文章编号:1674-7070(2014)06-0515-05

戴慧[1]　阚建飞[1]　李伟仁[2]　周卫东[2]

多变量时间序列滑动窗口异常点的检测

摘要

针对多变量时间序列(MTS)的异常点的探测问题,提出了采用由粗到细的二次探测方案.基于滑动窗口数据的置信区间,构造了变化趋势值特征和相对变化趋势值特征分别用于二次探测,同时研究了特征的快速提取算法.通过对OPEN3000数据监测系统采集的事故发生前后某市城南变电站各设备表的数据集进行异常点探测,结果表明提出的算法能够快速准确地探测出异常点的位置.

关键词

多变量时间序列;滑动窗口;异常点;置信区间

中图分类号 TP206

文献标志码 A

0　引言

时间序列是属性值在时间顺序上体现出来的特征数据集合,多变量时间序列(MTS)在工业界得到广泛关注.由于系统的观测变量之间具有关联性,需要对具有相关性的观测变量进行综合比较分析.如果变量的观测值偏离其他的观测值太远,可能是由其他机制导致的不正常数据,这样的观测值被称为异常点.多变量时间序列是时间序列的子序列,需对时间序列数据集进行处理和比较分析,挖掘出具有异常点的多变量时间子序列集合.在工业界对事故的分析和预测的研究中,为调度运行人员提供电网系统的设备健康状态评价和电网的故障辅助分析决策支持显得非常重要,而前兆数据就是挖掘出的包含异常点的MTS.由于系统采集的各种实时、连续、有序的时间序列值是典型的数据流,具有无边界性的特征,根据数据流的特点和MTS相关性的要求,本文提出了一种对MTS进行异常点探测的算法,建立了滑动窗口的相对变化趋势的模型用于异常点的提取和状态监测,研究了快速的滑动窗口数据特征提取算法,用于对电网各设备事故发生之前的非正常状态模式下的探测.

目前,时间序列异常点的探测研究大部分针对单变量时间序列,包括基于距离的算法[1]、基于离群指数的算法[2]、基于偏差的算法[3]和基于小波变换的算法[4]等.针对MTS,文献[5]提出了基于滑动窗口的MTS异常数据的挖掘算法,但该算法在检测效率和运算效率上都有待改进.本文提出由粗到细的二次探测方案,构造新的基于滑动窗口置信区间的特征,并研究特征的快速提取算法,从而提升异常点的探测精度和加快速度.

1　问题的提出

本文是对OPEN3000数据监测系统采集的各种电力数据流进行异常点探测.OPEN3000系统采集的设备表中变量之间的相关性符合MTS的特征要求和符合数据流无限性和实时性的特征要求,同时也符合数据流在实际应用中的需要.本文要解决的问题是将OPEN3000数据采集系统中监测与记录的电流数据描述成MTS,然后采用滑动窗口的方法在MTS中找出含有异常数据的MTS子序列.

假设一个长度为 k 含有 n 个变量MTS为 $x(t)=(x_1(t),x_2(t),\cdots,x_n(t)),1\leqslant t\leqslant k$,其变量 i 的第 j 个长度为 l 的滑动窗口为 $s_{i,j}=$

收稿日期 2014-09-06

资助项目 江苏省政府留学奖学金基金;国家自然科学基金(61103141);江苏省高校自然科学基金(13KJB520015)

作者简介

戴慧,女,博士生,讲师,研究方向为智能电网和智能大数据.1520759669@qq.com

1　南京工程学院　计算机工程学院,南京,210013

2　德州大学阿灵顿分校　电子工程系,达拉斯,美国,76013

$(x_i(j),x_i(j+1),\cdots,x_i(j+l-1))$.该滑动窗口其实就是对应一段长度为 l 的 MTS 子序列,因此总共可以得到$(k-l+1)$ 个 MTS 子序列.要解决的问题就是找出这些 MTS 子序列中含有异常的一些子序列.

