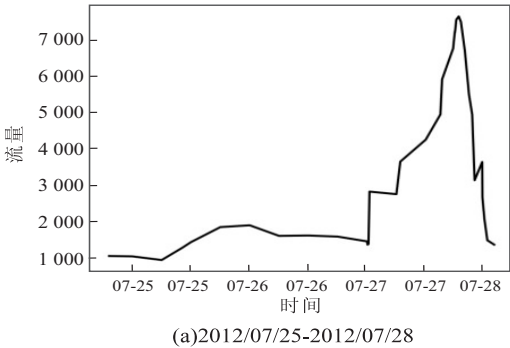


图 2 异常序列(1)

利用人工合成数据进行实验,其中人工合成数据是在原始数据的基础上增加了三个异常模式,增加的三个异常模式分别为 2005 年 6 月 13 日至 2005 年 6 月 21 日、2014 年 9 月 12 日至 2014 年 9 月 16 日、2009 年 7 月 18 日至 2009 年 7 月 28 日,人工合成数据时间序列如图 3 所示。

分别用文中的异常检测方法和相关性分析的检测方法进行实验,通过多次实验发现,相关性分析的方法当 $e_1=0.6$, $e_2=0.03$ 时检测结果最好,检测结果如表 2 所示。

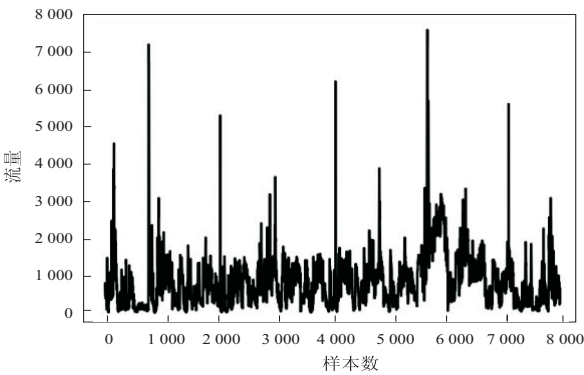
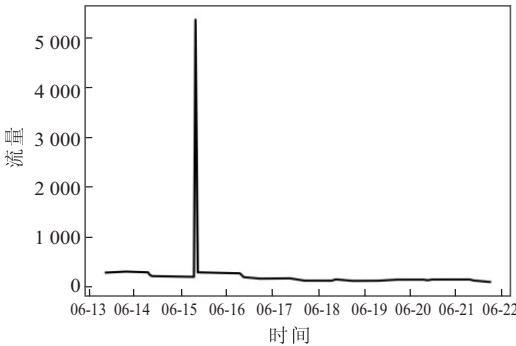


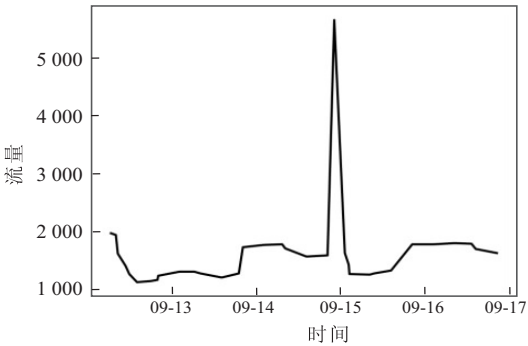
图 3 人工合成数据时间序列

表 2 文中方法与相关性分析方法实验结果对照(合成数据)

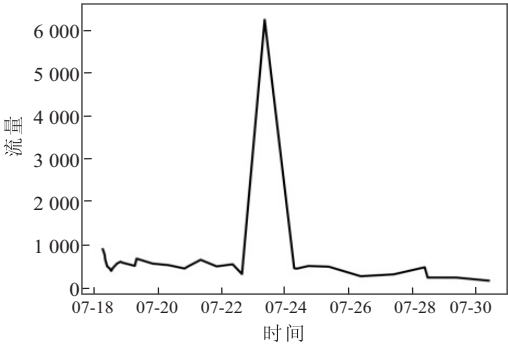
类簇	文中方法			相关性分析方法		
	OF	$\text{len}(C_i)$	异常时间段	e_1	e_2	异常时间段
$C(0)$	0.59	3	2002/07/04–2002/07/05	0.6	0.03	2016/06/03–2016/06/09
			2010/09/19–2010/09/20			2002/07/04–2002/07/05
			2014/09/12–2014/09/16			2009/07/18–2009/07/28
$C(2)$	1.15	1	2003/07/29–2003/07/31			2003/07/29–2003/07/31
$C(4)$	0.70	2	2005/06/13–2005/06/21			2012/07/25–2012/07/28
			2009/07/18–2009/07/28			2012/07/28–2012/07/29
$C(5)$	1.46	1	2012/07/25–2012/07/28			
$C(6)$	1.19	1	2012/07/28–2012/07/29			



