

# 基于异常因子的时间序列异常模式检测

刘雪梅,王亚茹

(华北水利水电大学 信息工程学院,河南 郑州 450045)

**摘要:**时间序列中的异常模式能够提供大量有意义的信息,由于时间序列数据量大、含噪音、维度高,直接在原始时间序列数据中进行异常模式挖掘要花费大量的时间和空间代价。常用的时间序列分段线性表示法,易受阈值和分段数目的影响。对此,根据实际工程监测中时间序列的特征,将不限定分段数目与子序列长度的方法相结合,基于斜率及最大时间跨度,将原始时间序列分割成长度不同的子序列,提取子序列的极值差、斜率、均值等特征值,并映射到三维特征空间,在该特征空间中计算正常模式间的距离,以正常模式间距离为标准,求出各子序列的异常因子,检测异常模式。为验证该算法的有效性,采用南水北调工程安全监测中的实测数据和人工合成数据进行测试,取得了较好的效果。

**关键词:**时间序列;分段线性表示;异常模式;异常因子;子序列;特征空间

**中图分类号:**TP31

**文献标识码:**A

**文章编号:**1673-629X(2018)03-0093-04

doi:10.3969/j.issn.1673-629X.2018.03.019

## Anomaly Pattern Detection in Time Series Based on Outlier Factor

LIU Xue-mei, WANG Ya-ru

(School of Information Engineering, North China University of Water Resources and  
Electric Power, Zhengzhou 450045, China)

**Abstract:** Anomaly patterns of time series can provide a lot of meaningful information. Because of the large amount of data, noises and high dimension for time series, anomaly pattern mining in the original time series directly will take much time and space. The commonly used piecewise linear representation methods are vulnerable to the threshold and the number of segments. For this, based on the time series in engineering monitoring, we combine the method of not limiting the number of segments and the length of the subsequence and segment the time series based on slope and time span. Then the extreme difference, slope and mean value of these sections are extracted and transformed into the three-dimensional feature space, where the distance of the normal pattern is calculated as the standard to solve the outlier factors of each subsequence for detection of anomaly patterns. We demonstrate the effectiveness of the proposed algorithm through an application to an actual dataset from South-to-North Water Transfer Project as well as an artificial dataset, with better results.

**Key words:** time series; piecewise linear representation; anomaly pattern; outlier factor; subsequence; feature space

## 0 引言

南水北调工程是缓解我国北部地区水资源紧张,优化水资源配置的一项战略性基础设施工程。在工程安全监测中有一类数据是按照发生的时间顺序保存的,这类数据叫做时间序列。在时间序列大量的数据中,有些极少出现的子序列与其他子序列有显著的不同,使得人们怀疑它是由不同的机制产生的,这些子序列称为异常模式<sup>[1]</sup>。在工程安全中,异常模式往往更能够帮助人们认识事物。因此,从海量数据中挖掘出

异常模式,对保证南水北调工程的安全具有重要意义。

## 1 时间序列上进行异常模式挖掘

时间序列具有高维性、海量性、含有大量噪声等特征,直接在原始时间序列上进行异常模式挖掘要花费大量的时空代价,会影响算法的可靠性。

目前常用的时间序列表示法主要有频域表示法<sup>[2]</sup>、奇异值表示法<sup>[3]</sup>、分段线性表示法<sup>[4-5]</sup>、符号化表示法<sup>[6]</sup>。文献<sup>[7]</sup>中,通过离散傅里叶变换,将时间

收稿日期:2017-04-06

修回日期:2017-08-16

网络出版时间:2017-12-05

基金项目:国家科技重大专项(2014ZX03005001);河南省高校科技创新团队支持计划(13IRTSTHN023);郑州市科技创新团队支持计划(131PCXTD595)

作者简介:刘雪梅(1965-),女,博士,教授,研究方向为数据挖掘、计算机图形学、虚拟现实;王亚茹(1990-),女,硕士研究生,研究方向为数据挖掘。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20171205.0903.020.html>