2 算法设计

本文提出算法的框图如图 1 所示.下面对各个模块涉及的算法和模型进行详细介绍.

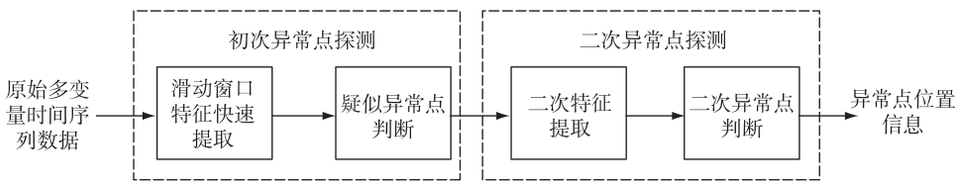

图 1 提出算法的框图

Fig. 1 Block diagram of a two-step outlier detection algorithm

2.1 滑动窗口特征快速提取及疑似异常点判断

2.1.1 滑动窗口变化趋势值特征

对于每个变量的每个滑动窗口,考虑采用其变化趋势值作为特征进行异常点的初步判断,而变化趋势值定义为每个窗口置信区间的距离半径.变量 x_i 的第 j 个长度为 l 的滑动窗口 $s_{i,j}=(x_i(j),x_i(j+1),\cdots,x_i(j+l-1))$ 的置信区间的距离半径为

$$d_{i,j}=\left|\hat{\theta}_{i,j}-\tilde{\theta}_{i,j}\right|/2, \tag{1}$$

式(1)中,置信上限 $\hat{\theta}_{i,j}=\bar{s}_{i,j}+\frac{\sigma_{i,j}}{\sqrt{l}}Z_{\alpha/2}$,其中 $\bar{s}_{i,j}$ 为均值,$\sigma_{i,j}$ 为均方差,随机变量 $Z\sim N(0,1)$,α 为置信水平,取 0.05,置信下限 $\tilde{\theta}_{i,j}=\bar{s}_{i,j}-\frac{\sigma_{i,j}}{\sqrt{l}}Z_{\alpha/2}$.

2.1.2 疑似异常点判断

基于 2.1.1 提取的特征,给出疑似异常点的判断模型:假设含有 n 个变量的 MTS 序列 $x(t)$,对于第 j 个目标窗口,如果存在某个变量 x_i,使得距离半径 $d_{i,j}>\tau$,其中 τ 为阈值,则认为 MTS 序列 $x(t)$ 在第 j 个窗口存在疑似异常点.

2.1.3 滑动窗口均值和标准方差的快速计算算法

由式(1)可知,特征提取的关键在于均值 $\bar{s}_{i,j}$ 和均方差 $\sigma_{i,j}$ 的计算.对于长度为 k 滑动窗口长度为 l 的 MTS,共有$(k-l+1)$个滑动窗口.通常 MTS 都具有较大的长度 k,从而$(k-l+1)$个滑动窗口的均值和均方差的计算量是非常可观的,因此,研究它们的快速计算显得非常必要.

对于变量 x_i 的第 j 个长度为 l 的滑动窗口 $s_{i,j}$,易知其均值和均方差 $\sigma_{i,j}$ 分别为

$$\bar{s}_{i,j}=\frac{\sum_{t=j}^{j+l-1}x_i(t)}{l},\quad \sigma_{i,j}=\sqrt{\frac{\sum_{t=j}^{t+l-1}(x_i(t)-\bar{s}_{i,j})^2}{l}}. \tag{2}$$

接下来推导第 $(j+1)$ 个滑动窗口的均值 $\bar{s}_{i,j+1}$ 和标准方差 $\sigma_{i,j+1}$ 与其前一个窗口均值 $\bar{s}_{i,j}$ 和均方差 $\sigma_{i,j}$ 之间的关系.由式(2) 可得:

$$\bar{s}_{i,j+1}=\frac{\sum_{t=j+1}^{j+l}x_i(t)}{l}=\frac{\sum_{t=j}^{j+l-1}x_i(t)+x_i(j+l)-x_i(j)}{l}=\bar{s}_{i,j}+\frac{x_i(j+l)-x_i(j)}{l}, \tag{3}$$

$$\sigma_{i,j}=\sqrt{\frac{\sum_{t=j+1}^{j+l}(x_i(t)-\bar{s}_{i,j+1})^2}{l}}. \tag{4}$$

