

基于大数据的时间序列异常点检测研究

程艳云,张守超,杨 杨

(南京邮电大学 自动化学院,江苏 南京 210023)

摘要:针对传统时间序列异常点检测方法在处理大量数据时检测精度与效率低下的缺陷,文中提出一种基于大数据技术的全新时间序列异常点检测方法。首先介绍了传统时间序列异常点检测方法并分析了其缺陷。其次介绍了基于大数据方法的理论推导,包括特征提取、奇异点检测及异常点判别,具体为采用大数据方法将海量序列分解为周期分量、趋势分量、随机误差分量及突发分量四个不同分量,对不同分量进行特征提取并根据特征提取结果进行奇异点检测,并在此基础上利用序列特点判别奇异点是否为异常点。最后通过实验分析对比验证大数据方法的可行性与效率。实验结果表明,基于大数据方法的时间序列异常点检测相比于传统的方法具有更高的检测精度与更快的检测速率。

关键词:异常点检测;时间序列;大数据;特征提取

中图分类号: TN915.07

文献标识码: A

文章编号: 1673-629X(2016)05-0139-06

doi: 10.3969/j.issn.1673-629X.2016.05.030

Research on Time Series Outlier Detection Based on Big Data

CHENG Yan-yun, ZHANG Shou-chao, YANG Yang

(College of Automation, Nanjing University of Posts and Telecommunications,
Nanjing 210023, China)

Abstract: According to the detection accuracy and efficiency limitation of traditional time series outlier detection methods when dealing with a large amount of data, a new time series outlier detection method is put forward, which is based on the big data technology. Firstly, the traditional time series outlier detection methods are introduced, analysis of their defects. Secondly, it introduces the theoretical derivation of big data method in this paper, which can be divided into feature extraction, abnormal detection and outlier distinguish. The massive series is decomposed into four different components, including periodic component, trend component, random error component and burst component. Then the feature is extracted to four components and abnormal detection is made according to the result of extraction. On this basis it determines whether abnormal point is outlier by series characteristic. Finally, the feasibility and efficiency of big data approach is verified by experiment analysis and comparison. The results show that the big data method has higher precision and rate compared with traditional methods.

Key words: outlier detection; time series; big data; feature extraction

0 引言

所谓异常点,即数据集中与数据的一般行为或对象不一致的数据对象,异常点一般也称作离群点^[1]。数据的不确定性是产生异常点的主要原因,数据的不确定性可分为存在的不确定性和值的不确定性两类^[2]。简而言之,就是数据测量和收集误差、自然变异、数据不确定性等原因导致异常点的出现。异常数据往往包含着很重要的信息,对实验结果与分析起到了重要作用,一方面不应该将异常数据简单地作为错

误数据来处理,另一方面应当寻找有效的方法去检测并挖掘这些异常点所隐含的意义。

时间序列是由记录值和记录时间组成的元素的有序集合^[3]。时间序列的分析主要包括趋势分量、季节性分量、突发分量以及随机误差分量,而趋势分量与随机误差分量是时间序列中异常点检测的重要研究方向。文中将在分析时间序列特性的基础上,比较传统时间序列异常点检测方法的各自优缺点,结合大数据算法,研究如何采用大数据方法来对时间序列进行异

收稿日期:2015-07-06

修回日期:2015-10-14

网络出版时间:2016-05-05

基金项目:江苏省自然科学基金(BK20140877, BE2014803)

作者简介:程艳云(1979-),女,副教授,硕士生导师,研究方向为自动控制原理、网络优化;张守超(1991-),男,硕士研究生,研究方向为大数据挖掘在通信网络中的应用。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20160505.0817.046.html>

常点的检测与分析,从而提高检测的效率与精度,为下一步的数据处理提供帮助。

1 传统时间序列异常点检测

为了减少异常点对实验结果的干扰,需要对异常点进行检测并处理。异常数据的挖掘主要使用偏差检测,包括聚类法、序列异常法、最近邻居法、多位数据分析法等^[4-5]。通过偏差检测可以在一定程度上发现异常点,但是也存在部分缺陷,比如导致两种不良的后果:(1)掩盖现象,即未能识别出某些真正的离群点;(2)淹没现象,即将正常点误判为离群点^[6]。

