

正交信号校正的自回归模型及其在动态过程监测中的应用

童楚东, 史旭华, 蓝艇

(宁波大学 信息科学与工程学院, 浙江 宁波 315211)

摘要: 针对采样数据的自相关性, 提出一种基于自回归(AR)模型的动态过程建模方法. 首先, 利用正交信号校正(OSC)消除用于AR模型回归的两数据集间的正交不相关信号; 然后, 在处理过的数据上进行偏最小二乘(PLS)回归建模. 该方法对模型潜隐成分和残差信息同时进行在线监测, 并借鉴贝叶斯推理方法将多个监测指标进行融合, 以易化触发故障警报的决策过程. 最后通过在田纳西-伊斯曼(Tennessee Eastman, TE)过程上的仿真实验验证了所提出方法的有效性.

关键词: 正交信号校正; 自回归模型; 动态过程监测; 偏最小二乘

中图分类号: TP277

文献标志码: A

Orthogonal signal correction based auto-regression model with application to dynamic process monitoring

TONG Chu-dong, SHI Xu-hua, LAN Ting

(Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China. Correspondent: SHI Xu-hua, E-mail: shixuhua@nbu.edu.cn)

Abstract: With respect to the auto-coorelation existed in sampled data, a dynamic process modeling method based on auto-regression(AR) model is proposed. Firstly, orthogonal signal correction(OSC) is used to eliminate the orthogonal uncorrelated components from the two datasets used for regresing AR model. Then, the partial least square(PLS) is employed to get the AR model on the basis of preprocessed process datasets. The latent components extracted by the regression model as well as the residuals are together monitored online. With the involvement of Bayesian inference, the resulted multiple monitoring statistics are combined into a single probabilistic index, and triggering fault alarm can thus be simplified. The simulation experiment on the Tennessee Eastman(TE) benchmark process demonstrates the effectiveness of the proposed method.

Keywords: orthogonal signal correction; auto-regression model; dynamic process monitoring; partial least square

0 引言

随着全球化市场竞争日趋激烈化, 有效地保证生产的正常运行是现代过程工业面临的主要挑战. 与此同时, 由于计算机技术的飞跃发展和工业规模的不断扩大, 工业过程能够测量和存储的数据信息越来越大, 而能够精确描述过程的机理模型却难以获得. 数据驱动的过程监测技术就是在这个背景下应运而生的, 其中针对多变量统计过程监测(MSPM)的研究已得到了工业界和学术界的广泛关注^[1-2]

MSPM方法的理论基础是主成分分析(PCA)和偏最小二乘(PLS)^[2], 但在近几年的发展中, 为了克服

PCA/PLS的缺陷, 学者们提出了各式各样的改进策略^[3-6]. 然而, 对于大多数工业过程而言, 都存在动态性, 即测量样本是时序相关的. 为解决动态过程监测的相关问题, 一些学者提出通过在过程数据矩阵中引入延时测量值构成增广矩阵, 然后对其进行多变量统计分析建模, 这就是所谓的动态PCA(DPCA)^[4]以及其他类似方法的基本出发点^[3,5]. 此外, Kerkhof等^[6]还结合利用自回归模型(AR)和PLS回归算法(AR-PLS)来消除数据自相关性和交叉相关性的影响, 解决了间歇动态过程监测问题. 虽然, 这种利用PLS回归算法训练得到AR模型的思路为动态过程监测问题提供了

收稿日期: 2015-06-11; **修回日期:** 2015-09-17.

基金项目: 浙江省自然科学基金项目(LY14F030004, LY16F030001); 浙江省科技厅公益技术应用研究项目(2015C31017); 宁波市自然科学基金项目(2013A610120).

作者简介: 童楚东(1988—), 男, 副教授, 博士, 从事数据驱动过程监测的研究; 史旭华(1967—), 女, 教授, 博士, 从事数据建模、化工过程优化及机器学习算法等研究.

