

一种基于聚类分析的 3MAD – MMMD 过失误差侦破方法

肖冬, 包晶晶

(东北大学 信息科学与工程学院, 辽宁 沈阳 110819)

摘 要: 软测量建模时所使用的数据集中若含有过失误差, 将在很大程度上影响所建模型的精确度. 因此, 在建模之前, 针对建模所使用的数据集, 提出了基于聚类分析的集成 3MAD – MMMD 过失误差的侦破方法. 在采集无缝钢管穿孔过程中不同变量不同时刻的数据后, 将其排列成数据矩阵. 首先运用 3MAD 算法剔除其中的单变量大误差得到新的数据矩阵, 之后采用欧氏距离公式求得新矩阵中同一变量的数据到其最近点的距离, 最后以所有变量最近距离的中位值 d_{med} 为检测标准, 对新的数据矩阵进行过失误差侦破处理. 实验和仿真图表明, 3MAD – MMMD 侦破方法有效地剔除了采集数据中的过失误差.

关 键 词: 软测量建模; 过失误差; 聚类分析; 3MAD; MMMD

中图分类号: TP 273

文献标志码: A

文章编号: 1005-3026(2015)08-1089-04

Detection of Gross Error Using 3MAD-MMMD Based on Cluster Analysis

XIAO Dong, BAO Jing-jing

(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: XIAO Dong, associate professor, E-mail: xiaodong@ise.neu.edu.cn)

Abstract: If there exist gross errors in the soft sensor modeling data, the accuracy of the model is largely affected. Therefore, for the data set to be used in the modeling process, a method of gross error detection of 3MAD-MMMD based on cluster analysis is proposed to process the data before modeling. The data of different variables in different time from the seamless pipe perforation process is collected. Then these data are arranged into a matrix. The 3MAD algorithm is used first to eliminate the large error of single-variables to get the new data matrix. Based on the Euclidean distance formula, the distance is then obtained from the data in matrix to another which is closest to it of the same variable. Finally, d_{med} , the median value of all variables' closest distance, is treated as testing standards to detect gross error of new data matrix. It can be seen from the experimental and simulation results that the gross errors in the collected data sets are effectively eliminated in the 3MAD-MMMD detection method.

Key words: soft sensor modeling; gross error; cluster analysis; 3 median absolute deviation; modified median minimum distance

目前我国工业企业的产品质量检测工作大部分依赖定时、离线的采样分析. 生产操作几乎完全依靠经验和感官知觉, 从而造成产品质量不稳定, 收益较低. 软测量技术^[1]被认为是解决这一问题的卓有成效的方法. 软测量技术的核心^[2]是建立工业对象的模型, 而初始模型是对过程变量的历

史数据进行辨识得到的; 因此建模所需数据精确与否, 在很大程度上决定了软测量建模的有效性和准确性. 为了保证模型的精度, 在建立软测量模型之前, 通过某种方法将真实信号从受误差影响的混合信号中分离出来, 这个过程称为过失误差侦破^[3].

收稿日期: 2014-07-08

基金项目: 国家自然科学基金资助项目(61203214); 辽宁省教育厅科学研究一般项目(L2013101).

作者简介: 肖冬(1978-), 男, 湖南涟源人, 东北大学副教授, 博士.

在实际工业过程中,数据采集过程无法精确到足以发现所有误差.其中,误差分为两种:随机误差和过失误差^[4-5].通常,随机误差对建模的影响较小且不能避免,只能尽量减小;而过失误差往往对软测量建模的影响很大,应予以剔除.引起过失误差的原因主要有:仪器的零点漂移,仪器操作故障,操作不稳定等.过失误差会对建模带来严重干扰,因此在建立软测量模型之前,需要将过失误差数据从模型数据中侦测并分离出来,这对成功建立精确的软测量模型非常重要.