序列从时域映射到频域,傅里叶变换会平滑掉具有重要特征的点,对非平稳的时间序列不适用。奇异值表示法的时空复杂度高。分段线性表示方法通过首尾相连的线段将时间序列分割成多个子序列,目前常用的主要有两种:一是限制分段数目。文献[8-9]中使用了分段聚集近似法(piecewise aggregate approximation),也称 PAA 算法。PAA 算法忽略了时间序列的特征值,出现了较大的拟合误差。第二种方法是通过限制分段误差将时间序列划分成长度不等的子序列,分段误差的阈值对分段的影响较大。而分段结果的好坏直接影响到异常检测的准确性。

通过以上分析得出,如何选择合适的分段数目是限制分段数目算法存在的问题,如何选择重要点及阈值是不限制分段数目方法的难点。结合两者的优缺点,提出了基于斜率及子序列的最大时间跨度,不限制分段数目进行时间序列分割,同时为了尽可能减少阈值对结果的影响,限制每段子序列的最大时间跨度,实现对分段数目最大值的限制,将不限制分段数目与限制分段数目相结合。

目前用于异常检测的方法可分为基于模型的检测方法<sup>[10]</sup>、基于聚类的检测方法<sup>[11]</sup>、基于异常点检测方法<sup>[12-13]</sup>、基于密度的检测方法<sup>[14-15]</sup>。基于异常点检测方法较为简单,但时间序列的高维性使该方法失效。基于密度的异常检测方法精度高,但时空复杂度高。基于聚类的方法对于发现频繁模式比较适用。基于模型的方法,建立模型和参数的估计存在一定的困难。

在时间序列分段线性表示的基础上,文中提取子序列的斜率、均值、极值差,将时间序列映射到该特征空间,每个子序列就对应到该特征空间中的一个点,用特征值构成的三元组表示各子序列在特征空间的位置,在此基础上计算各模式间的距离。通过一定的处理,得到正常模式间的距离,比较每个模式的距离与正常模式距离的比值,提取异常模式。

## 2 基于异常因子的时间序列异常模式探测算法

### 2.1 时间序列分段线性表示

定义 1 重要点<sup>[12]</sup>:给定时间序列  $T = (t_1, t_2, \dots, t_m)$ ,若  $t_i (1 \leq i \leq m)$  为极值点并满足以下条件之一,则称其为重要点。

- (1)  $t_i$  是时间序列的起点;
- (2)  $t_i$  是时间序列的终点;
- (3)  $(t_i - t_{i-1}) * (t_{i+1} - t_i) < 0$ 。

对于时间序列模式表示,只需保留引起模式变化的重要点,这样既能保留时间序列的形状特征,又能实

现大幅度的压缩。文中提出的分割算法,重点为引起斜率变化幅度较大及子序列达到最大时间跨度的点。分段线性表示方法就是用  $K$  条重要点相连的直线段来表示时间序列。由于是近似的表示时间序列,因此会平滑掉一些数据,使数据的管理更加高效。采用不限制分段数目与限制分段数目的方法相结合,基于斜率选择合适的分段点进行分割。

基于斜率的时间序列分割算法将相邻的两点作为一个最小分段,计算相邻两段斜率的差值与阈值进行比较,若小于阈值,则将两端合并,若大于阈值,则中间点为分割点。为了避免因阈值选择不当,平滑掉时间序列的主要特征,限制子序列的最长时间跨度,这也是文中的创新之处。算法 1 具体描述了基于斜率的时间序列分段线性表示方法。

算法 1:基于斜率的时间序列分段线性表示算法。

输入:(时间序列 array( $x_1, x_2, \dots, x_n$ ),  $d$ ),其中  $d$  为斜率误差阈值

输出:时间序列的分段线性表示

Step1:序列的第一个点加入重要点序列。

$s = (x_1, 1)$

Step2:分别计算以点  $x_i$  为端点的相邻两个线段的斜率。

$j=0; k=1; h=2$

for( $i=1$  to  $n$ )

$tg1[i] = (x_k - x_j) / (k - j)$

$tg2[i] = (x_h - x_k) / (h - k)$

Step3:判断点  $x_i$  是否为重要点,若是,则加入重要点集合  $s$ 。

if( $fabs(tg1[i]) - tg2[i] > d$ )

