第47卷第8期 中国科学技术大学学报 Vol.47,No.8
2017年8月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Aug. 2017

文章编号:0253-2778(2017)8-0644-09

一种基于贝叶斯后验的异常值在线检测及置信度评估算法

孙栓柱[1,2],宋 蓓[2],李春岩[1,2],王 皓[2]

(1.江苏方天电力技术有限公司,江苏南京 211102;2.南京大学计算机软件新技术国家重点实验室,江苏南京 210023)

摘要:为识别一类更新速度快、变化趋势平缓、缺少人工类标的大数据量工业时间序列中所存在的异常值,提出了一种以贝叶斯后验为基础的异常值在线检测及置信度评估算法.算法将预测检测和假设检验相结合,首先建立时间序列自回归模型,然后对预测残差作基于贝叶斯原理的后验检验,用后验概率对数比确定序列中的异常值.为减少识别过程中的误判,在检测完成后,利用自组织映射神经网络计算状态转移概率,进一步对已标记的异常值进行置信度评估.通过定期更新模型,算法各参数能动态保持与数据变化规律同步,提高了检测的准确率.实验结果表明,该算法能够对时间序列异常值准确快速地进行在线检测,同时给出可靠的置信度评估,具有较高的实用价值.

关键词:时间序列;异常检测;贝叶斯后验;置信度评估

中图分类号:TP399 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2017.08.003

引用格式:孙栓柱,宋蓓,李春岩,等.一种基于贝叶斯后验的异常值在线检测及置信度评估算法[J].中国科学技术大学学报,2017,47(8):644-652.

SUN Shuanzhu, SONG Bei, LI Chunyan, et al. An online outlier detection and confidence estimation algorithm based on Bayesian posterior ratio[J]. Journal of University of Science and Technology of China, 2017,47(8):644-652.

An online outlier detection and confidence estimation algorithm based on Bayesian posterior ratio

SUN Shuanzhu[1,2], SONG Bei[2], LI Chunyan[1,2], WANG Hao[2]

(1. *Jiangsu Frontier Electric Technology Co. Ltd., Nanjing* 211102, *China*;
2. *State Key Laboratory for Novel Software Technology at Nanjing University, Nanjing* 210023, *China*)

Abstract: In order to satisfy the outlier detection requirements in one kind of high-speed, small-variance unlabeled industrial time series, an online outlier detection and confidence estimation algorithm based on Bayesian posterior ratio was proposed. The algorithm combined prediction and hypothesis testing, establishing the autoregressive model firstly and then using Bayesian posterior logarithm of residuals to identify outliers. To reduce misjudgment, the state transition probabilities were calculated by self-organizing map neural network and the reliability of detected outliers was evaluated afterwards. It updated models periodically to dynamically adapt to data changes, thus improving accuracy. Experimental results demonstrate that the online algorithm can effectively detect outliers in time series provide reliable

收稿日期:2017-05-26;**修回日期**:2017-07-14

基金项目:国家自然科学基金(61503178),江苏省自然科学基金(BK20150587)资助.

作者简介:孙栓柱,男,1973年生,硕士/教授级高级工程师.研究方向:电力行业节能减排技术及工业大数据挖掘.E-mail:sunshuanzhu@sina.com

通讯作者:王皓,博士.E-mail:wanghao@nju.edu.cn

confidence evaluation, bringing higher adaptability and practicability.

Key words: time series; outlier detection; Bayesian posterior ratio; confidence estimation

0　引言

时间序列具有数据量大、维数高、变化迅速等特点，蕴含极高的潜在价值，近年来受到越来越多学者的关注.目前的时间序列挖掘研究主要集中在序列分割、模式识别、序列分类及聚类等方向，其中大多数期望从带有时间属性的海量数据中挖掘出既定规则，往往将存在异常的点当作噪声进行简单剔除处理.事实上，这些少量的异常数据却隐藏着重要的信息，值得进一步挖掘分析.当前，时间序列异常检测技术已广泛应用于防信用卡诈骗[1]、反股市操控[2]、抗网络入侵[3]及医疗诊断[4]等.

最早出现的异常值检测方法利用统计假设检验来进行异常值识别[5]，它通过假设数据服从某种概率模型，根据不一致性判断异常.大多数情况下数据的分布难以事先得知，这使得该类方法的应用范围受限.Knorr 等[6]首先提出了基于距离的异常检测算法.他们计算集合中所有数据之间的距离，比较距离大小以确定异常发生在何处，但这种方法对参数选择较为敏感.Ramaswamy 等[7]推广了距离的概念，优化了时间复杂度.这类基于距离的方法从全局角度出发，忽略了局部可能存在的异常.考虑局部可能性，Breuning 等[8]提出了基于密度的检测算法，根据对象的局部邻域密度来描述离群程度，从而判断数据发生异常的概率.

已有的很多异常值检测研究往往针对无序数据集，对有序关联的时间序列数据并不适用；并且，仅仅将异常视为一种二元特性，即正常或异常是不准确的.为满足一类更新速度快、变化趋势平缓、缺少人工类标的大数据量工业时间序列异常值检测需求，本文提出了一种基于贝叶斯残差后验的异常值在线检测及置信度评估算法.通过将预测检测和假设检验相结合，在线辨识时间序列数据中的异常值，并在识别完成后，利用自组织映射神经网络状态转移概率，根据历史数据状态变化规律对异常值出现的可能性进行估算，实现置信度评估.这种方法不仅能识别各维度中存在的数据异常，还能识别各维度之间可能存在的关系异常.检测过程中，模型参数定期更新，以提高算法准确率.在人工数据集和燃煤机组烟尘超低排放浓度数据上的实验结果表明，该方法可以快速并有效地检测时间序列中存在的异常值并给出可靠的置信度评估，具有较高的实用价值.