至今,国内外就如何剔除工业过程中的过失误差数据已研究出一些较为完善的方法,主要是基于统计假设检验的检测方法,如整体检验法、测量残差检验法、节点检验法以及主成分检验法等.其中,整体检验法和节点检验法简单易行,可以有效地判定过失误差的存在,但无法定位;测量残差检验法可以侦破并确定过失误差的位置,但其在识别过程中依据的是可能会产生干扰的协调数据,容易出现“误判”;PCA 主元分析法^[6-7]考虑了误差的相关性,将相关变量转化成不相关的变量,并且可以有效地侦破和识别幅度较小的过失误差,但是需要建立相关模型,计算繁琐.而最近的主流方法 MMMD(modified median minimum distance)聚类分析法^[8]不需要事先通过对学习样本的学习建立数学模型,进而再通过训练来识别和检测过失误差,而是可以直接对工业过程中现场采集的数据进行分类,将显著误差剔除,以用于之后的软测量数据建模.

1 3MAD 方法简介

在数据向量 \mathbf{x} 中,满足式(1)的 x_i 为发生过失误差的数据点.

$$\left. \begin{aligned} |x_i - x_{\text{med}}| > 3 \cdot S_{\text{MAD}}, \\ S_{\text{MAD}} = 1.4826 \text{med}(|x_i - x_{\text{med}}|) \end{aligned} \right\} \quad (1)$$

式中: x_i 为数据向量 \mathbf{x} 中的数据点, $i = 1, 2, \dots, n$; x_{med} 为数据的中位数; $\text{med}(\cdot)$ 表示求取数据的中位值.

上述方法中,当数据向量 \mathbf{x} 服从正态分布时, S_{MAD} 是标准差的无偏估计.当存在大误差时, x_{med} 比平均值更能反映数据的中心位置.

2 MMMD 方法简介

基于平均最小距离的聚类算法是一种行之有效的过失误差侦破方法.该方法根据相似性的度

量方法,把原始数据聚集成不同的数据类,这样就能方便地把异类点和主体数据分开,从而实现过失误差侦破.将采集到的同一变量不同时刻的数据写为列向量,并将不同变量的列向量组成数据矩阵. MMMD 方法考虑数据矩阵中的所有数据,以列向量为单位,求得列向量中每个数据与同一向量中其他数据的欧氏距离,并取其最小值 l_i . 求数据矩阵所有列向量最小距离 (l_1, l_2, \dots, l_N) 的中位值记为 l_{med} ,以此中位值作为数据状态的判断准则.具体方法如下.

在 d 维空间中,给定具有 N 个数据对象的集合 X_1, X_2, \dots, X_N ,写成数据矩阵(2)的形式:

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2 \quad \cdots \quad \mathbf{X}_N] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dN} \end{bmatrix} \quad (2)$$

运用欧几里得算法计算数据之间的距离,计算公式为式(3)和式(4).

矩阵中任意列向量表示为 $\mathbf{X}_i = [x_{1i}, x_{2i}, \dots, x_{di}]$ ($i = 1, 2, \dots, N$). \mathbf{X}_i 中每个数据到同一列向量中其他数据的距离用 d_{ji} 表示:

$$d_{ji} = \left(\sum_{\substack{k=1 \\ k \neq j}}^d (x_{ji} - x_{ki})^2 \right)^{1/2} \quad (3)$$

$$j = 1, 2, \dots, d; i = 1, 2, \dots, N.$$

式中 x_{ji}, x_{ki} 表示同一列向量中不同的数据.

取每列向量中所求各距离的最小值,用 l_i 表示:

$$l_i = \min_{j=1}^d (d_{ji}), (i = 1, 2, \dots, N) \quad (4)$$

由式(4)求得最小距离向量为 $\mathbf{l} = [l_1 \quad l_2 \cdots l_N]$.取向量 \mathbf{l} 中各最小距离的中位值,用 l_{med} 来表示.若 $d_{ji} > l_{\text{med}}$,则原数据矩阵中对应的数据为过失误差数据;若 $d_{ji} < l_{\text{med}}$,则为正常数据.以上算法称为 MMMD 聚类算法.

3 3MAD - MMMD 过失误差侦破方法

1) MMMD 方法的具体步骤:

①输入样本数据矩阵 $\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2 \cdots \mathbf{X}_N]$.

②运用式(3),式(4)求最小距离向量 $\mathbf{l} = [l_1 \quad l_2 \cdots l_N]$.

③按从小到大顺序排列欧几里得公式所求的最小距离,得到新的向量 $\mathbf{l} = [l_1, l_2, \dots, l_N]$.

④根据步骤③所得到的最小距离的排列结

果,得到其中位值 l_{med} .

