

基于样本稀疏化高斯过程的发酵过程 软测量建模方法

何 坤 赵利强 王建林* 于 涛

(北京化工大学 信息科学与技术学院, 北京 100029)

摘 要: 提出了一种基于样本稀疏化高斯过程(GP)的发酵过程软测量建模方法。该方法将聚类 and 灰色关联度分析相融合,综合考虑样本点间欧式距离和各个特征向量对样本点间相似度的影响,通过剔除相似度比较大的样本点,实现训练样本集的稀疏化,降低了模型的计算复杂度。利用基于样本稀疏化的高斯过程构建青霉素发酵过程的软测量模型,同时得到青霉素浓度的预估值和表征预估值的不确定度,实验结果表明,本文所提方法与标准 GP 方法相比,在保证模型预测精度的前提下,减少了模型的训练时间。

关键词: 高斯过程;样本稀疏化;仿射传播聚类算法;灰色关联度分析

中图分类号: TN911.4

引 言

生物参量是反映发酵过程进程的重要指标,受生物传感技术发展水平的限制,一些关键的生物参量不易在线测量,难以实现发酵过程的优化控制,因此有必要对不易在线测量的关键生物参量进行软测量。作为软测量建模的典型方法,神经网络和支持向量机在参数选择^[1-2]以及对软测量模型不确定度的获取^[3]方面存在显著不足。高斯过程(GP)建模不仅可以很方便的得到模型不确定度,而且模型训练过程中超参数优化也易于实现。然而作为一种新兴的回归算法^[4-5],GP在发酵过程的软测量模型构建中应用较少。di Sciascio等^[6]对苏云金杆菌(*Bacillus thuringiensis*)发酵过程中生物参量的在线检测构建了高斯过程软测量模型,利用该模型对预测样本进行模型预估,得到了比较好的预估效果。Azman等^[7]利用高斯过程对生物系统进行黑箱建模,并将其成功应用于硝化厂的污水处理过程以及威尼斯湖的藻类生长预测过程。

在训练高斯过程模型中涉及矩阵的求逆运算,其计算复杂度和样本规模有很大的关联^[8],有必要对样本进行稀疏化处理。样本稀疏化的基本原则是

选取可表征原始样本的有效样本子集。近年来,研究者们在高斯过程的样本稀疏化方面进行了大量的研究^[9-11]。如 Seeger 等^[10]给出了高斯过程中运用贪婪近似原则选取有效样本子集的方法,Luo 等^[11]应用层次聚类的方法以样本点间的欧氏距离得到了稀疏样本子集。上述改进方法虽然可以对训练样本集合进行稀疏化,但是只考虑了欧式距离对数据间相似度的影响,并没有考虑样本点各个特征向量的影响。因此如何选取最具有代表性的训练样本子集是解决发酵过程中基于 GP 软测量建模方法计算复杂度问题的关键。

本文提出了一种基于样本稀疏化 GP 的软测量建模方法,采用聚类 and 灰色关联度分析相融合的方法实现训练样本集的稀疏化,并利用稀疏化后可代表训练样本集合的有效样本子集构建了基于 GP 的发酵过程软测量模型。

1 高斯过程建模方法

高斯过程的基础理论是贝叶斯理论方法,其回归建模的基本思想是:以贝叶斯理论为基础,首先设定隐函数 $f(x)$ 的先验概率分布,在获得样本信息之后,综合样本信息和含有未知参数的 $f(x)$ 的先验分布,利用贝叶斯定理,求得 $f(x)$ 的后验分布,根据后验分布推断未知样本的参数模型。假定训练样本表示为 $D = \{(\mathbf{x}_i, y_i)\}$,其中 $\mathbf{x}_i \in \mathbf{R}^l$, $y_i \in \mathbf{R}$,训练样本数为 n ,输入向量维数为 l 。对于新的测试样本 \mathbf{x}^* ,

收稿日期: 2013-04-22

第一作者: 女,1986年生,硕士生

* 通讯联系人

E-mail: wangjl@mail.buct.edu.cn

<http://www.journal.buct.edu.cn>

GP 模型的预测值为

$$\mathbf{y}^* = \mu(\mathbf{x}^*) = \mathbf{k}^T \mathbf{K}^{-1} \mathbf{y} \quad (1)$$

预测值对应的方差为

$$\sigma^2(\mathbf{x}^*) = \mathbf{k}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}(\mathbf{x}, \mathbf{x})^T \mathbf{K}^{-1} \mathbf{k}(\mathbf{x}, \mathbf{x}^*) \quad (2)$$

式(1)中, \mathbf{K} 表示协方差矩阵; \mathbf{k} 代表样本点输入值间的协方差函数; $\mathbf{k}(\mathbf{x}^*, \mathbf{x}^*)$ 表示测试输入与其自身的协方差矩阵。由式(1)和式(2)可知高斯过程模型可以表示为包含均值和方差的基于函数的一个高斯分布模型, 利用设定的协方差函数和选择的训练样本信息实现对测试样本的预测。

协方差函数的确定是高斯过程建模中最关键的步骤。对于协方差函数要保证其对应的协方差矩阵为对称半正定矩阵。本文所使用的协方差函数为

$$\mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_s^2 \exp \left(- \sum_{d=1}^l \frac{\|\mathbf{x}_d - \mathbf{x}'_d\|}{2l_d^2} \right) + \sigma_n^2 \delta_{ij} \quad (3)$$