1　相关研究

对于时间序列中的异常，学术界并未给出统一的定义.目前被普遍采用的是 Hawkins[9]给出的定义：异常是在数据集中偏离大部分数据的数据，这些数据疑似并非为随机误差所致，而是产生于完全不同的机制.从 20 世纪 80 年代起，异常检测作为数据挖掘的一个重要分支得到了广泛研究，学者提出的异常检测方法主要可以分为以下 5 类：

(Ⅰ)基于假设检验的异常检测方法.假设检验是最早用来发现异常样本的方法.它基于统计学基础，首先假设数据集服从某个已知分布或概率模型.模型确定后，若该数据集中某点与其服从的分布存在不一致，则此处可能发生异常.Abraham 等[10]提出时间序列中识别异常点的贝叶斯方法；之后，基于其他各种分布模型，提出了如 t 检验，偏度-峰度检验的异常值检测方法，但这类方法往往针对单个属性，难以处理高维数据；并且对大量分布特征未知的数据，先验假设并不一定成立.

(Ⅱ)基于距离的异常检测方法.基于距离的检测方法设定某种距离函数来对数据点之间的距离进行计算，当某个数据点与其余点距离过大时，将其视为异常.Knorr 等[6]首次提出了基于距离的异常检测算法，他们定义对象 q 为 DB(p,q)异常，如果存在大于 p 个点与目标 q 的距离大于 d.在此基础上，Ramaswamy 等[7]把距离推广到 k-近邻的距离 $D_k(p)$，将 $D_k(p)$ 值最大的前 n 个点定义为异常.这类方法易于理解和实现，但也存在一些不足.由于数据集中各对象之间的距离计算均从全局出发，若集合中存在多种不同分布时，效果并不好.包含不同密度子集的数据同样效果不佳，而且基于距离的异常检测方法对参数十分敏感，需要合理设置参数，应用受到一定限制.

(Ⅲ)基于密度的异常检测方法.鉴于异常往往存在于局部，Breuning 等[8]提出了基于密度的异常值检测算法.为方便数据集密度的度量，他们提出了

局部异常因子(local outlier factor,LOF)这一概念.对数据中的一个对象,给定参数确定最小邻居数 k 和邻域,计算对象的 k-距离,再计算此邻域中的可达距离和可达密度,定义邻域的可达密度与其自身可达密度之比(LOF).LOF 越大,意味着对象离群程度越高,是异常值的可能性也就越高.此后,有学者提出用增量式 LOF(I-IncLOF)[11]等异常程度的度量方法,提高了计算效率.这类方法克服了不同密度子集混合造成的检测错误,检测精度较高,既能够发现全局异常点,也能够发现局部异常点,但它们的算法时间复杂度较高,对相应参数的选择,也缺少统一可靠的标准.

(Ⅳ)基于聚类的异常检测方法:基于聚类的检测利用已有的聚类算法如 DBSCAN[12]等来识别时间序列中存在的异常.它们将数据集聚类为若干簇,不属于任何簇的数据点即视作异常点,较为简单直观.由于聚类分析与异常值检测存在本质上的差别,该方法通常只是聚类算法研究的附属产品,检测精度和算法效率都不高.

(Ⅴ)基于预测的异常检测方法.许多学者采用贝叶斯网络[13]、支持向量机[14]、神经网络[15]等模型对时间序列的历史数据建立预测模型,通过计算数据点预测值与实际值的偏离来判断异常.这类方法虽然简洁直观,但不适宜处理多元时间序列的异常检测问题.此外,基于预测的检测算法精度很大程度上依赖于预测模型本身的预测能力,因此较难确定合理阈值.

此外,Martins 等[16]基于最小二乘支持向量机和滑动窗口思想提出了一种时间序列在线异常检测算法.Johansen 等[17]定义了一系列与 Huber 跳跃及最小修剪方差统计量相关的渐进检测算法,包括一步 Huber 跳跃和向前搜索.刘芳等[18]提出了适用于调节系统震荡数据异常检测的自回归模型与小波相结合的在线异常值检测算法.杨志勇等[19]利用知识粒度方法来查找时间序列中的异常数据,减少检测过程时间花费,提高了检测效率.

本文将预测检测与假设检验相结合,实现了对一类更新速度快、变化趋势平缓、人工类标不足的大数据量工业时间序列的异常值在线辨识和置信度评估.该算法高效准确,既不要求数据满足正态分布,也无需事先给出异常标记,克服了传统异常值检测方法的不足,并能够为复杂的工业故障检测和诊断提供准确可靠的早期预警.

2 基于贝叶斯后验的异常值在线检测及置信度评估算法

2.1 算法描述

基于贝叶斯后验的异常值在线检测及置信度评估算法,结合了预测检测与假设检验.首先采用预测的思想对给定时间序列数据 $\{x_1,x_2,\cdots,x_N,\cdots\}$ 建立自回归模型,得到拟合残差序列,然后对残差序列作后验检验,识别原序列中的异常点.在检测完成后,利用自组织映射神经网络训练得到的状态转移矩阵对检测出的异常点进行置信度评估,最终确定数据的异常情况.算法总体上包括 3 个阶段:①模型离线训练;②异常值在线辨识;③模型批量更新.具体如算法 2.1 所示:

算法 2.1 基于贝叶斯后验的异常值在线检测及置信度评估算法

输入:时间序列历史数据$\{x_1,x_2,\cdots,x_N\}$.