(a)2005/06/13-2005/06/21



(b)2014/09/12-2014/09/16



(c)2009/07/18-2009/07/28

图 4 异常序列(2)

2.2 实验分析

通过实验表明,用实测数据进行实验时,文中方法检测出来的五个异常模式,从图 2(a) ~ (e) 可以看出该站的流量突然暴涨或者突然下降,这与该站的流量

模式不符合,故判定其为异常模式,但相关性分析的检测方法没有检测出 2010 年 9 月 19 日至 2010 年 9 月 20 日这段序列存在异常,从图 2(e) 可以看出,该段序列的流量快速上升又快速下降,属于异常模式;图 2

(f)表示 2016 年 6 月 3 日至 2016 年 6 月 9 日这段序列,在该序列中该站流量基本保持不变,不属于异常模式。用人工合成的数据进行实验时,文中方法检测出八个异常模式,从图 2(a)~(e)和图 4(a)~(c)可以看出这八个模式均属于异常模式,其中图 4(a)~(c)的异常模式与人工增加的异常模式吻合,相关性分析的方法只检测出六个异常模式,其中人工合成的三个异常模式,该方法只检测出一个,所以文中的异常检测方法要优于相关性分析的方法。

3 结束语

文中提出一种两阶段的水文时间序列异常检测方法。该方法通过分段线性表示、层次聚类、异常模式检测三个步骤来检测时间序列中存在的异常模式。为验证该算法的准确性,采用另一种相关性分析的异常检测方法进行对比,通过实验发现,文中方法能准确检测出异常模式,而且要优于相关性分析的检测方法。但文中方法利用阈值确定异常模式,使得该阈值对检测结果有一定影响,阈值的确定方式缺少灵活性,有待进一步改进。

参考文献:

[1] OBUCHOWSKI J, WYŁOMANŃSKA A, ZIMROZ R. The local maxima method for enhancement of time-frequency map[J]. Mechanical Systems & Signal Processing, 2014, 46(2): 389-405.

[2] 叶燕清. 多元时间序列数据挖掘相似性分析方法及应用研究[D]. 长沙:国防科学技术大学, 2015.

[3] 喻高瞻, 彭 宏, 胡劲松, 等. 时间序列数据的分段线性表示[J]. 计算机应用与软件, 2007, 24(12): 17-18.

[4] YAMANISHI K, TAKEUCHI J. Discovering outlier filtering rules from unlabeled data[C]//Proceedings of the 7th ACM SIGKDD international conference on knowledge discovery and data mining. [s. l.]: ACM, 2001: 389-394.

[5] KNORR E M, NG R T. Algorithms for mining distance-based outliers in large datasets [C]//International conference on very large data bases. [s. l.]: Morgan Kaufmann Publishers Inc., 1998: 392-403.

[6] 黄光球, 彭绪友, 靳 峰. 基于密度的异常挖掘方法研究与应用[J]. 微电子学与计算机, 2005, 22(3): 262-265.

[7] BREUNIG M M, KRIEGEL H, NG R T. LOF: identifying density-based local outliers [C]//ACM SIGMOD international conference on management of data. Dallas, Texas, USA: ACM, 2000: 93-104.

[8] 曹文平, 熊启军, 罗 颖, 等. 基于聚类的时间序列异常检测模型[J]. 金融科技时代, 2012(11): 100-101.

[9] 蒋盛益, 李庆华, 赵延喜. 一种两阶段异常检测方法[J]. 小型微型计算机系统, 2005, 26(7): 1237-1240.

[10] 刘雪梅, 王亚茹. 基于异常因子的时间序列异常模式检测[J]. 计算机技术与发展, 2018, 28(3): 93-96.

[11] 段明秀. 层次聚类算法的研究及应用[D]. 长沙: 中南大学, 2009.

[12] 张红梅, 丁 伟, 范艳峰. 一种改进的层次聚类算法在面包品质检验中的应用[J]. 微电子学与计算机, 2009, 26(7): 187-190.

[13] 詹艳艳, 徐荣聪. 时间序列异常模式的 k-均距异常因子检测[J]. 计算机工程与应用, 2009, 45(9): 141-145.

[14] 曹文平, 熊启军, 罗 颖, 等. 基于相关性分析的时间序列异常检测方法[J]. 信息系统工程, 2012(10): 131-132.

[15] 李海林. 基于动态弯曲的时间序列异步相关性分析[J]. 计算机应用研究, 2014, 31(7): 1976-1979.