新的解决方案,但是,传统 PLS 算法对过程数据直接进行回归建模,未曾考虑到剔除数据集间不相关因素的“噪声”影响.若能剔除这部分不相关的信息成分,则 AR 模型就能更好地描述过程数据间的自相关性,从而改善相应监测模型的故障检测效果.为此,将正交信号校正(OSC)^[7]方法融入 AR 模型拟合中,并将之应用于动态过程监测的研究具有重要意义.OSC 是一种应用广泛的数据预处理方法,它能有效地剔除与两数据集无关的信息成分(即正交信号),以期改善回归模型的预测能力^[8].Ghasemi 等^[8]已将 OSC 用于 PLS 回归建模(OPLS),验证了 OPLS 相比于 PLS 的优越性.受此启发,本文利用 OPLS 方法训练过程动态数据的 AR 模型(AR-OPLS),并建立相应的统计监测模型.此外,借鉴贝叶斯推理的信息融合能力^[9-10],进一步提出将 AR-OPLS 多个监测统计量通过概率的形式融合为一个指标,从而简化监测方法触发故障警报的过程.值得强调的是,引入 OSC 预处理数据主要不在于提高回归模型的预测效果,而在于剔除正交不相关信号后,利用 PLS 训练得到的 AR 模型能更好地描述自相关性,以期提高模型的故障检测效果.

1 动态过程监测方法

1.1 DPCA 过程监测模型^[4]

考虑过程数据序列相关性,对矩阵 $\mathbf{X} \in \mathbf{R}^{n \times m}$ 引入前 l 个时刻的测量值,可构成如下增广矩阵:

$$\mathbf{X}_a = \begin{bmatrix} \mathbf{x}_{1+l} & \mathbf{x}_l & \cdots & \mathbf{x}_1 \\ \mathbf{x}_{2+l} & \mathbf{x}_{1+l} & \cdots & \mathbf{x}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & \mathbf{x}_{n-1} & \cdots & \mathbf{x}_{n-l} \end{bmatrix} \in \mathbf{R}^{(n-l) \times (l+1)m}. \quad (1)$$

对 \mathbf{X}_a 进行奇异值分解,可建立相应的 PCA 统计模型,当然,也可以建立其他统计模型.

以 DPCA 模型为例,将监测模型用于在线监测新样本 \mathbf{x}_t 时,需要先对其增广,得到 $\mathbf{x} = [\mathbf{x}_t^T, \mathbf{x}_{t-1}^T, \cdots, \mathbf{x}_{t-l}^T]^T$; 然后利用 Hotelling T^2 统计量和 Q 统计量以及对应的控制限监测当前样本是否发生故障,即^[1]

$$T^2 \leq \frac{d(N-1)}{N-d} F_\alpha(d, n-d), \quad (2)$$

$$Q \leq g\chi_{h,\alpha}^2, \quad g = v/2a, \quad h = 2a^2/v. \quad (3)$$

其中: d 为潜变量个数, $N = n - l$ 为训练样本个数, $F_\alpha(d, n-d)$ 表示自由度为 d 和 $n-d$ 的 F 分布在置信度 α 下的取值, a 和 v 分别为训练数据集 Q 统计量的均值和方差, $\chi_{h,\alpha}^2$ 表示自由度为 h 的 χ^2 在置信度 α 下的取值.

1.2 AR-PLS 方法

1.2.1 AR 回归模型

多变量 AR 模型旨在利用 k 时刻样本 $\mathbf{x}_k \in \mathbf{R}^{m \times 1}$

的前 l 个时刻的测量值来预测 \mathbf{x}_k , 即

$$\mathbf{x}_k = \mathbf{C}^T \mathbf{x}_{k-1:k-l} + \mathbf{e}_k. \quad (4)$$

其中: $\mathbf{x}_{k-1:k-l} = [\mathbf{x}_{k-1}^T, \mathbf{x}_{k-2}^T, \cdots, \mathbf{x}_{k-l}^T]^T \in \mathbf{R}^{ml \times 1}$ 是这 l 个测量样本组成的向量, \mathbf{e}_k 为模型误差, $\mathbf{C} \in \mathbf{R}^{ml \times m}$ 为 AR 模型系数. 式(4)能有效地表示过程数据的时序相关性.

若将所有可用测量样本的 AR 模型整合在一起,则可用如下回归形式统一描述:

$$\mathbf{Y} = \mathbf{X}\mathbf{C} + \mathbf{E}. \quad (5)$$

其中: $\mathbf{Y} = [\mathbf{x}_{l+1}, \mathbf{x}_{l+2}, \cdots, \mathbf{x}_n]^T$ 可看成输出数据矩阵, $\mathbf{X} = [\mathbf{x}_{l:1}, \mathbf{x}_{l+1:2}, \cdots, \mathbf{x}_{n-1:n-l}]^T$ 为输入数据. 则拟合 AR 模型一般可采用 PLS 算法对其进行回归建模,具体介绍如下.