⑤定义 l_{med} 作为分界点,根据数据点到数据中心的距离,将数据集分为两类:当 $d_{ji} > l_{med}$, \mathbf{X} 为过失误差类;当 $d_{ji} < l_{med}$, \mathbf{X} 为正确数据类.

⑥结束过失误差侦破.

2) 3MAD - MMMD 方法的具体步骤.

假设软测量建模原始数据矩阵为 $\mathbf{X}_{M \times N}$, M 代表测量采样次数, N 代表测量变量个数.

①采用 3MAD 方法,利用主元分析式对 \mathbf{X} 的每一列进行单变量的过失误差侦破,只要有一个变量在 t 时刻发生了过失误差,就把 t 时刻的数据从建模样本中剔除,即 x_{ij} 的第 i 个元素超出了误差边界,将其整行剔除,最后得 $\mathbf{X}_{M_1 \times N}$.

②采用 MMMD 方法对 $\mathbf{X}_{M_1 \times N}$ 进行误差侦破,将发生过失误差的数据剔除,得到用于建模的正常数据 $\mathbf{X}_{M_2 \times N}$.

3) 3MAD - MMMD 方法的特点:

①首先用 3MAD 法对数据各个变量进行单变量过失误差侦破,避免大的单变量误差对之后的 MMMD 方法产生影响.

②作为聚类算法,MMMD 算法可以直接面对数据,不必考虑复杂的建模过程,这样作为数据的预处理步骤,可以减小数据错误处理的概率.

③利用 3MAD 法和 MMMD 法进行过失误差侦破,无论是单变量大误差还是不同变量中的过失误差都得以有效剔除.

4 3MAD - MMMD 过失误差侦破方法在导盘转速测量上的应用

无缝钢管穿孔过程是热轧无缝钢管变形的首道工序:将经过加热的管坯在穿孔机上穿孔成后壁毛管,导盘转速是其重要参数.管坯穿孔生产过程具有多时段、复杂非线性、动态多变量等特性,导盘转速难以测量,与过程变量之间的关系比较复杂,因此需要建立精度较好的模型来进行预报和控制.

本文选取了该过程中的 400 组数据用于数据侦破.在影响导盘转速的因素中选取了下列参数作为变量,分别对 3MAD 法,MMMD 法和本文的 3MAD - MMMD 算法的过失误差侦破算法进行了侦破仿真.参数如表 1 所示.

图 1 为 3 种方法针对导盘转速计算过程进行的过失误差侦破效果图.

由图 1b 可以看出,MMMD 算法可以检测出采集数据中的大部分过失误差,但对比图 1a 中

表 1 变量表
Table 1 Variable table

编号	变量	变量含义
1	x_1	上辊电流
2	x_2	上辊磁场
3	x_3	上辊压下给定
4	x_4	下辊压下给定
5	x_5	上辊倾角给定
6	x_6	上辊倾角实际值
7	x_7	下辊倾角给定
8	x_8	下辊倾角实际值
9	x_9	右导盘位置给定
10	x_{10}	右导盘位置实际值
11	x_{11}	左导盘位置给定
12	x_{12}	左导盘位置实际值
13	x_{13}	上辊入口侧温度
14	x_{14}	上辊出口侧温度

