文章编号：1671-7848(2006)04-0298-03

软测量技术的数据预处理方法研究

罗健旭，常 青
(华东理工大学 信息学院，上海 200030)

摘 要：针对软测量技术在线实施时的数据预处理问题，提出了基于聚类分析的过失误差侦破方法。该方法不需过程的先验知识和假设，直接面向数据，可十分方便地在线实现。将该方法与滑动平均滤波算法相结合，可以有效处理过程测量数据的过失误差和随机误差，从而提高软仪表估计的精度。在二元精馏塔底产品组分浓度软测量仪表在线进行的仿真中，应用该方法进行数据预处理，使进入软测量模型的过程数据更接近真实值，取得了很好的效果。

关 键 词：过失误差侦破；软测量；聚类分析
中图分类号：TP 273 文献标识码：A

Data Pre-processing in Soft Sensor Technology

LUO Jian-xu, CHANG Qing
(School of Information Science, East China University of Science and Technology, Shanghai 200030, China)

Abstract: Clustering technique is used to detect gross error of data pre-processing in soft sensor technology. The advantage of this method is no need of priory knowledge and assumption of the process. The gross error detection approach is integrated with moving time window average filter algorithm, so the random error and gross error can be handled simultaneously. The application result in a simulated binary distillation column shows that the proposed approach is a good data pre-processing method for soft sensing technology.

Key words: gross error detection; soft sensing; clustering

1 引 言

现场采集的工业过程测量数据由于各种原因不可避免地带有误差，它的存在会导致建立不正确的软测量模型，也会导致软仪表在实际在线应用时，得到偏离实际值很多的估计输出。因此，有效的数据预处理方法是保证软测量技术成功实施的前提。本文主要研究了软仪表在线实施阶段对动态测量数据的预处理，包括过失误差侦破和随机误差处理。

传统的数据校正方法是建立在过程模型基础上的；同时，求解的过程往往是迭代过程，速度很慢，难以在线应用。本文采用简单的滑动时间窗平均滤波算法平滑随机误差，同时应用聚类分析方法侦破过失误差。将二者有机结合，既能有效地侦破过失误差，又能抑制测量噪声，为软测量估计模型提供了可靠的过程数据。

2 滑动时间窗平均滤波

对于数据中的随机误差，可以采用滑动平均法进行抑制。滑动平均法是一种古典的数据处理方法，它对数据中频繁的随机起伏具有滤波作用，能够有效抑制随机误差。

设置长度为 H 的时间窗，如图 1 所示。

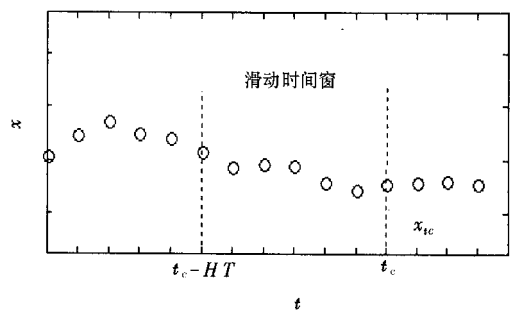

图 1 滑动时间窗

采样时间为 T，在 t_c 时刻得到最新的测量变量值 x_{tc}。对时间窗内的测量值采用最简单的求算术平均值的方法处理，得到平滑滤波后 $\overline{x}_{t_c}$ 的估计值。

$$\overline{x}_{t_c} = \sum_{t=t_c-HT}^{t_c} w_t x_t, \quad w_t = 1/H \tag{1}$$

采用滑动时间窗平均滤波方法处理压力变量，可以看出测量的随机性误差被有效地抑制，如图 2

收稿日期：2006-04-02； 收修定稿日期：2006-04-29
作者简介：罗健旭(1972-)，女，副教授，博士，主要从事复杂工业过程的先进控制等方面的教学与科研工作。

所示。

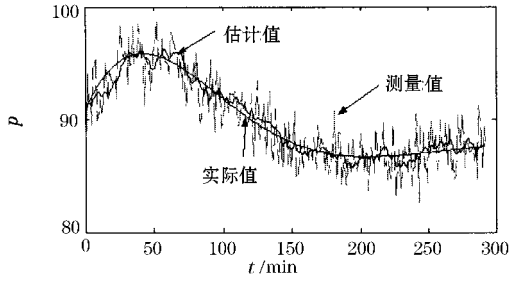

图 2　滑动平均算法处理压力测量变量

然而，当进入时间窗的测量数据 x_{t_c} 是一过失误差时，则将对数据的平滑滤波产生很大的影响，不仅要影响到当前的测量估计值，还将影响后续的 $H-1$ 个值的估计。