输出:时间序列历史数据中各点正常或异常,若异常则同时输出异常置信度.

1)对历史数据建立 p 阶自回归模型 AR(p),用训练出的模型预测后续数据值;

2)将 1)中得到的数据预测值与真实值对比,求得预测残差 $\{e_1,e_2,\cdots,e_N\}$;

3)用核密度估计计算 2)中预测残差序列的概率密度函数;

4)以历史数据作为输入,训练并建立自组织映射神经网络状态模型,进一步得到离散的状态序列$\{C_1,C_2,\cdots,C_K\}$和状态转移概率矩阵,矩阵中第 i 行第 j 列元素值代表从状态 C_i 转移到状态 C_j 的概率;

5)利用 2)中估计出的残差序列先验概率和条件概率,分别计算新来的数据点 x 正常和异常的后验概率;

6)以 5)中两者的后验概率对数比作为指标,判断新来数据点 x 是否为异常点,输出正常或异常;

7)对 6)中输出为异常的点,利用 4)中状态转移概率矩阵计算从前序数据点对应状态到该点对应状态的转移概率,输出置信度评分;

8)在每 N 个新数据到达并被检验完成后,返回 1),将训练窗口向后滑动 N 个位置,更新预测模型;返回 3),更新概率密度函数;返回 4),将此 N 个数据输入自组织映射神经网络,更新状态转移矩阵.

9)重复以上步骤,直至无新数据输入.

2.2 基于贝叶斯后验的异常值在线检测

基于贝叶斯后验的异常值在线检测首先选取一个大小为 L 的滑动窗口,建立 t 时刻数据 x_t 与前 AR(p)个历史数据 $\{x_{t-p},x_{t-p+1},\cdots,x_{t-1}\}$ 的 p 阶

自回归模型，预测 t 时刻的数据值.通过计算预测残差并对预测残差作基于贝叶斯的后验检验，确定时间序列中的异常值位置.此过程中，相应的概率计算采用无需过多先验信息的核密度估计来完成.

2.2.1　基于时间序列自回归模型的数据预测

时间序列自回归模型(auto-regressive，AR)[20]分析和表征时间序列数据之间的相互依赖性与相关性，是一种线性预测方法.

给定一个时间序列 $\{x_1,x_2,\cdots,x_t\}$，p 阶自回归模型 AR(p)将当前值 x_t 建模为它 p 个相邻历史值的线性组合加上常数项和随机误差.AR 模型用于本文算法第一阶段的数据拟合.

在建立 AR 模型的过程中，可通过向后滑动训练窗口改变训练集，对模型进行动态更新.用训练好的模型对时间序列数据进行预测，对比数据的预测值和真实值，计算得到预测残差：

$$e_t = x_t - \hat{x}_t \tag{1}$$

式中，e_t 为时刻 t 的预测残差；x_t、$\hat{x}_t$ 分别为时刻 t 的样本值和预测值.

2.2.2　基于核密度估计的残差概率密度函数估计

核密度估计(kernel density estimation，KDE)[21]是一种非参数估计方法，用于估计未知密度函数.与参数估计方法不同，核密度估计可以在不利用任何先验条件的情况下，根据数据样本对未知密度函数进行估计，实现所估结果与真实结果间具有最小均方积分误差的目标.

核密度估计通过将一个移动的单元格(核函数)陆续放置在每一个数据点的位置上通过叠加获得一条光滑的曲线.其中核函数的选择条件为单个峰值下的函数面积为 1.

假设 $x_1,x_2,\cdots,x_N$ 为独立同分布 F 的 N 个样本点，其概率密度为 f，则其核密度函数估计为

$$\hat{f}_h(x) = \frac{1}{N}\sum_{i=1}^{N} K_h(x-x_i) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right) \tag{2}$$

式中，$K(\cdot)$ 为核函数，通常满足对称性及 $\int K(x)\mathrm{d}x = 1$.核函数是一种加权函数，其中数据点 x_i 到 x 的距离 $(x-x_i)$ 影响点 x_i 在估计点 x 时的加权作用大小.离 x 点越近的样本点在估计时起到的加权作用越大，公式如下：

$$K(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} \tag{3}$$

式(2)中，$h\,(h>0)$ 是一个平滑参数，称为带宽.h 的选择对估计 $f(x)$ 有很大影响，当 h 很小时，只有特别接近 x 的点才能起到较大作用；随着 h 的增大，距离 x 较远的点在估计中的作用也随之增加.标准正态核函数的带宽 h 可由 Silverman 拇指法则[22]得到

$$h = \left(\frac{4}{3N}\right)^{\frac{1}{5}}\sigma \tag{4}$$

式中，σ 是样本标准差.

采用 KDE 方法得到了预测残差的概率密度函数后，就可以计算时刻 t 的残差概率 $p(e_{t-L}^{t})$.具体计算方法如下：假设经 AR 模型预测后所得残差序列为 $e_{t-L}^{t} = \{e_{t-L},e_{t-L+1}\cdots,e_t\}$，取时间段 $(t-L,t)$ 中任意时间点 i 的残差 e_i 所属值域区间的概率作为此点的概率，则

$$p(e_{t-L}^{t}) = \prod_{i=t-L}^{t} p(e_i) \tag{5}$$

2.2.3　基于贝叶斯原理的后验检验

用核密度估计得到相应概率值后，对 AR 模型中的预测残差序列进行后验检验.选取一个大小固定为 L' 的滑动窗口，检验当前数据点与其前 L' 个数据是否同时服从高斯分布 $N(0,v_{L'})$.若是，则认定窗口中数据未发生异常；反之，异常发生.检验假设如下：

H_0：t 时刻数据未发生异常，是正常点；

H_1：t 时刻数据为异常点.