时间序列的一个最重要特征就是具有时间属性,序列值之间必须按照时间先后顺序进行严格的排序。针对这一特性,产生了很多种时间序列异常点检测方法,主要分为以下几大类:

(1) 统计学方法^[7-9]。

主要包括基于统计的异常点检测算法、基于密度的方法、基于距离的异常点的检测算法等等,然而这类方法需要在多维空间中寻找异常点,并不适用于一维的时间序列,并且在使用统计学方法前必须得知道数据的分布模型,这就涉及到模型参数的问题,但是这些信息一般事先是不知道的。

(2) 机器学习方法^[10-12]。

机器学习方法主要可以划分为两大类:一是人工神经网络,二是支持向量机。两类方法也是各有优缺点:人工神经网络在处理小规模问题上具有很好的应用效果,但是对于大规模的问题,人工神经网络的构造将会非常复杂,因此不能很好地往大规模问题上扩展;相对于人工神经网络,支持向量机不仅具有相同的处理能力,而且在计算效率上也有很大的提高,但是支持向量机在理论方面或者在建立模型方面都相对比较复杂,因此在实际应用中存在一定的难度。

(3) 其他方法^[13-16]。

除了上述提及的两大类方法,还包括基于空间的方法、基于小波的方法、基于 AR 模型的方法等等。基于小波的方法虽然在查询性能上有所改进,但是对短期的异常模式无法检测;而基于 AR 模型的方法需要知道时间序列模型。

虽然时间序列异常点检测研究领域出现了很多算法,但是这些算法还不够成熟,尤其面对日益增加的数据量,传统的时间序列异常点检测方法在效率与精度上都达不到预期要求,所以必须采用新的方法来进行处理。

2 基于大数据的异常点检测算法

由于时间序列具有结构简单的特点,传统时间序

列异常点检测方法面临的主要问题在于时间序列特征难以提取,而且面对海量数据时,传统方法在处理能力上已显得力不从心。为了解决这两大问题,文中在时间序列基本特点的基础上,结合全新的大数据处理算法,提出一种新的时间序列异常点检测方法。通过分析隐藏在海量数据背后的特征,提取新的时间序列特征进行分析,从而提高预测效率与精度,改进预测速率,克服传统方法的缺陷。

时间序列由周期分量、趋势分量、突发分量及随机误差分量四个分量组成^[17],每个分量均具有不同的特征。文中将先对四个分量进行特征提取并根据特征提取结果进行奇异点检测,然后结合四个分量共同特点进行异常点检测,提高检测精度与速率。图 1 为基于大数据方法的异常点检测方法流程图。

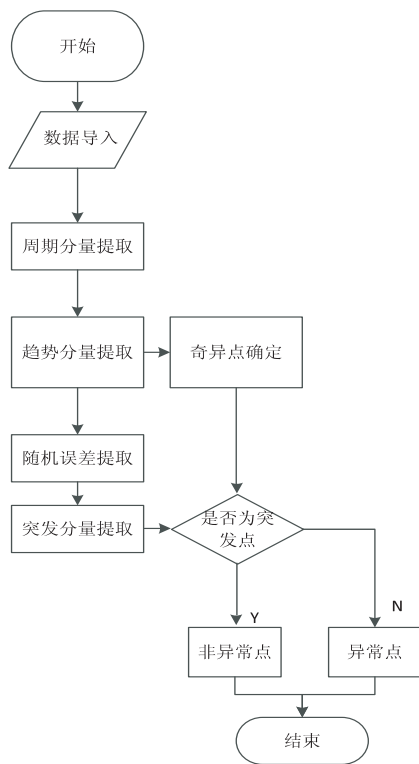


图 1 基于大数据的时间序列异常点检测流程图

1) 周期分量特征提取。

时间序列中周期分量的特征提取主要为时间序列周期的确定,一般的周期确定方法有傅里叶变换、小波变换、差分计算等^[18]。文中将根据时间序列周期特点,结合大数据处理方法,采用一种全新的方法来确定周期 L 。首先根据式(1)对时间序列 $X\{x_1, x_2, \dots, x_m\}$ (x_i 范围为 $V_1 - V_2$) 进行差分计算得到矩阵 A (其中 $m > 30 * 48, n > 50$), 如下所示:

$$A = \begin{bmatrix} x_2 - x_1 & x_3 - x_2 & \cdots & x_{n-1} - x_{n-2} \\ x_3 - x_1 & x_4 - x_2 & \cdots & x_n - x_{n-2} \\ \vdots & \vdots & \vdots & \vdots \\ x_m - x_1 & x_{m+1} - x_2 & \cdots & x_{n+m-3} - x_{n-2} \end{bmatrix} \quad (1)$$