1.2.2 PLS 回归算法

PLS 算法通过潜隐成分 \mathbf{Z} 来建立输入数据矩阵 \mathbf{X} 与输出数据矩阵 \mathbf{Y} 的回归关系, 即

$$\mathbf{X} = \mathbf{Z}\mathbf{Q}^T + \mathbf{E}_X, \quad \mathbf{Y} = \mathbf{Z}\mathbf{R}^T + \mathbf{E}_Y. \quad (6)$$

其中: \mathbf{Q} 和 \mathbf{R} 为载荷矩阵, \mathbf{E}_X 和 \mathbf{E}_Y 为残差矩阵. 通过式(6), 回归系数矩阵 \mathbf{C} 可计算为

$$\mathbf{C} = \mathbf{W}(\mathbf{Q}\mathbf{W})^{-1}\mathbf{R}^T = \mathbf{J}\mathbf{R}^T, \quad (7)$$

其中 \mathbf{W} 为 NIPALS 算法计算出来的加权矩阵.

1.2.3 过程监测方法

在式(6)的回归模型中,为了充分利用 AR-PLS 模型所得到的成分进行监测,可进一步对残差矩阵进行 PCA 建模,即

$$\mathbf{E}_X = \mathbf{T}_X \mathbf{P}_X^T + \mathbf{F}_X, \quad \mathbf{E}_Y = \mathbf{T}_Y \mathbf{P}_Y^T + \mathbf{F}_Y, \quad (8)$$

从而得到相应得分矩阵 \mathbf{T}_X 和 \mathbf{T}_Y , 以及载荷矩阵 \mathbf{P}_X 和 \mathbf{P}_Y , $\mathbf{F}_X \in \mathbf{R}^{(n-l) \times ml}$ 和 $\mathbf{F}_Y \in \mathbf{R}^{(n-l) \times m}$ 为对应的残差矩阵. 当将 AR-PLS 用于监测新样本 \mathbf{x}_t 时,首先利用其前 l 个时刻的延时测量值组成向量 $\mathbf{x}_{t-1:t-l} = [\mathbf{x}_{t-1}^T, \mathbf{x}_{t-2}^T, \cdots, \mathbf{x}_{t-l}^T]^T$, 计算潜隐成分

$$\mathbf{z}_t = \mathbf{J}^T \mathbf{x}_{t-1:t-l},$$

并建立如下 T^2 统计量:

$$T_z^2 = \mathbf{z}_t^T \mathbf{S}_z^{-1} \mathbf{z}_t, \quad (9)$$

其中 \mathbf{S}_z 为 \mathbf{Z} 的协方差矩阵; 然后需分别将残差 $\mathbf{e}_X = \mathbf{x}_{t-1:t-l} - \mathbf{Q}\mathbf{z}_t$ 和 $\mathbf{e}_Y = \mathbf{x}_t - \mathbf{R}\mathbf{z}_t$ 投影至对应的 PCA 模型中, 并建立相应的 T^2 和 Q 统计量

$$T_X^2 = \mathbf{e}_X^T \mathbf{P}_X \mathbf{\Lambda}_X^{-1} \mathbf{P}_X^T \mathbf{e}_X,$$

$$T_Y^2 = \mathbf{e}_Y^T \mathbf{P}_Y \mathbf{\Lambda}_Y^{-1} \mathbf{P}_Y^T \mathbf{e}_Y; \quad (10)$$

$$Q_X = \mathbf{e}_X^T (\mathbf{I} - \mathbf{P}_X \mathbf{P}_X^T) \mathbf{e}_X,$$

$$Q_Y = \mathbf{e}_Y^T (\mathbf{I} - \mathbf{P}_Y \mathbf{P}_Y^T) \mathbf{e}_Y. \quad (11)$$

其中 $\mathbf{\Lambda}_X$ 和 $\mathbf{\Lambda}_Y$ 分别为 \mathbf{T}_X 和 \mathbf{T}_Y 的协方差矩阵. 对应控制限的确定方法与式(2)和(3)相仿.