根据上述假设得到的似然分别为

$$p(e_{t-L'}^{t} \mid H_0) = \prod_{i=t-L'}^{t} p(e_i \mid v_{L'}) \tag{6}$$

$$p(e_{t-L'}^{t} \mid H_1) = \prod_{i=t-L'}^{t-1} p(e_i \mid v_{L'}) \times p(e_t \mid v_t) \tag{7}$$

式(6)表示在 H_0 假设下，t 时刻的数据 x_t 为正常点的似然概率.由于 H_0 假设 t 时刻数据未发生异常，因此 t 时刻数据点的方差与前 L' 个数据的方差 $v_{L'}$ 相同.式(7)表示在 H_1 假设下，t 时刻的数据 x_t 为异常点的似然概率，在此假设下 t 时刻数据的方差与前面 L' 个数据的方差不同，用 v_t 表示.

方差在零均值高斯概率函数中敏感度极高[23]，为了克服方差估计对异常值检测准确性的影响，采用边缘化处理方法，对方差进行积分处理.根据贝叶斯公式可得两个假设的后验概率分别为

$$p(H_0 \mid e_{t-L'}^{t}) = \frac{p(H_0)}{p(e_{t-L'}^{t})} p(e_{t-L'}^{t} \mid H_0) =$$

$$\frac{p(H_0)}{p(e_{t-L'}^{t})}\prod_{i=t-L'}^{t}p(e_i\mid v_{L'})=$$

$$\frac{p(H_0)}{p(e_{t-L'}^{t})}\int p(e_{t-L'}^{t}\mid v_{L'})\times p(v_{L'})\,\mathrm{d}v_{L'} \tag{8}$$

$$p(H_1\mid e_{t-L'}^{t})=\frac{p(H_1)}{p(e_{t-L'}^{t})}p(e_{t-L'}^{t}\mid H_1)=$$

$$\frac{p(H_1)}{p(e_{t-L'}^{t})}\prod_{i=t-L'}^{t-1}p(e_i\mid v_{L'})\times p(e_t\mid v_t)=$$

$$\frac{p(H_1)}{p(e_{t-L'}^{t})}\int p(e_{t-L'}^{t-1}\mid v_{L'})\times p(v_{L'})\,\mathrm{d}v_{L'}\times$$

$$\int p(e_t\mid v_t)\times p(v_t)\,\mathrm{d}v_t \tag{9}$$

式中，$p(H_0)$、$p(H_1)$ 分别为假设 H_0、H_1 的先验概率，满足 $p(H_0)+p(H_1)=1$，如果将异常值置信度设置为 95%，则 $p(H_0)=0.05$，$p(H_1)=0.95$. $p(e_{t-L'}^{t})$ 为拟合残差 $e_i(i=t-L',\cdots,t)$ 的先验概率，用 KDE 方法估计得到的残差概率密度函数计算.

两个假设边缘化处理后的后验概率计算公式为

$$p(H_0\mid e_{t-L'}^{t})=\frac{p(H_0)\,(2\pi)^{-\frac{L'+1}{2}}\Gamma\left(\frac{L'+1}{2}\right)}{p(e_{t-L'}^{t})A^{\frac{L'+1}{2}}} \tag{10}$$

$$p(H_1\mid e_{t-L'}^{t})=\frac{p(H_1)\,(2\pi)^{-\frac{L'+1}{2}}\Gamma\left(\frac{L'}{2}\right)\Gamma\left(\frac{1}{2}\right)}{p(e_{t-L'}^{t})A_1^{\frac{L'}{2}}A_2^{\frac{1}{2}}} \tag{11}$$

$$A=\frac{1}{2}\sum_{i=t-L'}^{t}e_i^2,\ A_1=\frac{1}{2}\sum_{i=t-L'}^{t-1}e_i^2,\ A_2=\frac{1}{2}e_t^2 \tag{12}$$

式中，$\Gamma(\cdot)$ 为伽马函数，通过下述近似方程计算：

$$\Gamma(a)\approx\sqrt{2\pi a}\ (a/e)^a \tag{13}$$

式中，$\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$.

式(10)、(11)分别代表残差为 $e_{t-L'}^{t}$ 时，t 时刻数据 x_t 未发生和发生异常的概率.进一步，引入指标后验概率对数比 $\varphi(t)$（简称后验概率比）衡量检验假设 H_0、H_1 之间的大小关系，并作为异常值判定的依据，计算公式如下：

$$\varphi(t)=\frac{\lg[p(H_1\mid e_{t-L'}^{t})]}{\lg[p(H_0\mid e_{t-L'}^{t})]} \tag{14}$$

由式(14)可以看出，如果 x_t 异常，则异常假设 H_1 的后验概率 $p(H_1\mid e_{t-L'}^{t})$ 远大于正常假设 H_0 的后验概率 $p(H_0\mid e_{t-L'}^{t})$，$\varphi(t)<1$，否则相反.设定后验概率对数比 $\varphi(t)$ 的检测阈值为 η，按照下式对数据序列进行异常值检验：

$$\varphi(t)\begin{cases}\leqslant\eta, x_t\text{ 是异常点}\\ =\text{else}, x_t\text{ 不是异常点}\end{cases} \tag{15}$$

式中，检测阈值 η 取值一般在 1 左右，通常取 0.95.