$j=i; k=i+1; h=i+2$

Then  $s = s + (x_i, i)$

Else  $k=i+1; h=i+2$

if( $k-j > D$ )// $D$  为子序列的最大时间跨度

$s = s + (x_i, i)$

Step4:最后一个点加入重要点序列。

$s = s + (x_n, n)$

Step5:输出分段线性表示的子序列。

$L(x) = \{ L(x_1, x_2), L(x_2, x_3), \dots, L(x_{n-1}, x_n) \}$

### 2.2 异常模式探测算法

设时间序列  $x_1 = (x_{11}, x_{12}, \dots, x_{1n})$  是时间序列  $x = (x_1, x_2, \dots, x_n)$  的子序列。

定义 2 模式极值差:子模式中的最大值和最小值之间的差值。

$$vd = x_{imax} - x_{imin} \quad (1)$$

定义 3 模式斜率:连接重要点的直线段的实际斜率。

$$tg = \frac{x_{i_n} - x_{i_1}}{n - 1} \quad (2)$$

定义 4 模式均值:子序列中各时间点数据均值。

$$\text{spx} = \frac{\sum_{j=i_1}^{i_2} x(j)}{n} \quad (3)$$

定义5  $p$  和  $q$  的距离:设  $p = (x_p, y_p, z_p)$ ,  $q = (x_q, y_q, z_q)$ , 则  $p$  和  $q$  之间的距离为:

$$\text{dist}(p, q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2 + (z_p - z_q)^2} \quad (4)$$

定义6 异常因子(lof):该模式距离与正常模式的距离的比值。

$$\text{lof}(i) = \frac{d_i}{D} \quad (5)$$

其中,  $D$  为正常模式间距离;  $d_i$  为第  $i$  段子序列的距离。

定义7 异常模式:如果异常因子大于给定的阈值,则为异常模式。

异常模式检测将子序列的斜率、均值、极值差组成的三维特征空间进行距离的计算,三者值域差别很大,但衡量时间序列都很重要,因此要将三者的值域进行规范化处理。设  $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$  为其中一个子序列,则利用式(6)将该组特征值规范化到值域为(0, 1)的区间。

$$\text{norm}(x_{li}) = \frac{x_{li} - x_{\min}}{x_{\max} - x_{\min}} \quad (6)$$

其中,  $x_{\max}$  和  $x_{\min}$  分别表示各特征值的最大值和最小值。

时间序列异常检测在模式间距离的基础上求出异常因子,通过判断异常因子是否超出给定的阈值来判断模式状态。先通过算法1将时间序列进行线性分割,通过式1~3计算出每段子序列的极值差、斜率和均值,利用式(6)将时间序列的每个特征值规范到(0, 1),将每段子序列看成该空间中的一个点,其坐标值为规范化后的斜率、均值和极值差。由定义可知异常模式的异常因子较大。传统的距离计算的方法需要计算模式与其他每个模式间的距离,复杂度高。相对于频繁模式,异常模式是极少出现的模式,因此在计算模式间距离时,无需计算一个子模式与其余每个子模式的距离,只需在一个周期内取一个子模式,计算一个子模式与所取子模式间的距离即可,将每个子模式与其他子模式间的距离的均值作为该模式的距离,再将每个模式的距离求均值作为正常模式的距离。利用式(5)计算出异常因子,在一定程度上降低了时间复杂度。

算法2:异常模式检测算法。

输入:  $((s_1, e_1), (s_2, e_2), \dots, (s_m, e_m), d)$

输出:异常模式。其中  $s_m$  为子序列的起始位置,  $e_m$  为终止位置。

万方数据

Step1:计算子序列间的模式距离距  $d[m]$

For  $i=1$  to  $m$

$d1[i] = d(i, T), d2[i] = d(i, 2 * T) \dots \dots dn[i] = d(i, n * T)$

Step2:将 Step1 中每个子序列与其他子序列之间的距离,去掉最大值后的均值作为子模式的距离  $d[i]$ 。所有子模式的距离排序后取中位数,各中位数的均值作为正常模式距离。