2 基于 AR-OPLS 的动态过程监测

2.1 基于 AR-OPLS 的动态数据建模

考虑到 \mathbf{X} 中会存在与 \mathbf{Y} 正交的成分, 若对 \mathbf{X} 进行数据预处理, 将与 \mathbf{Y} 不相关的信号成分剔除, 则随后建立的回归模型拟合效果会得到相应的改善. 在上述 AR-PLS 的基础上, 引入 OSC 剔除正交不相关的信号. OSC 的迭代计算过程如下.

- 1) 对矩阵 \mathbf{X} 进行 PCA 建模, 得到最大奇异值对应的主成分向量 \mathbf{t} ;
- 2) 正交化 \mathbf{t} 和 \mathbf{Y} , 即 $\mathbf{t}_{\text{new}} = [\mathbf{I} - \mathbf{Y}(\mathbf{Y}^T \mathbf{Y})^{-1} \mathbf{Y}^T] \mathbf{t}$;
- 3) 利用 PLS 得到加权向量 \mathbf{w} , 使之满足 $\mathbf{X} \mathbf{w} = \mathbf{t}_{\text{new}}$;
- 4) 更新 $\mathbf{t} = \mathbf{X} \mathbf{w}$;
- 5) 检验是否满足条件 $\|\mathbf{t} - \mathbf{t}_{\text{new}}\| / \|\mathbf{t}\| < 10^{-6}$, 若否, 则重复执行步骤 2) ~ 步骤 5), 若是, 则执行步骤 6);
- 6) 剔除正交成分 \mathbf{t} , 即 $\mathbf{X} = \mathbf{X} - \mathbf{t} \mathbf{t}^T$;
- 7) 重复执行上述步骤, 以剔除更多正交成分.

将剔除正交成分后的 \mathbf{X} 和 \mathbf{Y} 进行 PLS 建模, 以得到 OSC 处理后数据的 AR 模型, 即 AR-OPLS.

值得强调的是, AR-OPLS 模型用于监测 \mathbf{x}_t 时, 需首先按照上述步骤 6) 剔除 $\mathbf{x}_{t-1:t-l}$ 中与 \mathbf{x}_t 正交的成分; 然后与 AR-PLS 一样, 同样会建立 3 个 T^2 统计量和 2 个 Q 统计量对新采样数据进行监测. 一般而言, 任何一个监测统计量触发故障警报都代表过程进入非正常运行状态. 但是, 这样一来, 会使监测模型对较大的随机扰动更为敏感, 增加误报故障的机率. 为此, 进一步采用贝叶斯推理将这 5 个统计指标融合为一个概率型指标. 该概率型监测指标能综合考虑这 5 个监测指标的故障检测情况, 不仅能简化 AR-OPLS 模型触发故障警报的决策过程, 而且还能有效地避免误报故障现象的发生.

2.2 基于贝叶斯推理的概率型监测指标

假设监测指标 M_i (可以是 T^2 或 Q 统计量) 在置信度 α 下的控制限为 C_i , 新样本 \mathbf{x}_t 的故障概率为

$$\vartheta_{M_i}(F|\mathbf{x}_t) = \vartheta_{M_i}(\mathbf{x}_t|F)\vartheta_{M_i}(F)/\vartheta_{M_i}(\mathbf{x}_t), \quad i = 1, 2, \dots, 5. \quad (12)$$

概率 $\vartheta_{M_i}(\mathbf{x}_t)$ 定义如下:

$$\vartheta_{M_i}(\mathbf{x}_t) = \vartheta_{M_i}(\mathbf{x}_t|N)\vartheta_{M_i}(N) + \vartheta_{M_i}(\mathbf{x}_t|F)\vartheta_{M_i}(F). \quad (13)$$

其中: N 和 F 分别代表正常和故障, $\vartheta_{M_i}(N)$ 和 $\vartheta_{M_i}(F)$ 分别取值 α 和 $1-\alpha$. 条件概率 $\vartheta_{M_i}(\mathbf{x}_t|N)$ 和 $\vartheta_{M_i}(\mathbf{x}_t|F)$ 按下式计算:

$$\begin{aligned} \vartheta_{M_i}(\mathbf{x}_t|N) &= \exp(-M_i/C_i), \\ \vartheta_{M_i}(\mathbf{x}_t|F) &= \exp(-C_i/M_i). \end{aligned} \quad (14)$$