2.3 基于 SOM 状态模型的异常值置信度评估

在通过贝叶斯后验识别出异常值后，由于缺少数据是否确为异常的真实类标，还需要估计识别出异常点置信度，确定检测出的异常值有多大概率确为异常，以减少工业过程中的误判.在此阶段，利用自组织映射神经网络状态模型和状态转移概率矩阵确定从前一个数据状态转移到该异常点对应状态的概率，计算该点出现异常的可能性，进一步判断异常值假设成立的置信度.

2.3.1 SOM 状态转移模型

自组织映射神经网络(self organizing maps, SOM)[24]可以在一维或二维的处理单元阵列上形成输入信号的特征拓扑分布.它由输入层和输出层组成，输入层神经元为一维矩阵，接收网络的输入信号，其个数由输入向量个数决定；输出层的神经元按照一定的方式排列成一个二维节点矩阵，两层神经元之间通过权值相互连接.

SOM 神经网络对数据进行无监督学习聚类，训练时采用“竞争学习”的方式.输出层各神经元通过竞争与输入模式进行匹配，最后仅有一个神经元成为竞争的胜者，这个获胜的神经元代表对输入模式的分类.由于无监督学习的训练样本中不含期望输出也没有任何先验知识，因此 SOM 神经网络适用于对数据量大、不含类标签的数据作聚类分析.

具体来说，SOM 将整个时间序列 $\{x_1,x_2,\cdots,x_t\}$ 作为输入，序列 $C=\{C_1,C_2,\cdots,C_K\}$ 作为输出节点，将时间序列转化为线性空间中的离散点序列.对每一个时间点 t，离散点 $C_t\in\{C_1,C_2,\cdots,C_K\}$ 表示最接近 x_t 的状态，SOM 从本质上实现了对时间序列数据的状态聚类.

假设 SOM 模型输出神经元 i 代表的状态向量为 m_i，当训练样本 x 提供给网络后，样本与每个状态向量之间的欧氏距离，即计算样本与状态向量的相似度；然后网络根据相似度调整输出神经元状态向量，促使彼此相邻却不相似的神经元间距离最大，在训练结束时输出层能够对输入样本的数据分布进行最佳描述.神经元对应状态向量 m_i 按照下式进行更新：

$$m_i(t+1)=m_i(t)+h_{C(x),i}(x(t)-m_i(t)) \tag{16}$$

式中，t 为学习步长；$x(t)$ 为第 t 步中 x 的训练样本，$h_{C(x),i}$ 是递减的近邻函数，第一个下标 $C=C(x)$ 定义如下：

$$\forall i,\ \|x(t)-m_C(t)\| \leqslant \|x(t)-m_i(t)\| \tag{17}$$

式中，$m_C(t)$ 是第 t 步中所有神经元状态向量中与输入样本 $x(t)$ 最相似的神经元，称作最佳匹配单元.式(16)中的递减近邻函数通常采用高斯函数，形如：

$$h_{C(x),i}=\alpha(t)\exp\left(-\frac{\|r_i-r_C\|^2}{2\sigma^2(t)}\right) \tag{18}$$

式中，$0<\alpha(t)<1$ 是单调递减的学习系数；r_i 、r_C 是神经元的位置，$\sigma(t)$ 是近邻函数的宽度.

SOM 网络学习算法具体步骤如下：

step 1 随机生成输出层中神经元状态向量；

step 2 对输入向量 $x(t)$，遍历输出层的每个神经元，计算输入向量 $x(t)$ 和输出层神经元状态向量 m_i 之间的相似度，以距离最小的神经元作为最佳匹配单元；

step 3 按式(16)更新最佳匹配单元邻域内神经元的状态向量；

step 4 增加步长 t，然后返回 step 2，直至步长超出预先设定的循环次数.

2.3.2　神经元状态转移概率矩阵

SOM 状态模型训练结束后，可得到一个状态序列 $\{C_1,C_2,\cdots,C_K\}$ 和一个输出层神经元之间的状态转移矩阵，矩阵中第 i 行第 j 列的元素值 $p_{i,j}$ 代表从状态 C_i 转移到状态 C_j 的概率.假设某时间序列 $\{x_t,x_{t+1}\}$ 通过 SOM 神经网络转换得到对应的状态序列为 $\{C_i,C_j\}$，由于 x_{t+1} 出现在 x_t 之后，因此可以认为发生了一次从状态 C_i 到 C_j 的转移，定义转移概率为

$$p_{i,j}=\frac{\text{状态 } C_i \text{ 转移到状态 } C_j \text{ 的次数}}{\text{状态 } C_i \text{ 转移到所有状态的次数}} \tag{19}$$

2.3.3　基于状态转移概率的异常值置信度评估

上述状态转移概率矩阵中，对角线元素(状态不变)取值最大，靠近转移概率矩阵对角线的元素(只在近邻状态之间转移)数值次之，矩阵外围元素数值最小.由于平稳数据序列的最大转移概率仅在 0.6 左右，而其近邻状态转移概率降幅明显，有些甚至降至 0.1 左右，不同状态之间的转移概率相差不显著.并且在 SOM 模型输出的状态聚类数目增加时，状态转移组合数随之增加，导致状态转移概率下降，所以仅仅通过简单比较状态转移概率大小来判断是否为异常点并不可靠.本文采用最大-最小比较的思路引入异常打分函数，通过计算异常状态转移概率(最小概率)与最频繁发生的状态转移概率(最大概率)之比，得到一个更为显著直观的评价指标，对检测出来的异常点进行必要的置信度评估.