具有过失误差的动态测量数据按上述算法处理后，如图 3 所示。

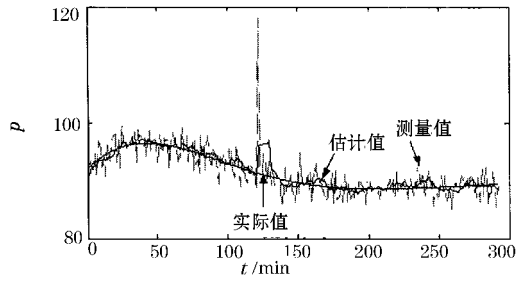

图 3　滑动平均算法处理压力测量变量

3　基于聚类分析的过失误差在线侦破方法

过失误差的存在使得上述滑动时间窗平均滤波算法在一个时间窗的长度内，对测量值的估计产生大的偏差。因此在采用上述算法对测量变量去噪声之前，需要进行过失误差的侦破工作。本文采用基于聚类分析的方法对测量数据进行过失误差侦破。

聚类的目的是把彼此相近的数据集合在一起，成为一类。对于实际的工业过程，变量的变化是连续的、一致的、相关的，而过失误差则是不相关的异常数据，它不服从大多数测量数据的统计分布，通常是整体数据中很小的一部分，远离主体数据。因此能够采用聚类分析的方法来区分出过失数据和正常的主体数据。这种方法的优点就是不需要关于过程的先验知识及假设。

1）一种适于过失误差侦破的聚类算法—MMD 算法　Yin 和 Chen[1] 提出了一种新颖的聚类算法：求取每一数据对象到距离它最近的邻居点的距离，当这一距离满足相似性测度(在一定的距离内)时，将其划分到离它最近的邻居点所属的类；若这一距离不满足相似性测度，则认为此数据对象为噪声。下面定义衡量相似性测度的一个量—平均最小距离(MMD，Mean Minimum Distance)。

定义 1　在 d 维空间，给定具有 N 个数据对象的集合，X_1，X_2，…，N_N，$X_i=(x_{i1}, x_{i2}, \cdots x_{id})$，MMD 表示每一对象到其最近的邻居点的距离的平均值，定义如下：

$$\mathrm{MMD}=\frac{1}{N}\sum_{i=1}^{N}\min_{j\neq i}[(\sum_{k=1}^{d}(x_{ik}-x_{jk})^2)^{1/2}] \qquad (2)$$

某一数据点 X_i 到其最近点的距离 $\mathrm{dist}(x_t)$ 用 d_i 来表示，若 $d_i>2\mathrm{MMD}$，则 X_i 为噪声数据；若 $d_i\leqslant 2\mathrm{MMD}$，则 X_i 为正常数据，将其划分到距离它最近的数据点所属的类。本文将此聚类算法称为 MMD 算法。用该算法处理二维数据，聚类结果如图 4 示。

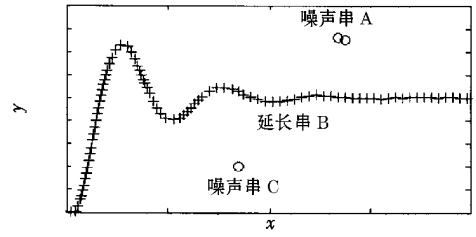

图 4　基于 MMD 算法的二维空间数据聚类

2）基于 MMD 算法的测量数据预处理方法　可以看出该聚类算法能够检测到噪声数据，因而特别适用于侦破过失误差[2~4]。以此算法为基础，将上述滑动时间窗平均滤波算法修正如下：

$$\overline{x}_{t_c}=\sum_{t=t_c-HT}^{t_c}w_t x_t \qquad (3)$$

式中，如果 $\mathrm{dist}(x_t)\leqslant 2\mathrm{MMD}$，$w_t=1/H$；如果 $\mathrm{dist}(x_t)>2\mathrm{MMD}$，$w_t=2\mathrm{MMD}/H\cdot\mathrm{dist}(x_t)$；MMD 为时间窗内的数据点的平均最小距离。

$\mathrm{dist}(x_t)$ 的含义在此处作了修正，以便更适于过失误差侦破和误差处理。$\mathrm{dist}(x_t)$ 不是时间窗内某一数据点 x_t 到其在时间窗内最近点的距离，而是 x_t 到时间窗内所有数据的平均值点的距离。这一修正是因为原始的 MMD 算法无法检验时间窗内的两个距离相近的过失误差。

等权值平均滤波算法改为非等权值后，若动态测量数据中出现了过失误差，它被 MMD 算法侦破出来后，会被赋予较小的权值，从而减小对估计的影响。用结合 MMD 过失误差侦破技术的滑动时间窗平均滤波算法处理数据，结果如图 5 所示。

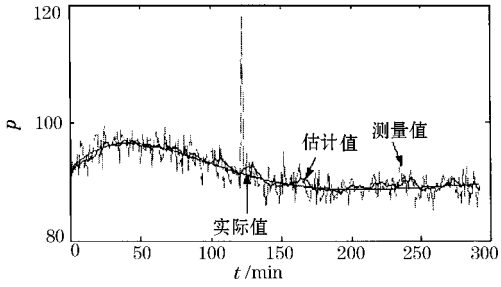

图 5 基于 MMD 算法的滑动平均算法处理压力变量

4 仿 真

1)仿真精馏塔 Skogestad 和 Morari[5] 提出的精馏塔,塔 A 曾被很多学者用来研究二元精馏塔产品组分浓度的推断估计[6,7]。这个塔有 40 个理论塔板(包括再沸器)和一个全冷凝器,第 21 塔板为进料板。为了模拟实际生产过程,给精馏塔增加了两个温度控制回路,用以控制塔顶、塔底产品成分。动态仿真条件如下:进料成分 z_F 和进料流量 F 作为扰动,用伪随机二进制信号作为进料成分的随机扰动,其幅值在静态值的 + 10 % 之内;进料流量每 2 h 做一次阶跃变化,幅度 + 10 %,但流量的总波动限制在静态值的 + 20 % 内。