对矩阵 A 的每一行进行线性拟合,参数分别记为 $(a_1, a_2, \dots, a_{m-1})$, $(b_1, b_2, \dots, b_{m-1})$, 将 A 的每一行下标分别代入对应的 $Y = aN + b$ 中得到对应的 A' , 如式(2)所示。

采用最小二乘法(见式(3))计算 A 与 A' 每行最小误差,其中首次出现最小误差的行数即为周期 L 。

$$A' = \hat{a}N + \hat{b} =$$

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{m-1} \end{bmatrix} \begin{bmatrix} 1 & 2 & \cdots & n-2 \\ 1 & 2 & \cdots & n-2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 2 & \cdots & n-2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{m-1} \end{bmatrix} \quad (2)$$

$$\forall i \in (1:m-1), \text{Error}(i) = \sum_{j=1}^{n-2} (A_{ij} - A'_{ij})^2, \text{则}$$

$$\text{Error} = \begin{bmatrix} (x_2 - x_1 - a_1 - b_1)^2 + (x_3 - x_2 - 2a_1 - b_1)^2 + \cdots + (x_{n-1} - x_{n-2} - (n-2)a_1 - b_1)^2 \\ (x_3 - x_1 - a_2 - b_2)^2 + (x_4 - x_2 - 2a_2 - b_2)^2 + \cdots + (x_n - x_{n-2} - (n-2)a_2 - b_2)^2 \\ \vdots \\ (x_m - x_1 - a_{m-1} - b_{m-1})^2 + (x_{m+1} - x_2 - 2a_{m-1} - b_{m-1})^2 + \cdots + (x_{n+m-3} - x_{n-2} - (n-2)a_{m-1} - b_{m-1})^2 \end{bmatrix} \quad (3)$$

2) 趋势分量特征提取。

周期分量特征的正确提取是进行趋势分量特征提取的前提条件。通过周期分量特征提取得到周期 L , 将时间序列 X 以周期 L 进行划分得到矩阵 B , 其中 x_M 为时间序列 X 最后一个数据, x_M 之后数据均为空值 NA。

$$B = \begin{bmatrix} x_1 & x_2 & \cdots & x_i & \cdots & x_L \\ x_{L+1} & x_{L+2} & \cdots & x_{L+i} & \cdots & x_{2L} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N \div L + 1} & x_{N \div L + 2} & \cdots & x_M & \text{NA} & \cdots \end{bmatrix} \quad (4)$$

矩阵 B 具有两个方向的特征坐标,同一行内所有点代表处于同一周期的所有时间点的集合,同一列内的所有点代表处于不同周期同一位置的点的集合。将矩阵 B 的每一列依次取出,得到共计 L 个时间序列 $\{x_1, x_{L+1}, \dots, x_{N \div L + 1}\}$, $\{x_2, x_{L+2}, \dots, x_{N \div L + 2}\}$, \dots , $\{x_i, x_{L+i}, \dots, x_M\}$, $\{x_{i+1}, x_{L+i+1}, \dots, \text{NA}\}$, \dots , $\{x_L, x_{2L}, \dots, \text{NA}\}$, 分别记做 $L_1 \sim L_L$ 。其中,每个 L_i 序列均有 N 个数据(x_M 所在列之后序列最后一位数值均为 NA)。

图2展示了时间序列 X 及经过趋势分量提取之后的序列 X' 。图(a)中, X 坐标表示时间点, Y 坐标表示序列值大小;图(b)中, X 坐标表示不同周期, Y 坐标表示单周期长度, Z 坐标表示序列值大小。