Step3:根据式(5)求出异常因子。

Step4:异常因子超出阈值,输出异常模式。

### 3 实验

#### 3.1 实验结果

文中算法采用实测数据和合成数据进行验证。实测数据集采用南水北调工程渗透压力斜测仪检测数据,断面桩号:SH(3)+699,日期为2015年10月18日到2016年10月21日,p8-2的监测数据。原始时间序列如图1所示。

利用算法1对时间序列进行分段线性表示,结果如图2所示。

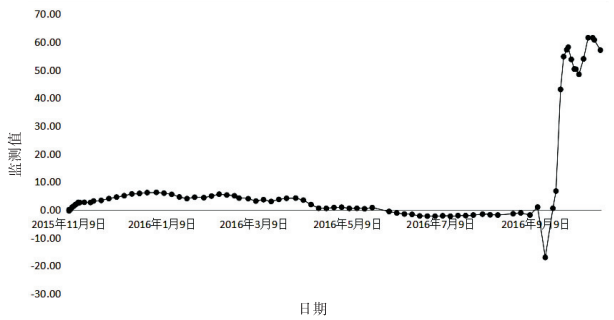


图1 南水北调工程监测数据原始时间序列

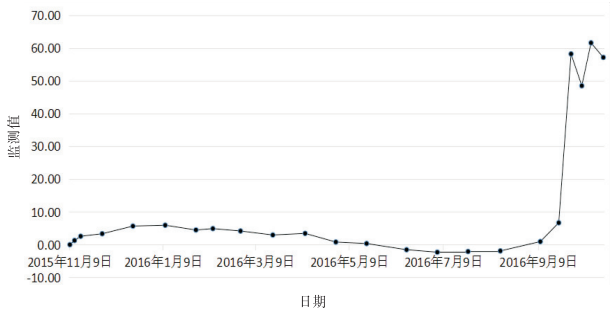


图2 分段线性表示后的时间序列

利用人工合成数据进行检验,原始时间序列如图3所示。

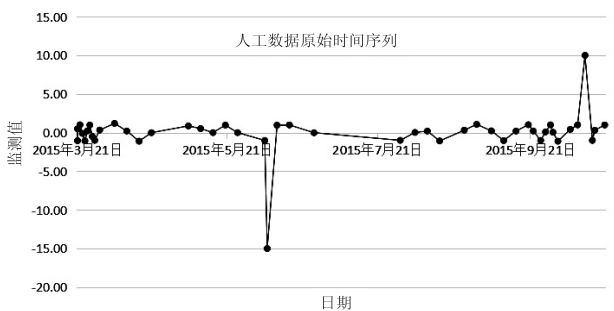


图3 人工合成数据原始时间序列

利用算法 1 对人工数据进行分段线性表示,结果如图 4 所示。

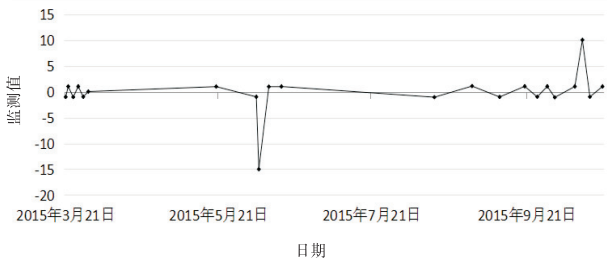


图 4 人工合成数据分段线性表示

当斜率阈值  $d$  取不同值时,对实测数据异常检测的输出结果如表 1 所示。

表 1  $d$  取不同值时的实验结果(实测数据)

$d$	压缩率/%	异常因子	异常时间段
1	60.24	3.224 59	2015/9/22–2015/9/25
2	68.67	3.015 69	2015/9/22–2015/9/25
3	69.87	3.287 75	2015/9/22–2015/9/25
4	72.29	3.946 28	2015/9/22–2015/9/25
>4	72.29	3.946 28	2015/9/22–2015/9/25

