基于残差统计的时间序列加性离群点检测算法研究

张 玲,刘 波
(国家数字交换系统工程技术研究中心,北京 100094)

摘 要:针对时间序列,提出了一种基于残差统计的加性离群点检测算法,利用 AR 模型对时间序列进行前向与后向拟合;采用了数据相对变化率判别法减少离群点对拟合的影响;根据假设检验原理,以高斯分布统计检验对残差进行统计分析并最终确定离群点。仿真结果表明,该方法对离群点检测有较高的准确性。

关键词:时间序列;离群点;AR 模型;高斯分布
中图分类号:TP311.11 文献标识码:A DOI:10.16157/j.issn.0258-7998.2015.09.023

中文引用格式:张玲,刘波. 基于残差统计的时间序列加性离群点检测算法研究[J].电子技术应用,2015,41(9):85-87,91.
英文引用格式:Zhang Ling,Liu Bo. Residuals statistics-based additive outlier detection algorithm for time series[J].Application of Electronic Technique,2015,41(9):85-87,91.

Residuals statistics-based additive outlier detection algorithm for time series

Zhang Ling,Liu Bo
(China National Digital Switching System Engineering and Technological Research Center,Beijing 100094,China)

Abstract: We propose a residuals statistics-based additive outlier detection algorithm for one-dimensional time series, The basic idea is using time series AR model for forward and backward fitting. In order to reduce the influence of outlier, we use data's relative change rate to preliminary judge the outlier. According to hypothesis testing theory and Gauss distribution statistic testing, we find out the outliers. The simulation results show that the this method has good performance on outlier detection.
Key words: time series; outlier; AR model; Gauss distribution

0 引言

在时间序列数据挖掘中,不可避免地存在一些远离序列一般水平的极端大值和极端小值,或者与其他序列样本点一般行为或特征不一致的点值,这些点被称做离群点。离群点的产生可能是采样中的误差,也可能是被研究对象本身由于受各种偶然非正常的因素影响而引起的。一方面,离群点的存在会影响时间序列模式表示,可能使数据挖掘陷入混乱,导致在随后的数据处理过程中产生偏差或误导;另一方面,离群点可以提供一些潜在的重要信息。目前,时间序列离群点检测作为对数据进行挖掘处理的第一步,已经成为该研究领域的重要方向之一,并广泛应用于通信流量监测、工业故障诊断、金融贸易等方面。

时间序列中的离群点有很多类型,按照出现的个数,可以分为孤立离群点和成片离群点,按照产生的影响可以分为加性离群点 AO(Additive Outlier)、更新离群点 IO(Innovational Outlier)、水平移位离群点 LS(Level Shift Outlier)和暂时变更离群点 TC(Temporary Change Outlier)[1]。本文主要对时间序列中的加性离群点检测方法进行研究,并在此基础上提出了一种基于残差统计的检测方法,仿真结果表明该方法在检测加性离群点方面具有较好的性能。

1 离群点检测方法研究

针对无序的数据集,离群点检测方法主要有基于统计的方法、基于距离的方法[4]、基于密度的方法[5]和基于偏离的方法。近年来,不少研究人员提出了专门针对时间序列的离群点检验算法,主要有统计诊断方法、贝叶斯方法、遗传算法、人工神经网络、小波检测等。国内也有相关人员对此做了深入的研究[2-5]。文献[6]提出了基于粗糙集理论的序列离群点检测方法,它利用粗糙集理论中的知识熵和属性重要性等概念来构建三种类型的序列,并通过分析序列中元素的变化情况来检测离群点。文献[7]通过建立多变量时间序列数据相似度矩阵,对相似度矩阵进行转换以最大化数据之间的相关性,并采用随机游走模型计算数据点之间的连接系数来检测数据点上的异常。文献[8]指出离群点与它所在时间段内的其他数据不具有相似性,从时序图上看,离群点相对于它相邻区域内的数据具有很强的跳跃性,进而提出基于数据相对变化率的时间序列离群点识别方法。

2 基于残差统计的加性离群点检测算法

2.1 问题提出

对于时间序列,离群点可能会隐藏在时间序列的趋势、季节或其他变化中,增加了检测难度。以图 1 所示的

时间序列为例，两个时间序列都处于上升趋势，A 点明显偏离了整个趋势，应判定为离群点；B 点虽然与前向时刻点在幅度变化率上发生了较大变化，但符合后向时刻点的变化趋势，是一个正常时间序列点，因此不应判定为离群点。