假设 x_t 为待评估的异常值，其前一个数据点为 x_{t-1}，通过训练好的 SOM 神经网络得到其对应的状态分别为 C_t 、C_{t-1}.通过查询状态转移概率矩阵可以找出状态 C_{t-1} 转移到状态 C_t 以及转移到最有可能转移状态 C_l 的概率 $p_{t-1,t}$ 、$p_{t-1,l}$，由此得到异常打分函数：

$$s=1-\frac{p(\text{状态 } C_{t-1} \text{ 转移到状态 } C_t)}{p(\text{状态 } C_{t-1} \text{ 转移到状态 } C_t)}=1-\frac{p_{t-1,t}}{p_{t-1,l}} \tag{20}$$

由上式可知，x_t 为异常值的可能性越大，其从前一个数据状态转移到 x_t 对应状态的概率 $p_{t-1,t}$ 越小，而 $p_{t-1,l}$ 是一个固定值，这导致比值变小，异常打分函数得分 s 增大.打分函数 s 的大小代表着数据异常发生的概率.通过计算此函数，可求得贝叶斯后验检测过程中已检测异常点的异常置信度，输出相应的异常评分.

3　实验分析

为验证本文提出的异常值检测及置信度评估算法的有效性，选择人工数据集和燃煤机组烟尘超低排放浓度数据集进行实验，将算法的检测结果与基于 AR 残差的异常值检测算法[25]、基于残差后验的异常值检测算法[23]进行对比.

3.1　评价指标

采用以下两个评价指标，对比人工数据集中已知真实异常情况和燃煤机组烟尘超低排放浓度数据集经甄别后确定异常情况，分析算法对时间序列异常值的检测效果.

$$\text{准确率}=\frac{\text{检测出的真实异常值个数}}{\text{样本真实异常值总数}}\times 100\% \tag{21}$$

$$\text{召回率}=\frac{\text{检测出的真实异常值个数}}{\text{检测出的异常值总数}}\times 100\% \tag{22}$$

3.2　人工数据集实验结果

利用高斯函数随机生成 500 组均值方差为 1 的白噪声信号，并加入 6 个异常点，如图 1 所示.

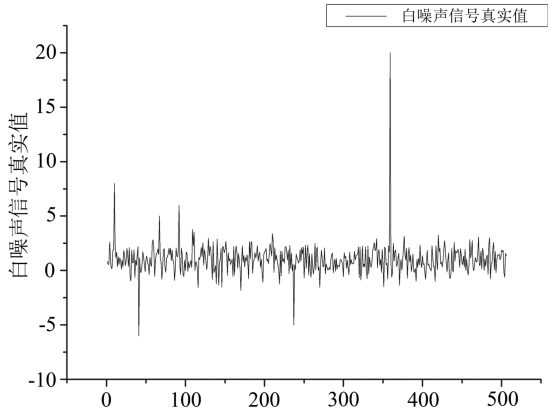

图 1 人工数据集

Fig.1 Synthetic data

使用基于贝叶斯后验的异常值在线检测及置信度评估算法进行实验，设置参数为滑动窗口大小 $L=90$、后验概率比检验阈值 $\eta=0.8$，得到此白噪声数据后验概率比如图 2 所示.

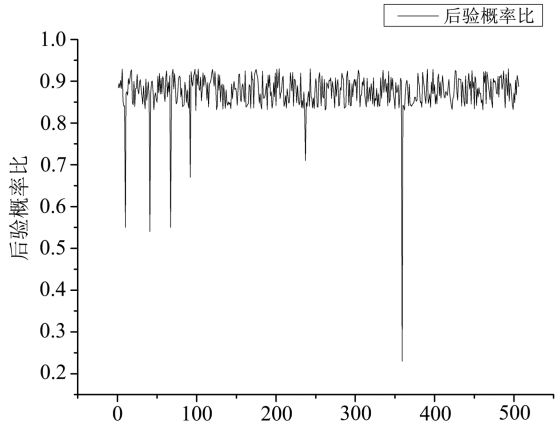

图 2 人工数据集后验比

Fig.2 Posterior ratio of synthetic data

从图 2 可以看出，异常点处的后验概率比明显小于非异常点.在设定检测阈值 $\varphi=0.82$ 的情况下，能够检测出所有插入的异常点.

进一步，我们进行 SOM 建模，得到状态模型转移概率矩阵，求得插入各异常点的得分，即异常可能性分别为 92%、90%、83%、81%、78%、100%.

在人工数据集上使用基于 AR 残差的异常值检测算法、基于残差后验的异常值检测算法进行检测.对比本文算法得到结果见表 1.

由表 1 可知，在人工数据集上，本文算法没有错检和漏检，具有较高的准确性.

3.3 燃煤机组烟尘超低排放浓度数据集实验结果

选取一台装机容量为 330 MW 的燃煤机组(烟尘超低排放工艺为：干式除尘器＋脱硫后湿式除尘器)进行算法验证.以该机组 2015 年 7 月至 2016 年 6 月的烟尘超低排放浓度数据作为训练样本，2016 年 6 月 30 日当天的数据作为测试样本，辨识烟尘排放浓度序列中的异常值.图 3 为待检测数据示意图.