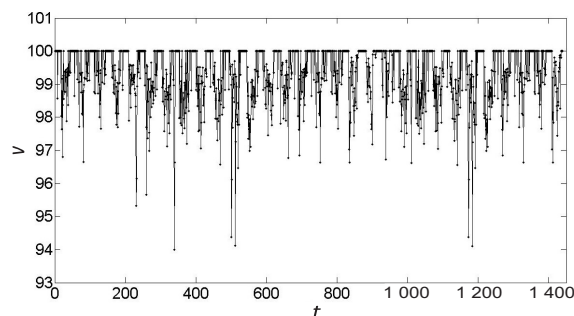
对所有序列 L_i 分别进行如下操作:

(1) 如果序列中存在 NA, 则将 NA 剔除, 序列长度变为 $N-1$;

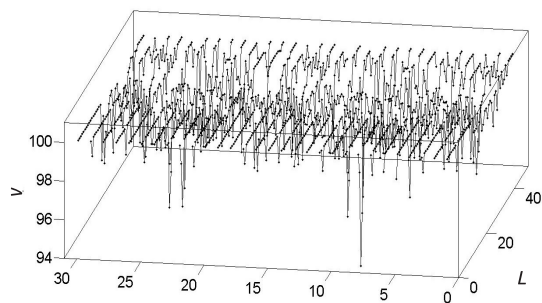
(2) 对处理后的 L_i 进行聚类分析^[19], 离群点划入奇异点 E 。

对于序列 L_i 内的所有点,其本质为序列 X 内所有周期内相同位置的点的集合,排除突发分量和随机误差影响,理论上具有相同的分布特性。若序列 L_i 趋势分量为固定值,则 L_i 内所有点处于同一条水平直线上,该直线之外的所有点则均认为是奇异点;若序列 L_i

趋势分量按照一定的规律分布,则不按照该规律分布的点视为奇异点;若序列 L_i 趋势分量为随机分布,则需要先找出随机分布范围 $[\min, \max]$, 在该范围之外的点均为奇异点。



(a) 周期特征提取前时间序列 X



(b) 趋势分量提取后序列 X

图2 时间序列 X 及趋势特征提取过后 X'

3) 随机误差分量 R_i 特征提取。

传统方法一般认为时间序列随机误差分布函数服从正态分布,其均值为0,方差则根据实际情况确定。而在文中方法中,将根据时间序列的实际情况来确定随机误差分量的分布函数,具体方法如下所示:

首先将矩阵 B 的每一行依次取出,得到共计 N 个时间序列 $\{x_1, x_2, \dots, x_L\}$, $\{x_{L+1}, x_{L+2}, \dots, x_{2L}\}$, \dots , $\{x_{N \div L + 1}, x_{N \div L + 2}, \dots, x_M, \text{NA}, \dots\}$, 分别记做 $N_1 \sim N_N$, 其中每个 N_i 序列均有 L 个数据(N_N 内 x_M 后所有数值均为 NA)。然后对序列 X 进行统计分析,得到序列 X 的

分布模型 $X \sim F(x)$ ($V_1 \leq x \leq V_2$)。对于所有的序列 N_i , 按照 $F(x)$ 分布模型可以得到一组全新的序列 N'_1, N'_2, \dots, N'_N , 对新生成的序列按从小到大排序, 然后对于原有的序列 N_1, N_2, \dots, N_N 内元素从小到大排序并与序列 N'_1, N'_2, \dots, N'_N 内元素依次相减, 得到 N 个随机误差序列, 记为 $R_i(i)$ 。根据 $R_i(i)$ 可以得到每个序列 N_i 的随机误差分布模型, 记做 $N_{e1} \sim F_{e1}(r), N_{e2} \sim F_{e2}(r), \dots, N_{eN} \sim F_{eN}(r)$ 。

$$F_{ei} = \sum_{r_{\min}}^{r_{\max}} P_i(r) \quad (5)$$

根据所有的 $R_i(i)$, 可以得到序列 X 的随机误差分布模型, 记为 $X_e \sim F_e(r)$ 。

$$F_e = \sum_{r_{\min}}^{r_{\max}} P(r) \quad (6)$$

根据随机误差分布模型, 即可得到序列 X 的随机误差分布范围, 为下一步的判别奇异点是否为异常点做好准备工作。

4) 突发分量 B_i 特征提取。

突发分量特征提取是判别奇异点是否为异常点的前提, 分别对 $N_1 \sim N_N - 1$ 行做如下操作:

(1) 对于序列 L_i , 依次计算每个序列的平均值, 记为 \bar{L}_i ;

(2) 分别按照式(7)和式(8)计算序列 N_i 与 \bar{L}_i 差值平方和及 N_i 的随机误差和。

$$\text{Sum}(i) = \sum_{j=1}^L (N_{ij} - \bar{L}_j)^2 \quad (7)$$

$$\text{Sum}(i)' = \sum_{r_{\min}}^{r_{\max}} P_i(r) * r^2 \quad (8)$$

(3) 如果 $\text{Sum}(i) > \text{Sum}(i)'$, 则序列 N_i 内数据均为突发点, 否则序列 N_i 内数据不为突发点。

对序列 N_N 先将 NA 剔除, 得到长度为 $M - N * (L - 1)$ 的序列 N'_N , 对序列 N'_N 进行如上操作, 其中需要将 L 变为处理后 N_N 的实际长度, 判别 N_N 内数据是否为突发分量。

突发分量特征 B_i 提取之后, 判别奇异点 E 是否属于 B_i 或者 R_i 范围内, 若是, 则该奇异点不是异常点, 若否, 则该奇异点为异常点。方法总体步骤如下所示:

for $i = 1:m$

计算 A 得到周期 L

end

then

计算得到 $B(N * L)$

for $i = 1:L$

对 B 的每列进行趋势分量提取;

分析得到奇异点 $E: \{e_1, e_2, \dots, e_n\}$;

end

提取突发分量特征 B_i ;

then

提取随机误差分量特征 R_i ;

for E

if (e_i 属于 B_i 或者在 R_i 范围内)

e_i 为非异常点

else

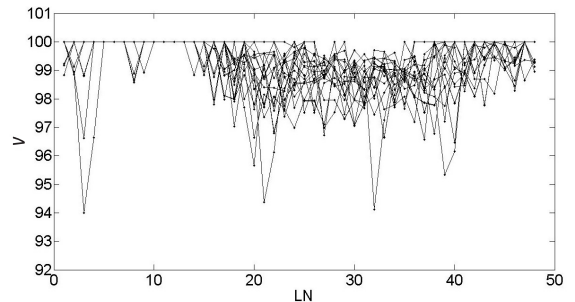
e_i 为异常点

end

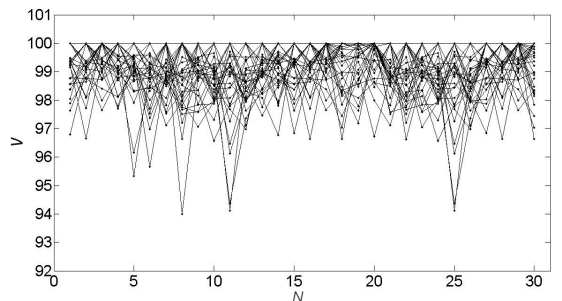
3 实验与结果

在通信网络中, 各项核心性能指标 (KPI) 均以时间序列形式表示。以单一小区为例, 单一 KPI 一年数据量长度为 $48 * 365$ 。文中将以通讯网络中时间序列为例, 分析并比较传统时间序列异常点检测方法与基于大数据的时间序列异常点检测方法各自的优缺点。

任取某一小区某一 KPI (RRC 设置成功率) 半年数据为例, 取 $m = 400$, $n = 100$ 进行差分处理得到矩阵 A ($400 * 100$), 并对矩阵 A 的每一行进行线性拟合并采用最小二乘法计算误差得到 Error 矩阵, 取矩阵 Error 首次出现最小值行数记为周期 L , 得到最优参数 $L = 48$, 按照 $L = 48$ 得到矩阵 B 。对矩阵 B 以周期为单位画图 (按行) 和以相同时间点作图 (按列), 见图 3。