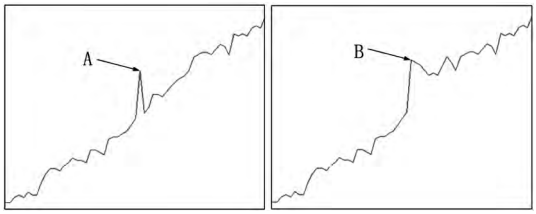

图 1 受加性离群点“干扰”的时间序列与正常时间序列

本文以一维时间序列为研究对象，提出了一种基于残差统计的加性离群点检测算法，基本思想是利用 p 阶 AR 模型对时间序列进行前向与后向拟合，得到每个时间点拟合残差。采用了邻域区间变化率判别法对离群点进行初判，初判的疑似离群点不参与拟合运算。最后根据高斯分布假设检验的方法对残差进行统计分析并最终确定离群点。

定义待检测时间序列数据样本为 $x_t, t=1,2,3,4\cdots M, x_t \in R$，并做如下假设：

(1)离群点随机分布；

(2)正常数据的数量远大于离群点数量。

2.2 算法描述

2.2.1 邻域区间变化率

定义 1 邻域区间变化率：时间序列各时刻点与相邻前后时刻的幅度变化率。设时刻 t 的邻域区间变化率为 δ_t，则：

$$\delta_t=|(x_t-x_{t-1})+(x_t-x_{t+1})|$$

对所有 δ_t 进行考虑，选定门限 δ，δ 值的计算可以采用平均法或加权计算等。若 $\delta_t>\delta$，则将 x_t 标志为 LK 点（疑似离群点），否则标志为 uLK 点（非疑似离群点）。

离群点相对于它前后相邻数据都会有较大变化，因此邻域区间变化率要同时对前向时刻和后向时刻进行考虑。定义 LK 点和 uLK 点是为了在拟合过程中尽量减少离群点的影响，对疑似离群点不作拟合参考。

2.2.2 AR 模型拟合与参数计算

拟合常用的模型有 AR 模型、MA 模型、ARIMA 模型等。AR 模型一般用于拟合平稳的时间序列，而时间序列从局部来看近似一个平稳的过程，并且 AR 模型结构相对简单，拟合精度较高，因此本文选用 p 阶自回归 AR 模型。为了准确反应各检测点的局部变化属性，并减少离群点对参数估计的影响，本文在文献[9]所采用的两窗口模型基础上，提出了改进的窗口计算模型，基本原理是：检测窗口仅包含 t 时刻待检测点，前向学习窗口和后向学习窗口位于检测窗口邻近两侧，宽度为 N，并且 $N>p$，根据前向和后向学习窗口中的数据分别对 t 时刻待检测点进行前向和后向拟合，采用剪枝思想，若学习窗口中包含疑似离群点 LK，则该点退出学习窗口不参与计算，其余时间轴上的 uLK 点向 t 时刻整体移位并填满窗口。如图 2 所示。

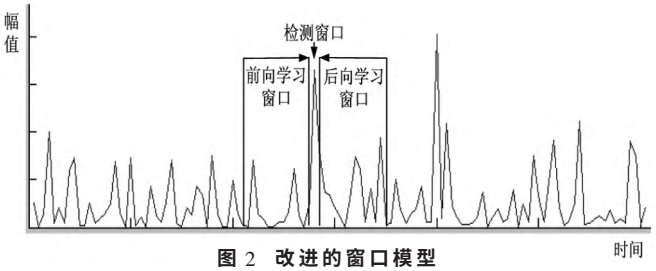

图 2 改进的窗口模型

前向拟合得到 t 时刻前向拟合残差 $\varepsilon 1_t$ 为：

$$\varepsilon 1_t=x_t-\hat{x}_t, t \in Z$$

$$\hat{x}_t=\sum_{j=1}^{p}\alpha_j x_{t-j}$$

后向拟合得到 t 时刻后向拟合残差 $\varepsilon 2_t$ 为：

$$\varepsilon 2_t=x_t-\hat{x}_t, t \in Z$$

$$\hat{x}_t=\sum_{j=1}^{p}\beta_j x_{t+j}$$

其中 $\varepsilon 1_t$ 和 $\varepsilon 2_t$ 服从 $N(0,\sigma^2)$，$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_p)$ 为前向自回归系数，$\beta=(\beta_1,\beta_2,\cdots,\beta_p)$ 为后向自回归系数。最后得到时刻 t 的拟合残差：$\varepsilon_t=\varepsilon 1_t+\varepsilon 2_t$。