最终的概率监测指标 BIC 可按加权形式融合为一, 即

$$\text{BIC} = \sum_{i=1}^5 \left\{ \vartheta_{M_i}(\mathbf{x}_t|F)\vartheta_{M_i}(F|\mathbf{x}_t) / \sum_{i=1}^5 \vartheta_{M_i}(\mathbf{x}_t|F) \right\}. \quad (15)$$

当 $\text{BIC} > 1 - \alpha$ 时, \mathbf{x}_t 为非正常样本, 反之为正常样本.

3 TE 过程仿真研究

TE 仿真模型因其反应过程和生产结构的复杂性, 已经成为测试不同控制方法和过程监测策略的标准实验平台^[1,10], 它可以仿真模拟 21 种不同的故障类型, 详细资料可参见文献 [1]. 选取过程连续测量的 33 个变量作为监测变量, 详情可参见文献 [10]. 仿真数据集源于 <http://web.mit.edu/braatzgroup/links.html>.

离线建模阶段, 利用正常工况下的 960 个样本建立 DPCA、AR-PLS 和 AR-OPLS 的过程监测模型, 以作对比分析用. 其中: 时滞 $l = 2$, DPCA 模型选取的主成分个数通过 $\text{CPV} > 85\%$ 来确定, PLS 模型的潜隐变量个数通过 10-fold 交叉验证确定, OSC 迭代剔除 2 个正交成分, 置信限 $\alpha = 99\%$.

利用 DPCA、AR-PLS 和 AR-OPLS 这 3 种动态过程监测模型对 TE 过程 21 种故障工况进行监测, 并计算相应的故障漏报率, 详情列于表 1 中. 值得指出的是, 基于贝叶斯推理的概率型监测指标同样适用于传统 AR-PLS 监测模型. 此外, 故障 3、9 和 15 对过程数据影响甚微, 很多文献都证实这 3 类故障很难被检测出来^[1,10], 因此, 这 3 种故障不予考虑. 在表 1 中, 取得最小漏报率的指标已用粗体标出. 显然, AR-OPLS 方法在绝大多数故障类型上能取得优于传统 DPCA 和 AR-PLS 方法的监测结果. 尤其是针对故障 5、19 和 20, 故障漏报率得到大幅度下降. 虽然 AR-OPLS 方法在故障 1、2、18 和 21 上的监测效果不如传统 DPCA

表 1 TE 过程 21 类故障漏报率 %

No.	DPCA		AR-PLS	AR-OPLS
	T^2	Q	BIC	BIC
1	0.75	0.13	0.25	0.25
2	1.50	2.75	1.88	2.13
4	94.00	0.00	0.00	0.00
5	76.00	55.00	63.25	0.00
6	1.13	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00
8	2.63	3.00	2.25	2.38
10	75.13	48.88	49.38	40.63
11	72.88	6.00	5.38	5.13
12	0.88	3.63	0.75	0.50
13	5.75	4.63	4.50	4.38
14	0.13	0.00	0.00	0.00
16	90.38	48.00	51.13	40.38
17	22.75	2.25	2.25	2.25
18	11.13	9.38	9.50	9.50
19	77.25	33.38	41.38	17.75
20	61.13	36.38	42.88	29.25
21	54.25	49.50	53.13	53.88

方法优越,但是相应的漏报率相差微乎其微.为了更好地体现 AR-OPLS 相对于 DPCA 和 AR-PLS 方法的优越性,将故障 5 的过程监测结果显示于图 1 中.很显然,传统 PCA 方法对故障 5 存在很大的漏报,而 AR-OPLS 方法能一直持续触发故障警报.