表 1 人工数据集不同算法结果比较

Tab.1 Results comparison of synthetic data

检测算法	异常值数	检测出的异常值数	错检数	漏检数
基于贝叶斯后验的异常值检测	6	0	0	0
基于 AR 残差的异常值检测	6	12	6	0
基于残差后验的异常值检测	6	7	1	0

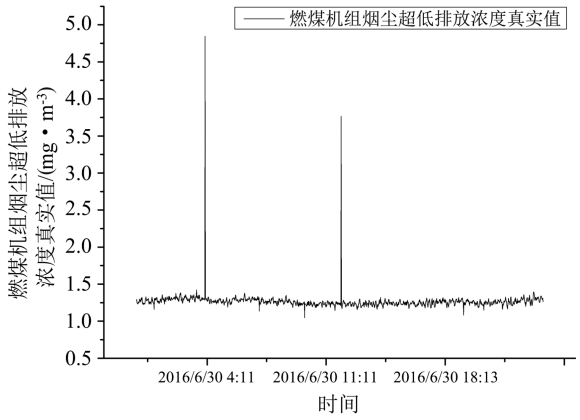

图 3 烟尘排放浓度数据集

Fig.3 Soot emission concentration

给定的烟尘数据整体变化趋势较为平缓，但有两处较为明显的突变.经过甄别，确定满足变化率绝对值大于 80%，即

$$\gamma=\left|\frac{x_t-x_{t-1}}{x_{t-1}}\right|\times 100\%\geqslant 80\% \qquad (23)$$

测试样本中两个该类异常点分别位于 2016/6/30 12:06:00(变化率绝对值为 208.50%)、2016/6/30 4:02:00(变化率绝对值为 270.51%).

使用基于贝叶斯后验的异常值在线检测及置信度评估算法进行实验，设置参数同 3.2 节，得到测试机组烟尘排放浓度数据后验概率比如图 4 所示.

从图 4 可以看出，仅两处明显突变的贝叶斯后验概率比低于阈值 0.8，查询可知确为 2016/6/30 4:02 及 2016/6/30 12:06 两个数据点，与人工甄别结果相符.

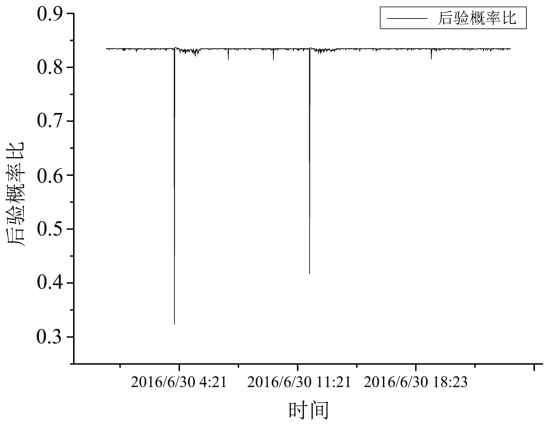

图 4　烟尘排放浓度数据后验概率比

Fig.4　Posterior ratio of soot emission concentration

为了评估所检测异常值结果的置信度，用相同训练样本进一步进行 SOM 建模，得到状态转移概率结果如表 2 所示.

表 2　烟尘排放浓度数据状态转移概率矩阵

Tab.2　State transition probabilities matrix of soot emission concentration

状态	C_1	C_2	…	C_{22}	…	C_{36}
C_1	0.330	0.000	…	0.002	…	0.000
C_2	0.000	0.571	…	0.000	…	0.003
C_3	0.000	0.180	…	0.000	…	0.004
C_4	0.003	0.000	…	0.003	…	0.000
⋮	⋮	⋮	…	⋮	…	⋮
C_8	0.000	0.015	…	0.000	…	0.138
⋮	⋮	⋮	…	⋮	…	⋮
C_{22}	0.005	0.001	…	0.615	…	0.001
⋮	⋮	⋮	…	⋮	…	⋮
C_{36}	0.000	0.006	…	0.000	…	0.745

对于 2016/6/30 4:02 这一数据点，查询表 2 计算得到其打分函数为

$$s=1-\frac{p_{t-1,t}}{p_{t-1,t}}=1-\frac{p_{22,8}}{p_{22,22}}=1-\frac{0}{0.615}=1.$$

这说明根据历史数据分布，该点发生异常的概率为 100%.同理计算得到 2016/6/30 12:06 数据点为异常的置信度为 91%.

为了降低实验的偶然性，另取 10 台燃煤机组烟尘超低排放浓度数据建模，根据上述步骤在线检测异常值并对异常值作置信度评估，采用式(21)作为真实异常的识别统计量，得到检测结果如表 3 所示.

将此算法在 3.3 测试机组上与基于 AR 残差的异常值检测算法、基于残差后验的异常值检测算法进行对比，得到结果如表 4 所示.由表 4 可知，本文提出的算法无论对人工数据集还是烟尘排放浓度数据集，都表现出了比其他算法更好的效果.这说明该算法对大数据量、更新快速、变化缓慢的工业时间序列具有优秀的处理能力和在线检测性能.

表 3　多台机组平均检测结果

Tab.3　Average detection results of more units

平均准确率(%)	平均召回率(%)	最小准确率(%)	最小召回率(%)
98.22	93.17	85.35	87.15

表 4　烟尘排放浓度数据集不同算法结果比较

Tab.4　Results comparison of soot emission concentration

检测算法	异常值数	检测出的异常值数	错检数	漏检数
基于贝叶斯后验的异常值检测	2	2	0	0
基于 AR 残差的异常值检测	2	5	3	0
基于残差后验的异常值检测	2	3	1	0

综上可知，本文提出的基于贝叶斯后验的异常值检测算法能够有效地对时间序列数据中的异常值进行在线快速检测，并且具有较高的准确率.

4　结论

本文提出了一种基于贝叶斯后验的异常值在线检测及置信度评估算法.首先，利用自回归模型实现对时间序列数据的预测，用贝叶斯后验概率比对预测残差进行异常值检测；然后，利用自组织映射状态模型对检测出的异常值进行可靠性评估，检验异常值判断的置信度.实验证明，该算法能够有效地对更新速度快、变化趋势平缓、人工类标不足的大数据量工业时间序列数据进行异常值在线辨识.