在计算残差之前，首先要对自回归系数进行估计。AR 模型的自回归系数在预测误差功率最小条件下满足 Yule-Walker 方程[10]，以前向自回归系数 $(\alpha_1,\alpha_2,\cdots,\alpha_p)$ 为例：

$$\begin{bmatrix} r_x(0) & r_x(1) & \cdots & r_x(p-1) \\ r_x(1) & r_x(0) & \cdots & r_x(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(p-1) & r_x(p-2) & \cdots & r_x(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} r_x(1) \\ r_x(2) \\ \vdots \\ r_x(p) \end{bmatrix}$$

采用自相关法，根据 t 时刻前向窗口内的观测数据样本 $x_{t-1}, x_{t-2}, \cdots, x_{t-N}$ 计算自相关函数 $r_x(0), r_x(1), \cdots, r_x(p)$ 估计值，窗口外的计算样本值假设为 0，自相关函数计算如下：

$$\hat{r}_x(k)=\frac{1}{N}\sum_{j=0}^{N-1} x_{t-N+j}\times x_{t-N+j+k}, k=0,1,2,\cdots,p$$

上述线性方程的求解按如下形式：

$$\begin{bmatrix} \hat{r}_x(0) & \hat{r}_x(1) & \cdots & \hat{r}_x(p-1) \\ \hat{r}_x(1) & \hat{r}_x(0) & \cdots & \hat{r}_x(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_x(p) & \hat{r}_x(p-2) & \cdots & \hat{r}_x(0) \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_p \end{bmatrix} = \begin{bmatrix} \hat{r}_x(1) \\ \hat{r}_x(2) \\ \vdots \\ \hat{r}_x(p) \end{bmatrix}$$

后向自回归系数 $(\beta_1,\beta_2,\cdots,\beta_p)$ 的计算同上，其自相关函数计算为：

$$\hat{r}_x(k)=\frac{1}{N}\sum_{j=0}^{N} x_{t+j}\times x_{t+j+k}, k=0,1,2,\cdots,p$$

2.2.3 高斯统计检测

基于假设检验理论，在一定的显著性水平下，拟合

残差 ε_t 近似服从高斯分布，即 $\varepsilon \sim N(u,\sigma^2)$。并且在假设2前提下，高斯分布作为残差统计模型对离群点判决同样具有较高置信度。在此，选择高斯分布做为统计模型，ε_t 的概率密度为：

$$f(\varepsilon)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{1}{2\sigma^2}(\varepsilon-u)^2\right]$$

(u,σ^2) 采用最大似然估计，似然函数为：

$$L(u,\sigma^2)=\prod_{i=1}^{M}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{1}{2\sigma^2}(\varepsilon_i-u)^2\right]$$

$$=(2\pi)^{-M/2}(\sigma^2)^{-M/2}\exp\left[-\frac{1}{2\sigma^2}\prod_{i=1}^{M}(\varepsilon_i-u)^2\right]$$

分别令：$\begin{cases}\dfrac{\partial}{\partial u}\ln L=0\\ \dfrac{\partial}{\partial \sigma^2}\ln L=0\end{cases}$

得到 (u,σ^2) 的最大似然估计值为：$\hat{u}=(1/M)\sum_{i=1}^{M}\varepsilon_i=\bar{\varepsilon}$，$\widehat{\sigma^2}=(1/M)\sum_{i=1}^{M}(\varepsilon_i-\bar{\varepsilon})^2$。

计算时间序列每个样本点的似然残差概率分布 $f(\varepsilon_i)$，选定一个显著水平上的临界值 F，对 x_i 是否为离群点做出决策：

$$x_i\begin{cases}f(\varepsilon_i)<F, \text{离群点}\\ \text{else}, \text{正常序列点}\end{cases}$$