对于式(4)中的分子,由式(3)可得:

$$\sum_{t=j+1}^{j+l}(x_i(t)-\bar{s}_{i,j+1})^2=\sum_{t=j+1}^{j+l}\left(x_i(t)-\bar{s}_{i,j}-\frac{x_i(j+l)-x_i(j)}{l}\right)^2=\sum_{t=j+1}^{j+l}(x_i(t)-\bar{s}_{i,j})+\sum_{t=j+1}^{j+l}\left(\frac{x_i(j+l)-x_i(j)}{l}\right)^2-\frac{2(x_i(j+l)-x_i(j))}{l}\sum_{t=j+1}^{j+l}(x_i(t)-\bar{s}_{i,j})=$$

$$\sum_{t=j}^{j+l-1}(x_i(t)-\bar{s}_{i,j})^2+(x_i(j+l)-\bar{s}_{i,j})^2-(x_i(j)-\bar{s}_{i,j})^2+\frac{(x_i(j+l)-x_i(j))^2}{l}-\frac{2(x_i(j+l)-x_i(j))^2}{l}=$$

$$l\sigma_{i,j}^2+(x_i(j+l)-x_i(j))\left(x_i(j+l)+x_i(j)-2\bar{s}_{i,j}-\frac{x_i(j+l)-x_i(j)}{l}\right). \tag{5}$$

将式(5)代入式(4)可得:

$$\sigma_{i,j+1}=(\sigma_{i,j}^2+((x_i(j+l)-x_i(j))(l(x_i(j+l)+x_i(j)-2\bar{s}_{i,j})-x_i(j+l)+x_i(j))l^{-2})^{\frac{1}{2}}. \tag{6}$$

式(3)和式(6)给出了前后2个窗口均值和均方差之间的关系,也可通过前一窗口的均值和均方差来计算当前窗口的均值和均方差,而且比式(2)所示的直接计算更快速.由表1可以看出提出的快速计算算法与窗口大小无关,明显优于直接算法.

表1 直接计算和提出的快速计算的时间复杂度

Table 1 Time complexity comparison between direct computation and the proposed rapid computation

算法	均值		均方差	
	加法	乘法	加法	乘法
直接计算	$l-1$	1	$l+1$	$l-1$
快速计算	2	1	5	6

2.2 二次特征提取及二次判断

由于前文提取的变化趋势值特征只考虑了当前窗口,而忽视了相邻窗口之间的关联,所以需引入相对变化趋势值特征对疑似异常点进行二次判断.

基于前文所提取的每个滑动窗口的变化趋势值特征,定义该滑动窗口的相对变化趋势值(二次特征)为

$$\Omega_{i,j}=(d_{i,j}-d_{i,j-1})/d_{i,j-1}. \tag{7}$$

给出二次判断模型:假设含有n个变量的MTS序列$x(t)$存在m个具有相关性的变量x_i,$i=1,2,\cdots,m$,设这m个变量的第j个目标窗口的相对变化趋势值为$\Omega_{i,j}$,如果满足下述条件则认为MTS序列$x(t)$在第j个窗口存在异常:

$$\max_{l\leqslant i,j\leqslant m}|\Omega_{i,j}-\Omega_{i',j}|>\gamma, \tag{8}$$

其中γ为阈值.