(a) $L - V$ 维度矩阵 B



(b) $N - V$ 维度矩阵 B

图 3 $L = 48$ 矩阵 B 数据分布图

提取矩阵 B 的每一列得到不同的时间子序列 L_i , 对于所有的序列 L_i , 判别其趋势分量特征。若趋势分量为固定值, 采用聚类或线性拟合进行奇异点确定^[20]; 若趋势分量为规律分布, 根据规律进行奇异点

确定;若趋势分量为随机分布,根据分布函数进行奇异点确定。

图4分别展示了矩阵 B 的三种不同趋势分量分布奇异点确定方法。

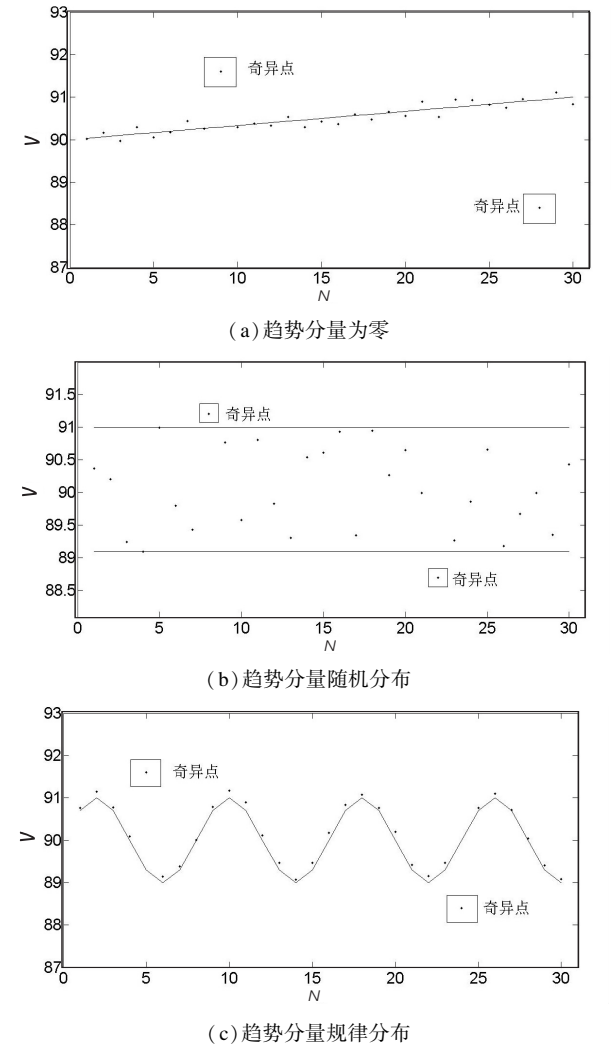


图4 矩阵 B 三种奇异点确定方法

如图(a)中所示,对于序列 N_i 内所有点理论均为固定值,即所有点的集合为一条直线,直线之外归为奇异点;图(b)中所有点为随机分布,找出分布的上下区间,取上区间的前5%和下区间的后5%的点记为奇异点;图(c)中,表示趋势分量呈一定的周期分布,在此周期之外的点为奇异点。

奇异点 E 提取过后,进行随机误差分量特征提取,按照大数据算法公式,可以得到整体分布函数为:

$$X \sim F(x) = \begin{cases} 0 & x < 87 \\ \sum_{x_i \leq x} p_k & 87 \leq x < 100 \\ 1 & 100 \leq x \end{cases}$$

统计每个周期的分布函数,通过计算可以得到随机误差分布函数为:

$$X_e \sim F_e(x) =$$

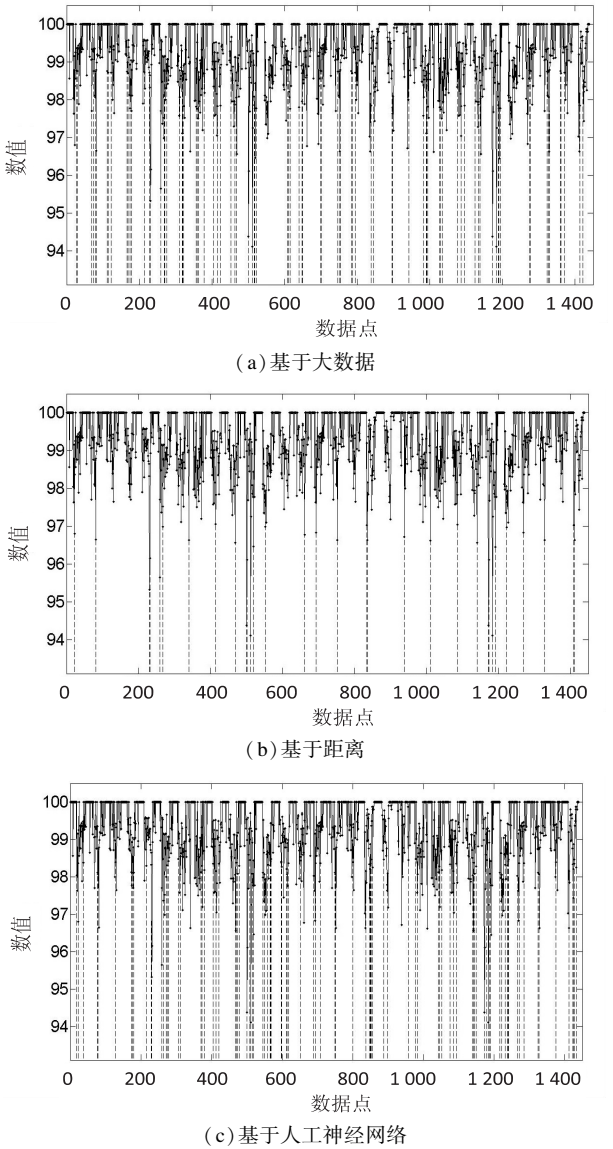
$$\begin{cases} 0 & -0.8 > x \\ \int_{-0.8}^x \left(\frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} + \left(1 - \int_{-0.8}^{0.8} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt \right) * \frac{(t+0.8)}{1.6} \right) dt & -0.8 \leq x < 0.8 \\ 1 & 0.8 \leq x \end{cases}$$

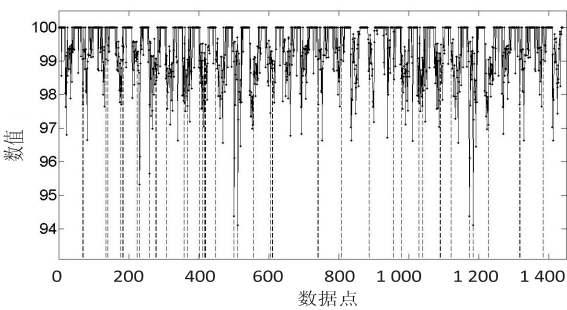
随机误差分布函数确定完成之后,进行突发分量特征提取。按照算法要求先计算每个序列 L_i 的均值得到 L ,计算 Sum 和 Sum',判别得到突发分量。

突发分量特征提取结束之后,进行最后一步计算,判别奇异点是否为突发分量或者在随机误差范围内。

图5为基于大数据方法的异常点检测算法、基于距离的异常点的检测算法^[21]、基于人工神经网络的异常点检测方法^[22]、基于AR模型方法的异常点检测方法^[23]结果对比图,其中虚线线条对应点即为异常点。

表1为上述四种方法对于不同数据量(短期:30 * 48个数据量;长期:12 * 30 * 48个数据量)的检测精度、检测效率、检测速率对比结果。





(d) 基于 AR 模型

图 5 不同方法异常点检测结果对比图

检测精度 = 正确的检测结果数 / 异常点总数值 * 100%

检测效率 = 正确的检测结果数 / 检测结果数 * 100%

检测速率 = 完成一次检测所需时间 (s)

表 1 短/长期检测精度、效率、速率结果对比

方法	短期精度	短期效率	短期速率	长期精度	长期效率	长期速率
	/%	/%	/s	/%	/%	/s
大数据方法	92.2	96.3	20	92.0	95.8	156
基于距离方法	90.5	83.9	13	89.9	79.7	150
人工神经网络	85.6	90.8	12	80.4	88.7	300
AR 模型法	87.4	92.3	13	82.6	90.1	273