3 仿真

为了验证本文所提算法的有效性，以局域网内某主机通信流量监测数据为对象进行测试。通信流量监测是网络管理的重要内容，通过流量监测，可以全面透视网络的流量控制，快速定位和发现网络故障，并保障关键应用的稳定运行，减少泄密风险。一般情况下，主机通信流量的具体业务包括 Web、Telnet、SNMP、请求应答数据包等，在仿真实验中，通过随机加入异常事件，比如网络拥塞、数据分发等来模拟加性离群点。

图 3 所示为某日上午 8:00-12:00 的某主机通信流量监测数据，单位为 KB/min，数据样本 200 个，离群点 5 个。窗口宽度取 15，模型阶数取 4，拟合残差分布情况如图 4 所示。由图看出，拟合后，离群点的残差值与正常的浮动范围相比有较大偏移。

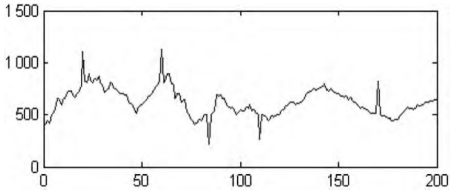

图 3 加入 AO 的通信流量监测数据

为了验证算法对离群点数量的鲁棒性，在 200 个流量监测数据样本点中分别随机加入 5、10、15、20 个离群点，拟合计算的窗口宽度取 15，模型阶数取 4，概率判决临界值分别取 0.95、0.95、0.9、0.9。在仿真测试中并未使用离群点数量先验知识。在此定义两个检测指标：

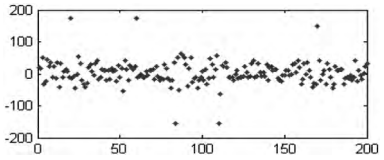

图 4 拟合残差

检出率：检测出的真实离群点数量与实际离群点数量之比。

误检率：检测出的错误离群点数量与实际离群点数量之比。

检测统计结果如表 1 所示。结果显示，当实际离群点数量在样本中的比重小于 0.05 时，算法能对离群点进行完全有效地检测，当实际离群点数量在样本中的比重大于 0.1 时，检出率下降，误检率有所上升，但此时离群点的发生不再是小概率事件，根据加性离群点对时间序列产生的影响上看，它不符合加性离群点特征。因此，本文所提算法对检测时间序列中的加性离群点有较好的性能，同时，在实际应用中证明该算法对其他类型离群点的检测也有一定的鲁棒性。

表 1 不同离群点数量下算法有效性检测

离群点数量	检出率	误检率
5	1	0
10	1	0
15	0.86	0.07
20	0.85	0.1

4 结论

本文针对时间序列中的加性离群点检测，提出了一种基于残差统计的检测算法。该算法利用 AR 模型计算每个样本点拟合残差，通过统计分析残差的概率分布来判别离群点。通过对局域网某主机通信流量监测数据的仿真结果显示，该算法在检测加性离群点方面是有效的，结果有较高的置信度。此外，在对拟合残差进行分析时，除了本文采用的统计模型方法外，还可以采用基于密度的聚类的方法。另外如何检测时间序列中其他类型的离群点也是值得研究的内容。

参考文献

[1] 胡云，王崇骏，谢俊元，等.社群演化的隐健迁移估计及演化离群点检测[J].软件学报，2013，24(11)：2710-2720.

[2] Hu Tianming，Sung Sam Yuan.A trimmed mean approach to finding spatial outliers[J].Intelligent Data Analysis，2004，8(1)：79-95.

[3] ALARCON-AQUINO V，BARRIA J A.Anomaly detection in communication networks using wavelets[J].Communications，IEEE，2001，148(6)：355-362.

[4] 刘耀宗，张宏，孟锦，等.基于小波密度估计的数据流离群点检测[J].计算机工程，2013，39(2)：178-181.

[5] 江峰，杜军威，葛艳，等.基于粗糙集理论的序列离群点检测[J].电子学报，2011(2)：345-350.

[6] 李权，周兴社.一种新的多变量时间序列数据异常检测方法[J].时间频率学报，2011，34(2)：154-158.

(下转第 91 页)

参考文献

[1] ERKIP S E, AAZHANG B. User cooperation diversity-part I: system description[J]. IEEE Trans. Communication, 2003, 51(11): 1927-1938.

[2] LANEMAN J N, TSE D N C, WORNELL G W. Cooperative diversity in wireless networks: efficient protocols and outage behavior[J]. IEEE Trans. Information Theory, 2004, 50(12): 3062-3080.