2.3 相关性模型

二次判断考虑的是具有相关性的变量之间的相对变化趋势值特征,因此定义2个变量序列之间的相关性(广义相关性)[6-10].对于长度为k滑动窗口长度为l的MTS序列$x(t)$中2个变量序列$x_i(t)$和$x_{i'}(t)$,定义它们之间的相关性系数R为

$$R=\frac{\sum_{j+1}^{k-l+1}R_j}{k-l+1}, \tag{9}$$

其中R_j为$x_i(t)$和$x_{i'}(t)$的2个相应第j个滑动窗口$s_{i,j}$和$s_{i',j}$之间的相关性系数:

$$R_j=\frac{\sum_{t=j}^{j+l-1}(x_i(t)-\bar{s}_{i,j})(x_{i'}(t)-\bar{s}_{i',j})}{\sqrt{\sum_{t=j}^{j+l-1}(x_i(t)\bar{s}_{i,j})^2\sum_{t=j}^{j+l-1}(x_{i'}(t)-\bar{s}_{i',j})^2}}. \tag{10}$$

由式(9)易知相关性系数的取值范围为$|R|\leqslant 1$.$|R|$越接近于1,表明变量之间的相关程度越高,它们之间的关系越密切.

2.4 算法步骤

输入:长度为k含有n个变量的MTS;滑动窗口的长度为l.

输出:含有异常数据的MTS子序列时间范围.

算法步骤如下:

1) 采用式(3)快速计算每个变量的($k-l+1$)个滑动窗口的均值;

2) 采用式(6)快速计算每个变量的($k-l+1$)个滑动窗口的均方差;

3) 采用式(1)求解每个变量的($k-l+1$)个滑动窗口的置信区间距离半径;

4) 将距离半径与阈值τ进行比较,初步确定含有突变点的MTS子序列的位置;

5) 采用式(7)计算具有相关性的变量在含有突变点的滑动窗口的相对变化趋势值;

6) 依据式(8)二次确定含有异常点的MTS子序列的时间范围.

3 实验仿真

3.1 实验数据集

数据集为OPEN3000系统于2013年9月2日1:10至2013年9月3日1:05对江苏省某市城南变电站每隔5 min对8个设备表采集的数据集,共有288个采样点,每个采样点包含了8个设备表的信息.事故发生的具体时刻为2013年9月2日13:05:55.用MTS来表示采集的数据:$x(t)=(x_1(t),x_2(t),\cdots,x_8(t))$,$1\leqslant t\leqslant 288$,其中$x_1(t)$表示高有功,$x_2(t)$表示高A相电流,$x_3(t)$表示高B相电流,$x_4(t)$表示高C相电流,$x_5(t)$表示低有功,$x_6(t)$表示低A相电流,$x_7(t)$表示低B相电流,$x_8(t)$表示低C相电流值.

3.2 实验结果及分析

根据系统对数据的采样频率以及经验理论将滑动窗口的大小设计为1 h,即长度l为12.将初次异常点判断的阈值τ设定为0.6,阈值γ设为0.01.实验环境:操作系统是Windows XP,CPU为2.40 GHz,内存2 GB,硬盘60 GB,采用MatlabR2006b进行代码编译.通过计算8个变量各滑动窗口置信区间的距离半径,发现只有变量高有功存在2个突变点,如图2所示.对应的时间范围为2013年9月2日9:05—9:10和2013年9月2日23:50—23:55.

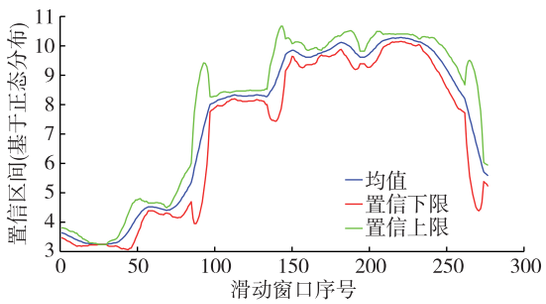

图 2　高有功变量的滑动窗口置信区间

Fig. 2 Confidence interval of sliding window with high active variables

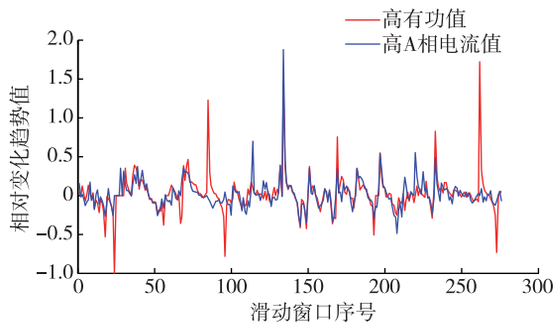

图 3　高有功值与高 A 相电流相对变化趋势

Fig. 3　Variation of high active value(red) and high A-phase current (blue) at 110 kV Grid Transformer Substation