本文提出的算法依次对每一维度进行在线检测，不仅能识别各维度中存在的数据异常，还能识别各维度之间可能存在的关系异常.对检测结果进行置信度评估，进一步提高了准确性，能够有效减少误报.此外，由于无需用过多先验知识的核密度估计进行概率估计，不要求数据分布满足正态性，也无需事先给出数据异常标记，降低了样本预处理的难度和工作量，提升了算法的通用性.由于自组织映射神经网络可以实现高维数据到低维数据的转换，满足参数繁多、记录数庞大的工业高维数据处理场景，使算

法具有较好的适用性和扩展性.同时,由于实现了在线检测功能,并提供了模型批量更新算法,模型能及时根据数据实际变化情况进行调整,进一步提高了检测的准确性、可靠性和适应性.总体上说,这种算法有着广阔的应用前景.

参考文献(References)

[1] PAWAR A D, KALAVADEKAR P N, TAMBE S N. A survey on outlier detection techniques for credit card fraud detection [J]. IOSR Journal of Computer Engineering, 2014, 16(2): 44-48.

[2] GOLMOHAMMADI K, ZAIANE O R. Time series contextual anomaly detection for detecting market manipulation in stock market[C]// IEEE International Conference on Data Science and Advanced Analytics. Pairs, France: IEEE Press, 2015: 1-10.

[3] KIM G, LEE S, KIM S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection [J]. Expert Systems with Applications, 2014, 41(4): 1690-1700.

[4] SCHIFF G D, VOLK L A, VOLODARSKAYA M, et al. Screening for medication errors using an outlier detection system[J]. Journal of the American Medical Informatics Association, 2017, 24(2): 281-287.

[5] BILLOR N, HADI A S, VELLEMAN P F. BACON: Blocked adaptive computationally efficient outlier nominators [J]. Computational Statistics & Data Analysis, 2000, 34(3):279-298.

[6] KNORR E M, NG R T. Algorithms for mining distance-based outliers in large datasets [C]// Proceedings of the 24th International Conference on Very Large Data Bases. San Francisco: Morgan Kaufmann Publishers,1998: 392-403.

[7] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets [C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, USA: ACM Press, 2000, 29(2): 427-438.

[8] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: Identifying density-based local outliers [C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, USA: ACM Press, 2000, 29(2): 93-104.

[9] HAWKINS D M. Identification of Outliers [M]. London: Chapman and Hall, 1980.

[10] ABRAHAM B, BOX G E P. Bayesian Analysis of Some Outlier Problems in Time Series[J]. Biometrika, 1979, 66(2):229-236.

[11] KARIMIAN S H, KELARESTAGHI M, HASHEMI S. I-IncLOF: Improved incremental local outlier detection for data streams [C]// Proceedings of the 16th CSI International Symposium on Artificial Intelligence and Signal Processing. Shiraz, Fars, Iran: IEEE Press, 2012: 23-28.

[12] 潘渊洋, 李光辉, 徐勇军. 基于DBSCAN的环境传感器网络异常数据检测方法[J]. 计算机应用与软件, 2012(11): 69-72.

[13] HILL D J, MINSKER B S, AMIR E. Real-time Bayesian anomaly detection for environmental sensor data [C]// Proceedings of the 32nd Congress-International Association for Hydraulic Research. 2007, (2): 503.

[14] ERFANI S M, RAJASEGARAR S, KARUNASEKERA S, et al. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning [J]. Pattern Recognition, 2016, 58(C): 121-134.

[15] JADIDI Z, MUTHUKKUMARASAMY V, SITHIRASENAN E, et al. Flow-based anomaly detection using neural network optimized with GSA algorithm[C]// Proceedings of the 33rd International Conference on Distributed Computing Systems Workshops. Philadelphia, USA: IEEE Press, 2013: 76-81.

[16] MARTINS H, PALMA L, CARDOSO A, et al. A support vector machine based technique for online detection of outliers in transient time series[C]// 10th Asian Control Conference. Kota, Kinabalu: IEEE Press, 2015: 1-6.

[17] JOHANSEN S, NIELSEN B. Asymptotic theory of outlier detection algorithms for linear time series regression models [J]. Scandinavian Journal of Statistics, 2016, 43(2): 321-348.

[18] 刘芳, 毛志忠. 过程控制时间序列中异常值的动态检测[J]. 控制理论与应用, 2012, 29(4): 424-432.

[19] 杨志勇, 朱跃龙, 万定生. 基于知识粒度的时间序列异常检测研究[J]. 计算机技术与发展, 2016, 26(7): 51-54.

[20] BOX G E P, JENKINS G M, REINSEL G C, et al. Time Series Analysis: Forecasting and Control [M]. John Wiley & Sons, 2015.

[21] LACOUR C, MASSART P, RIVOIRARD V. Estimator selection: A new method with applications to kernel density estimation[J]. arXiv preprint, 2016, arXiv:1607.05091.

[22] ANDERSSON B, DAVIER A A. Improving the bandwidth selection in kernel equating[J]. Journal of Educational Measurement, 2014, 51(3): 223-238.

[23] 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.

[24] GUIDO D, TEUVO K.Visual Explorations in Finance: With Self-Organizing Maps [M]. Springer Science & Business Media, 2013.

[25] TAKEUCHI J I, YAMANISHI K. A unifying framework for detecting outliers and change points from time series [J]. Journal of Taiyuan Normal University, 2006, 18(4): 482-492.