[3] SU W, SADEK A K, LIU R J K. SER performance analysis and optimum power allocation for decode-and-forward cooperation protocol in wireless networks[C]. In Proc. IEEE WCNC, 2005, 2: 984-989.

[4] MENG Y, JING L, SADIADPOUR H. Amplify-forward and decode-forward: the impact of location and capacity contour[C]. Military Communications Conference, 2005, 11(3): 1609-1615.

[5] Xu Lei, Zhang Hongwei, Li Xiaohui, et al. Optimum relay location in cooperative communication networks with single AF relay[J]. Communications, Network and System Sciences, 2011, 4: 147-151.

[6] GURRALA K K, DAS S. Impact of relay location on the performance of multi-relay cooperative communication[C]. IJCNWC, 2012, 2(2): 2250-3501.

[7] LIN F, LI Q H, LUO T, et al. Impact of relay location according to SER for amplify-and-forward cooperative communications[C]. 2007 IEEE International Workshop on Anti-counterfeiting, Security, Identification. Xiamen, 2007, 4.

[8] XUE K, HONG X, CHEN L, et al. Performance analysis and resource allocation of heterogeneous cognitive gaussian relay channels[C]. Global Communications Conference (GLOBECOM), 2013 IEEE, 2013: 1167-1172.

[9] TERA A D, GURRALA K K, DAS S. Power allocation for AF cooperative relaying using particle swarm optimization[C]. In Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference on. IEEE, 2014, 3.

(**收稿日期**: 2015-03-21)

作者简介:

刘钧彬(1982-), **男**, **讲师**, **博士研究生**, **主要研究方向**: **无线传感器网络**、**通信与信号处理**。

丁凡(1982-), **男**, **讲师**, **博士研究生**, **主要研究方向**: **无线传感器网络**、**通信与信号处理**。

(**上接第** 84 **页**)

Remote Sensing Letters, 2007, 4(4): 659-663.

[5] PIERCCINI M, LUZI G, ATZENI C. Terrain mapping by ground-based interferometric radar[J]. IEEE Transactions on Geoscience and Remote Sensing, 2001, 39(10): 2176-2181.

[6] RODELSPERGER S, BECKER M, GERSTENECKER C, et al. Digital elevation model with the ground-based SAR IBIS-L as basis for volcanic deformation monitoring[J]. Journal of Geodynamics, 2010(49): 241-246.

[7] NOON D, HARRIES N. Slope stability radar for managing rock fall risks in open cut mines[C]. Proceedings of the 3rd CANUS Rock Mechanics Symposium, 2007.

[8] LU B, ZHANG X, SONG Q, et al. A vehicle based SFCW SAR for differential interferometry[C]. Proceedings of the APSAR 2011, 2011: 691-694.

[9] YANG X L, WANG Y P, QI Y L, et al. Experiment study on deformation monitoring using ground-based SAR[C]. Apsar, 2013: 285-288.

(**收稿日期**: 2015-05-18)

作者简介:

蔡永俊(1989-), **男**, **博士研究生**, **主要研究方向**: **合成孔径雷达信号处理与系统研究**、**全极化合成孔径雷达信息处理等**。

张祥坤(1972-), **男**, **研究员**, **主要研究方向**: **合成孔径雷达信号处理与系统研究**、**微波遥感理论与技术等**。

姜景山(1936-), **男**, **研究员**, **博士生导师**, **中国工程院院士**, **主要研究方向**: **微波遥感理论与技术研究**、**机载遥感信息实时传输**。

(**上接第** 87 **页**)

[7] **周勇**. **时间序列时序关联规则挖掘研究**[D]. **成都**: **西南财经大学**, 2008.

[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(9): 2080-2087.

[9] **皇甫堪**, **陈建文**, **楼生强**. **现代数字信号处理**[M]. **北京**: **电子工业出版社**, 2003.

[10] **薛安荣**, **鞠时光**, **何伟华**, **等**. **局部离群点挖掘算法研究**[J]. **计算机学报**, 2007(8): 1455-1463.

(**收稿日期**: 2014-03-21)

作者简介:

张玲 (1976-), **国家数字交换系统工程技术研究中心**, **高级工程师**, **博士生**, **主要研究方向**: **关联规则挖掘**。

刘波 (1982-), **国家数字交换系统工程技术研究中心**, **工程师**, **主要研究方向**: **网络入侵检测**。