由于突变点可能是正常范围的电流值升高所导致的,并不一定是异常点,需进行二次异常点探测.二次探测是对提取 MTS 数据中具有相关性的变量的相对变化趋势值特征进行判断,所以需先分析变量之间的相关性.对于采集的 MTS 电力数据 $x(t)$ 中的每个滑动窗口,采用相关性模型计算 8 个变量之间的相关性系数(表 2).表 2 表明这 8 个变量之间均具有较强的相关性,相关性系数的绝对值均不小于 0.98.这与电力系统的经验是一致的.

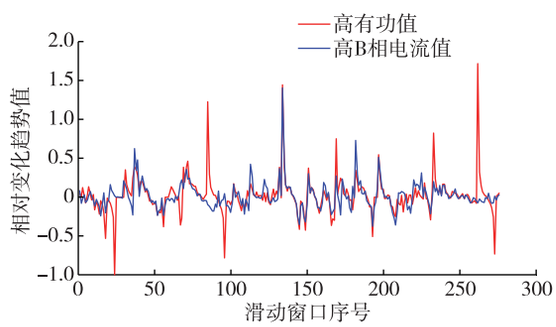

图 4　高有功值与高 B 相电流相对变化趋势

Fig. 4　Variation of high active value(red) and high B-phase current (blue) at 110 kV Grid Transformer Substation

图 3—5 展示了高有功和高 A、B、C 相电流之间的相对变化趋势.可知,存在 2 个点的相对变化趋势值的差值均超过了阈值 $\gamma = 0.01$,时间范围为 2013 年 9 月 2 日 9:05—9:10 的 MTS 子序列和 2013 年 9 月 2 日 23:50—23:55 的 MTS 子序列,所以这 2 个 MTS 子序列就是异常子序列,存在异常点.

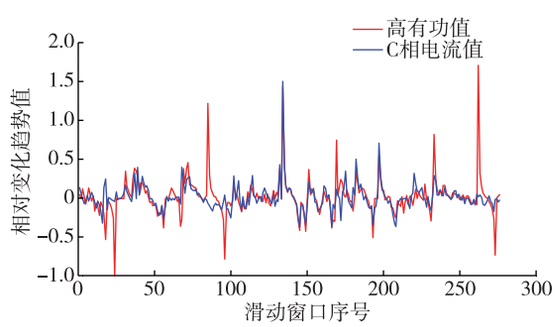

图 5　高有功值与高 C 相电流相对变化趋势

Fig. 5　Variation of high active value(red) and high C-phase current (blue) at 110 kV Grid Transformer Substation

4　结论

针对存在异常的 MTS 数据流,提出了一种由粗到细的两阶段探测策略,同时还研究了快速滑动窗口特征的提取算法.针对某市城南变电站各设备表采集的异常数据的实验结果很好地说明了提出算法的有效性.

表 2　MTS 中 8 个变量之间的相关性系数值

Table 2　Correlation values between 8 MTS variables

	高有功	高 A 相电流	高 B 相电流	高 C 相电流	低有功	低 A 相电流	低 B 相电流	低 C 相电流
高有功	1	0.98	0.98	0.98	−0.98	0.98	0.98	0.98
高 A 相电流	0.98	1	0.99	0.99	−0.99	0.99	0.99	0.99
高 B 相电流	0.98	0.99	1	0.99	−0.99	0.99	0.99	0.99
高 C 相电流	0.98	0.99	0.99	1	−0.99	0.99	0.99	0.99
低有功	−0.98	−0.99	−0.99	−0.99	1	−0.99	−0.99	−0.99
低 A 相电流	0.98	0.99	0.99	0.99	−0.99	1	0.99	0.99
低 B 相电流	0.98	0.99	0.99	0.99	−0.99	0.99	1	0.99
低 C 相电流	0.98	0.99	0.99	0.99	−0.99	0.99	0.99	1