通过实验结果可以看出,基于大数据的时间序列异常点检测方法在短期异常点检测中与传统方法相比在检测精度、检测效率上有一定的改进,但在检测速率上稍微逊色一点;但是在面对大量的长期数据时,基于大数据的时间序列异常点检测方法在检测精度与检测效率上均比其他方法有很大的提高,在检测速率上相比于其他方法,基于大数据的方法具有更快的速率。

4 结束语

面对日益增长的数据量,文中采用大数据方法,基于时间序列特征提出了一种全新的时间序列异常点检测方法,并通过实验分析该方法的可行性与效率,达到了预期要求。同时作为刚刚兴起的大数据方法,还有许多需改进的地方,将来的工作需要对算法做进一步的改进,提高短期预测速率、长期预测精度与效率等。

参考文献:

[1] 曹忠虔. 时间序列异常检测的研究[D]. 成都:电子科技大学,2012.

[2] 郭春. 基于数据挖掘的网络入侵检测关键技术研究[D]. 北京:北京邮电大学,2014.

[3] Box G E P. 时间序列分析——预测与控制[M]. 上海:机械工业出版社,2011.

[4] 杨金伟. 基于距离和信息熵的不确定异常点检测研究[D]. 昆明:云南大学,2011.

[5] 刘良旭,乐嘉锦,乔少杰,等. 基于轨迹点局部异常度的异常点检测算法[J]. 计算机学报,2011,34(10):1966-1975.

[6] 刘丹丹,陈启军,森一之. 线性回归模型的多离群点检测方法及其节能应用[J]. 信息与控制,2013,42(6):765-771.

[7] 胡世杰,钱宇宁,严如强. 基于概率密度空间划分的符号化时间序列分析及其在异常诊断中的应用[J]. 振动工程学报,2014,27(5):780-784.

[8] 苏卫星,朱云龙,胡琨元,等. 基于模型的过程工业时间序列异常值检测方法[J]. 仪器仪表学报,2012,33(9):2080-2087.

[9] 杨越,胡汉平,熊伟,等. 一种基于超统计理论的非平稳时间序列异常点检测方法研究[J]. 计算机科学,2011,38(6):93-95.

[10] 王佳玮. 决策支持中基于时间序列数据的异常点检测[D]. 合肥:中国科学技术大学,2014.

[11] 陈敏. 基于BP神经网络的混沌时间序列预测模型研究[D]. 长沙:中南大学,2007.

[12] 崔万照,朱长纯,保文星,等. 基于模糊模型支持向量机的混沌时间序列预测[J]. 物理学报,2005,54(7):3009-3018.

[13] 庄雪鹏. 基于小波的时间序列中异常点的检测[D]. 南京:南京大学,2013.

[14] 张建平,李斌,刘学军,等. 基于Hadoop的异常传感数据时间序列检测[J]. 传感技术学报,2014,27(12):1659-1665.

[15] 王骏,钟富礼,王士同,等. 基于移相加权球面单簇聚类的周期时间序列异常检测[J]. 自动化学报,2011,37(8):984-992.

[16] 张玉飞,董永贵. 一种时间序列异常检测用参数化熵滤波器[J]. 机械工程学报,2011,47(22):13-18.

[17] 张蕾. 非线性时间序列的高阶统计特征提取和趋势分析[D]. 沈阳:沈阳航空航天大学,2012.

[18] 龚祝平. 混沌时间序列的平均周期计算方法[J]. 系统工程,2010,28(12):111-113.

[19] 韩娜. 聚类算法在时间序列中的研究与应用[D]. 广州:广东工业大学,2011.

[20] 闫秋艳,夏士雄. 一种无限长时间序列的分段线性拟合算法[J]. 电子学报,2010,38(2):443-448.

[21] Rasheed F, Alhajj R. A framework for periodic outlier pattern detection in time-series sequences[J]. IEEE Transactions on Cybernetics,2014,44(5):569-582.

[22] Buzzi-Ferraris G, Manenti F. Outlier detection in large data sets[J]. Computers and Chemical Engineering,2010,35:388-390.

[23] Li S T, Cheng Y C. A stochastic HMM-based forecasting model for fuzzy time series[J]. IEEE Transactions on Cybernetics,2010,40(5):1255-1266.