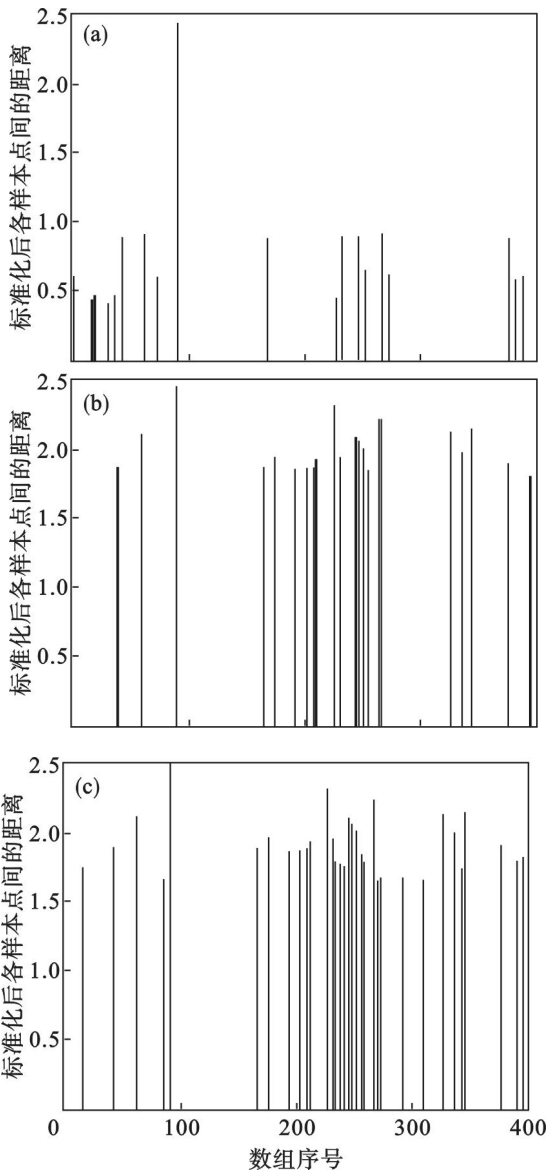


图 1 不同算法的侦破效果图

Fig. 1 Effectiveness of different algorithms

(a)—3MAD; (b)—MMMD; (c)—3MAD - MMMD.

3MAD 算法侦破效果可以看出,单一 MMMD 算法仍然漏报了很多单变量大误差.而图 1c 表明,采用 3MAD – MMMD 聚类新方法则侦破出数据中的所有过失误差.由此可知,新方法优于经典 MMMD 聚类算法.表 2 为 3 种方法仿真后的误差比较,由该表发现,3MAD – MMMD 法较好地融合了两种算法的优点,准确识别并剔除了数据矩阵中的过失误差.

表 2 不同算法的误差比较

Table 2 Comparison of errors with different algorithms

侦破方法	建模误差
3MAD	3.021 3
MMMD	2.986 2
3MAD – MMMD	2.672 6
原始数据	3.148 4

5 结 论

本文针对工业过程中采集数据矩阵存在过失误差的问题,在分析经典 MMMD 算法的基础上,引入了 3MAD – MMMD 新聚类算法.该方法先侦测数据矩阵中的单变量误差,然后再运用 MMMD 算法对新矩阵进行侦测.通过实验仿真图可见,新方法的过失误差侦破效果明显更好;通过误差比较可知其侦破精度高,为后续的软测量建模打下了坚实的基础.

参考文献:

- [1] Tham M A, Montague G A. Soft-sensors for process estimation and inferential control[J]. *Automotive Engineer*, 1991, 1(1): 3 – 14.
- [2] Wei J T, Chen F T, Jiang M L. Application of neural network in fault diagnosis of radar device[J]. *Modern Electronics Technique*, 2012, 12(19): 131 – 134.
- [3] Serth R, Heenan W. Gross error detection and data reconciliation in steam metering systems[J]. *American Institute of Chemical Engineering Journal*, 1986, 32(12): 733 – 742.
- [4] 王建勇,郝勇生,彭兴.基于质量平衡的煤气数据校正应用在能源管控系统中的应用[J]. *冶金自动化*, 2013, 37(2): 138 – 142.
(Wang Jian-yong, Hao Yong-sheng, Peng Xing. The application of the data correction of gas in the energy system of pipeline control based on mass balance[J]. *Metallurgical Industry Automation*, 2013, 37(2): 138 – 142.)
- [5] Zhang Z J, Shao Z J, Chen X. Quasi-weighted least squares estimator for data reconciliation[J]. *Computers and Chemical Engineering*, 2010, 34(2): 154 – 162.
- [6] Yang X S, He X S. Intelligence and smart optimization algorithms[J]. *Basic Sciences Journal of Textile Universities*, 2013, 26(3): 287 – 296.
- [7] Rajab J M, Matjafri M Z, Lim H S. Combining multiple regression and principal component analysis for accurate predictions for column ozone in peninsular Malaysia[J]. *Atmospheric Environment*, 2013, 71(1): 36 – 43.
- [8] Malumfashi A. Education expenditure and economic growth in Nigeria: co-intergration and correction technique[J]. *International Journal of Research in Commerce, Economics and Management*, 2012, 2(8): 34 – 37.