斜率阈值  $d$  取不同值时,对人工数据异常检测结果如表 2 所示。

表 2  $d$  取不同值时的实验结果(异常数据)

$d$	压缩率/%	异常因子	异常时间段
1	39.13	3.224 56	2016/6/5–2016/6/10 2016/10/10–2016/10/16
2	54.37	3.015 69	2016/6/5–2016/6/10 2016/10/10–2016/10/16
3	71.21	3.946 82	2016/6/5–2016/6/10 2016/10/10–2016/10/16
>3	71.21	3.946 82	2016/6/5–2016/6/10 2016/10/10–2016/10/16

3.2 实验分析

结果表明,用实测数据和合成数据均能正确检测出异常模式。基于斜率的时间序列分段线性表示有一个参数:斜率差值的阈值。实验表明,检测结果受阈值影响较小。随着  $d$  值的增大,分段数目越少,压缩率越高。对于南水北调工程安全检测数据,当  $d > 4$  时压缩率保持不变,检测结果受阈值影响较小;对于人工合成数据,当  $d > 3$  时,压缩率保持不变,检测结果受阈值影响较小,两种数据验证均取得了正确的检测结果。

4 结束语

针对如何在大量的时间序列中提取极少出现的异常模式,将时间序列进行线性分割,将不限分分段数目

与子序列长度的方法相结合,提出了基于斜率与最大时间跨度的分段算法。提取了时间序列的极值差、斜率、均值三个特征值,将其映射到特征空间,降低了时间序列的维数,实现了较高的压缩率。通过实测数据与合成数据进行实验,均能高效地检测出异常时间段,证明了该算法的有效性与可行性。

参考文献:

[1] 贾国栋.多相关周期性时间序列上的异常模式关联规则挖掘[D].沈阳:东北大学,2010.

[2] 谭宏强,牛强.基于滑动窗口及局部特征的时间序列符号化方法[J].计算机应用研究,2013,30(3):796–798.

[3] KORN F, JAGADISH H V, FALOUTSOS C. Efficiently supporting ad hoc queries in large datasets of time sequences[J]. ACM SIGMOD Record,1997,26(2):289–300.

[4] 陈帅飞,吕鑫,戚荣志,等.一种基于关键点的时间序列线性表示方法[J].计算机科学,2016,43(5):234–237.

[5] 曹文平,罗颖,熊启军,等.基于二次回归的时间序列分割算法[J].计算机光盘软件与应用,2012(18):157.

[6] 刘博,郭建胜.改进的多元时间序列符号化表示方法研究[J].计算机仿真,2015,32(1):314–317.

[7] OBUCHOWSKI J, WYŁOMANŃSKA A, ZIMROZ R. The local maxima method for enhancement of time – frequency map and its application to local damage detection in rotating machines[J]. Mechanical Systems and Signal Processing,2014,46(2):389–405.

[8] KEOGH E, CHAKRABARTI K, PAZZANI M, et al. Dimensionality reduction for fast similarity search time series databases[J]. Knowledge and Information Systems,2008,3(3):263–286.

[9] GEORGOULAS G, KARVELIS P, STYLIOS C D, et al. Automating the broken bar detection process via short time Fourier transform and two – dimensional piecewise aggregate approximation representation[C]//IEEE energy conversion congress and exposition. [s. l.]:IEEE,2014:3104–3110.

[10] 李敏,刘轲,罗惠琼,等.基于混合高斯模型的异常检测算法改进[J].计算机应用与软件,2014,31(6):198–200.

[11] 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.

[12] 苏卫星,朱云龙,刘芳,等.时间序列异常点及突变点的检测算法[J].计算机研究与发展,2014,51(4):781–788.

[13] 尚华.两类时间序列模型的异常值检测研究[D].北京:首都经济贸易大学,2016.

[14] 李少波,孟伟,曺晶晶.基于密度的异常数据检测算法 GSWCLOF[J].计算机工程与应用,2016,52(19):7–11.

[15] 孙梅玉.基于距离和密度的时间序列异常检测方法研究[J].计算机工程与应用,2012,48(20):11–17.