参考文献

References

[1] 廖国琼,李晶.基于距离的分布式 RFID 数据流孤立点检测[J].计算机研究与发展,2009,47(5):172-179
LIAO Guoqiong,LI Jing.Distance-based outlier detection for distributed RFID data streams [J]. Journal of Computer Research and Development,2009,47(5):172-179

[2] 郑斌祥,席裕庚,杜秀华.基于离群指数的时序数据离群挖掘[J].自动化学报,2004,30(1):70-77
ZHENG Binxiang,XI Yugeng,DU Xiuhua.Outlier mining for time series data based on outlier index[J].Acta Automatica Sinica,2004,30(1):70-77

[3] 谭庆,张瑞玲.基于局部偏离因子的孤立点检测算法[J].计算机工程,2008,34(17):59-61
TAN Qing,ZHANG Ruiling.Outlier detection algorithm based on local deviation factor [J]. Computer Engineering,2008,34(17):59-61

[4] 文琪,彭宏.小波变换的离群时序数据挖掘分析[J].电子科技大学学报,2005,34(4):556-558
WEN Qi,PENG Hong.Analysis of time series outlier mining based on wavelet transform[J].Journal of University of Electronic Science and Technology of China ,2005,34(4):556-558

[5] 翁小清,沈钧毅.基于滑动窗口的多变量时间序列异常数据的挖掘[J].计算机工程,2007,33(12):102-104
WENG Xiaoqing,SHEN Junyi.Outlier mining for multivariate time series based on sliding window [J]. Computer Engineering,2007,33(12):102-104

[6] Zhang Y,Meratnia N,Havinga P. Outlier detection techniques for wireless sensor networks:A survey[J].IEEE Communications Surveys & Tutorials,2010,12(2):159 -170

[7] Lee J-G,Han J W,Li X L.Trajectory outlier detection:A partition-and-detect framework[C]//IEEE 24th International Conference on Data Engineering,2008:140-149

[8] Yang K,Shahabi C.A PCA-based similarity measure for multivariate time series[C]//Proceedings of the Second ACM International Workshop on Multimedia Databases,2004:65-74

[9] Agyemang M.LSC-Mine:Algorithm for mining local outliers[C]//Khosrow-Pour M.Innovations Through Information Technology,2004,doi:10.4018/978-1-59140-261-9.ch002

[10] Krämer J,Seeger B.Semantics and implementation of continuous sliding window queries over data streams[J]. ACM Transactions on Database Systems,2009,34(1):19-26

Outlier detection for sliding window of multi-variable time series

DAI Hui[1] KAN Jianfei[1] LEE Weiren[2] ZHOU Weidong[2]

1 School of Computer Engineering,Nanjing Institute of Technology,Nanjing 210013

2 Department of Electronic Engineering,University of Texas in Arlington,Dallas 76013

Abstract This paper proposes a two-step detection scheme that begins thick and ends thin,to mine the outliers of multivariable time series (MTS).According to the confidence interval of the data in sliding window,characteristics of both variation trend value and relevant variation trend value were constructed,which were then used in the two detection processes.Meanwhile,the rapid extraction algorithm for characteristics is studied.The outlier detection scheme is then applied to mine outliers before and after an accident happened at a 110 kV Grid Transformer Substation in Jiangsu province.Data sets of various equipment tables,which were collected by OPEN3000 data surveillance system,were checked by the proposed detection scheme,and experiment result indicates that this algorithm can rapidly and precisely locate the outliers.

Key words multivariate time series;sliding window;outliers;confidence interval