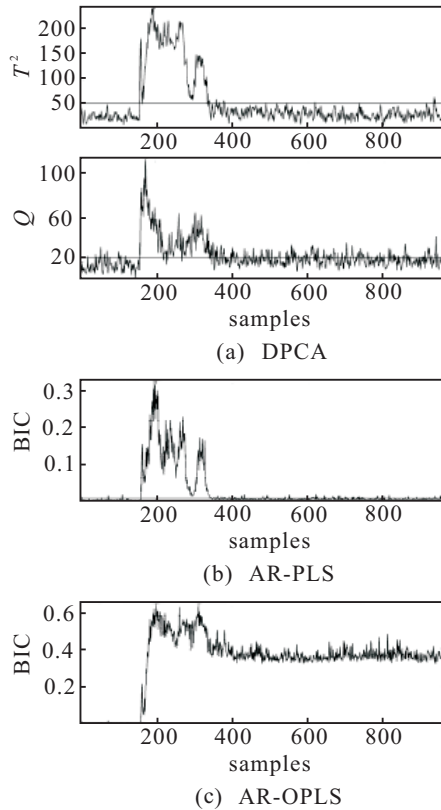


图 1 故障 5 的过程监测结果

在分析了 AR-OPLS 方法的故障监测效果后,需继续验证该方法的故障误报率.一般而言,较低的故障漏报率对应着较高的故障误报率,即将正常数据样本错判为故障.利用另一组正常工况下采集到的 500 个数据检验这 3 种方法统计监测指标的误报率,对应结果如表 2 所示.从表 2 可以看出,AR-OPLS 的故障误报率低于 DPCA 和 AR-PLS 方法.表 1 和表 2 的对比分析结果充分验证了基于 AR-OPLS 的动态过程监测方法的优越性和有效性.

表 2 正常测试数据的故障漏报率 %

	DPCA		AR-PLS	AR-OPLS
	T^2	Q	BIC	BIC
Index value	0.63	3.24	2.41	1.41

4 结 论

本文针对 AR-PLS 模型训练过程中未考虑两数据集间存在正交信号成分的问题,引入 OSC 方法对数据进行预处理.对剔除正交信号后的数据进行 PLS 回归建模,提出了一种基于 AR-OPLS 的动态过程监测方法.在 TE 过程上的仿真研究充分说明了 AR-OPLS

方法优于传统 DPCA 和 AR-PLS 方法.然而 AR-OPLS 仍是一种线性监测方法,如何将其扩展成非线性形式是未来研究中需要解决的问题.此外,故障诊断的研究还需进一步深入.

参考文献(References)

- [1] Chiang L H, Russell E L, Braatz R D. Fault detection and diagnosis in industrial systems[M]. London: Springer-Verlag, 2001: 6-13.
- [2] Ge Z, Song Z, Gao F. Review of recent research on data-based process monitoring[J]. Industrial & Engineering Chemistry Research, 2013, 52(10): 3543-3562.
- [3] 赵忠盖,刘飞.动态因子分析模型及其在过程监控中的应用[J].化工学报,2009,60(1): 183-186.
(Zhao Z G, Liu F. Modeling using dynamic factor analysis and its application in process monitoring[J]. J of Chemical Industry and Engineering, 2009, 60(1): 183-186.)
- [4] Ku W, Storer R H, Georgakis C. Disturbance detection and isolation by dynamic principal component analysis[J]. Chemometrics & Intelligent Laboratory Systems, 1995, 30(1): 179-196.
- [5] Komulainen T, Sourander M, Jämsä-Jounela S L. An online application of dynamic PLS to a dearomatization process[J]. Computers & Chemical Engineering, 2004, 28(12): 2611-2619.
- [6] Kerkhof P V D, Gins G, Vanlaer J, et al. Dynamic model-based fault diagnosis for (bio)chemical batch processes[J]. Computers & Chemical Engineering, 2012, 40: 12-21.
- [7] Svensson O, Kourti T, MacGregor J F. An investigation of orthogonal signal correction algorithms and their characteristics[J]. J of Chemometrics, 2002, 16(4): 176-188.
- [8] Ghasemi J, Niazi A. Spectrophotometric simultaneous determination of nitroaniline isomers by orthogonal signal correction-partial least squares[J]. Talanta, 2005, 65(5): 1168-1173.
- [9] Jiang Q, Wang B, Yan X. Multiblock independent component analysis integrated with Helinger distance and Bayesian for non-Gaussian plant-wide process monitoring[J]. Industrial & Engineering Chemistry Research, 2015, 54(9): 2497-2508.
- [10] Tong C, Palazoglu A, Yan X. Improved ICA for process monitoring based on ensemble learning and Bayesian inference[J]. Chemometrics & Intelligent Laboratory Systems, 2014, 135: 141-149.

(责任编辑:曹洪武)