2)软测量模型 采用 PLS 模型对塔 A 的塔底产品浓度进行估计。选取进料流量 F 和回流量 L 及若干塔板温度 T 作为辅助变量,T = (4,11,20,29,36) 5 个塔板温度。仿真 10 h,采样时间 1 min,获取了 600 组输入输出数据,训练 PLS 模型。

3)软仪表在线应用 将上述训练好的模型投入使用。对进料流量、回流量加入信噪比为 4 的白噪声,同时占流量数据总量 1 % 的数据是过失误差。对温度加入信噪比为 10 的白噪声。在线应用 10 h,用本文介绍的数据预处理方法处理进入 PLS 模型的辅助变量,软测量估计结果及数据处理结果如图 6,图 7,图 8 所示。

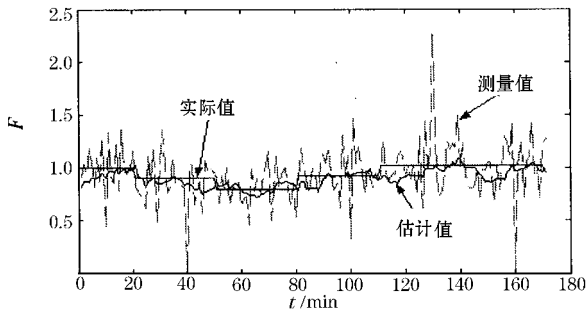

图 6 对带噪声及显著误差的进料流量的数据预处理

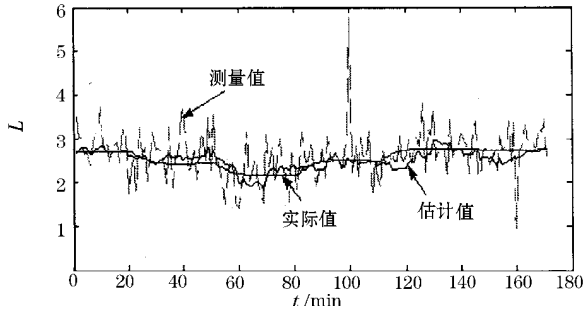

图 7 对带噪声及显著误差的回流量的数据预处理

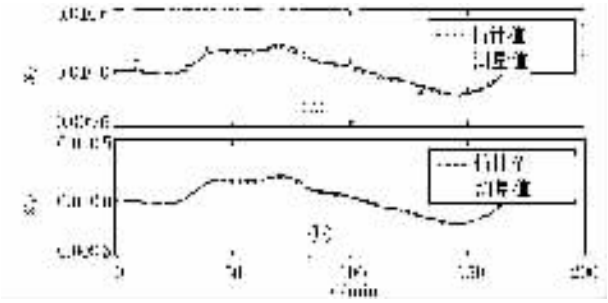

图 8 软仪表估计结果

5 结 语

本文应用聚类分析方法侦破过程数据的过失误差,并将该方法与滑动时间窗平均滤波算法相融合。仿真实例表明,该算法能有效地处理过程数据的随机误差和过失误差,使进入软测量模型的过程数据更加接近真实值,从而提高了软测量模型的精度。

参考文献:

[1] Yin P Y, Chen L H. A new non-iterative approach for clustering[J]. Pattern Recognition Letters, 1994, 15(2): 125-133.

[2] Chen J, Romagnoli J A. A strategy for simultaneous dynamic data reconciliation and outlier detection[J]. Computers and Chemical Engineering, 1998, 22(1): 559-562.

[3] Abuelzeet Z H, Becerra V M, Robert P D. Combined bias and outlier identification in dynamic data reconciliation [J]. Computers and chemical engineering, 2002, 26(6): 921-935.

[4] 罗健旭,邵惠鹤. 软测量建模数据的过失误差侦破——一种基于聚类分析的方法[J]. 仪器仪表学报, 2005, 28(3): 238-241.

[5] Scogestad S, Morari M. Understanding the dynamic behavior of distillation columns [J]. Ind. & Eng. Chem. Res., 1988, 27(10): 1848-1862.

[6] Mejdell T, Skogestad S. Estimation of distillation compositions from multiple temperature measurements using partial-least-squares regression[J]. Ind & Eng Chem Res, 1991, 30(12): 2543-2555.

[7] Mejdell T, Skogestad S. Output estimation using multiple secondary measurements: high purity distillation[J]. AICHE J, 1993, 39(10): 1641-1653.