式(3)中, $\{\sigma_s, \sigma_n, l_d\} = \boldsymbol{\theta}$ 表示模型的超参数, σ_s 代表对先验知识的一个量度, σ_n 代表服从高斯分布的噪声信息; l_d 代表模型的尺度参数; δ_{ij} 表示 Kronecker 算子, 其中 i 和 j 相同时为 1, 不同时为 0。设定超参数的初始值, 通过式(4)所示极大化似然函数得到超参数的最优值

$$\ln p(\mathbf{y}|\mathbf{x}) = -\frac{1}{2} \mathbf{y}^T \mathbf{K}^{-1} \mathbf{y} - \frac{1}{2} \ln |\mathbf{K}| - \frac{n}{2} \ln 2\pi \quad (4)$$

在优化超参数过程中涉及目标函数对超参数的求导, 其导数表示为

$$\frac{\partial L}{\partial \theta_j} = -0.5 \mathbf{y}^T \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_j} \mathbf{K}^{-1} \mathbf{y} + 0.5 \text{tr} \left(\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_j} \right) = 0.5 \text{tr} \left[(\mathbf{K}^{-1} - \mathbf{a} \cdot \mathbf{a}^T) \frac{\partial \mathbf{K}}{\partial \theta_j} \right] \quad (5)$$

由式(5)知利用共轭梯度法对超参数寻优过程中涉及 $n \times n$ 维矩阵的求逆运算, 其计算复杂度和样本规模 n 相关, 可以通过稀疏化样本的方法降低模型的计算复杂度。

2 基于样本稀疏化 GP 软测量建模

2.1 样本稀疏化方法

样本稀疏化的本质是选取有效样本集合。随机选取的样本子集并不能完全表达原始样本的整体特征, 单纯采用聚类的方法选取样本子集不能完全识别样本点的各个特征向量对该样本点的贡献率, 导

致样本点间相似度量度的偏差出现。本文从对训练集近似表达的角度出发, 同时考虑各个样本点间每个特征向量对数据间相似度的影响, 采用聚类和灰色关联度分析相结合的方法对训练样本稀疏化。基本思路为, 首先利用聚类方法以样本数据间的欧式距离作为数据间相似性度量的原则, 对训练样本进行聚类分析; 其次将各个聚类中心作为参考序列分别对每个类别的样本进行灰色关联度分析, 将各个样本对参考列数据的关联度作为判别数据间相似度的依据; 最后选取所有的参考列数据作为稀疏样本集合。

2.1.1 仿射传播聚类算法

仿射传播聚类(AP)方法^[12]和 k 均值算法同属于 k 中心聚类算法, 传统的 k 均值算法由于初始聚类中心的选取比较敏感易陷入局部最优, AP 算法最大的优势在于将所有的样本点看作聚类中心, 在聚类过程中定义了两个变量, 吸引度 $r(n, m)$ 和归属度 $a(n, m)$, 通过竞争机制进行迭代更新, 更新规则为

$$r(n, m) = s(n, m) - \max_{m': m' \neq m} \{a(n, m') + s(n, m')\} \quad (6)$$

$$r(m, m) = s(m, m) - \max_{m': m' \neq m} \{a(m, m') + s(m, m')\} \quad (7)$$

$$a(n, m) = \min \left\{ 0, r(m, m) + \sum_{n': n' \neq n, m} \max \{0, r(n', m)\} \right\} \quad (8)$$

$$a(m, m) = \sum_{n': n' \neq m} \max \{0, r(n', m)\} \quad (9)$$

式(6)~(9)中, $r(n, m)$ 表示样本点 n 和样本点 m 之间的吸引度; $a(n, m)$ 表示样本点 m 对样本点 n 的归属度; $s(n, m)$ 表示样本点 n 和样本点 m 的相似度。

通过式(6)~(9)更新规则, 样本点之间进行信息互换, 最终使得所选择的能量函数达到极小, 通过设置固定的迭代次数是局部的变化程度加以终止。AP 聚类的方法以样本点间的欧式距离作为样本点间的相似度测量, 并没有考虑特征向量对相似度的影响。

2.1.2 灰色关联度分析方法

灰色关联度分析(GRA)是一种通过计算关联系数和关联度判定不同特征之间关联程度的方法^[13]。虽然用回归或者相关分析的方法也可以对数据间的关联程度作出决策, 但是一般需要较多的数据量, 对数据的分布特征也有特定的要求。相对

而言 GRA 方法实现原理比较简单,实现步骤易于理解。具体的实现方法如图 1 所示。

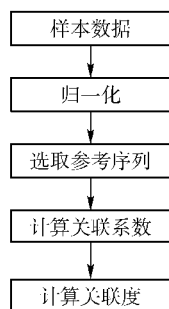


图 1 GRA 方法实现流程

Fig. 1 The flow diagram for GRA implementation

由于 GRA 方法可以通过计算特征向量间的关联度来表征参考列和样本列数据间的关联程度,因此该方法考虑了样本点间每个特征向量对相似度的影响。假设 3 个样本点到参考序列的欧式距离相等,如果样本点 1 和样本点 2 在特征向量方向上的数据的表现形式相似,就可以选择其中之一作为有效样本,剔除另外一个样本点,从而实现样本集合的稀疏化。

2.2 样本稀疏化 GP 软测量建模方法实现

利用本文提出的样本稀疏化 GP 方法构建软测量模型具体的实现流程如图 2 所示,具体实现步骤如下。

步骤 1 将样本数据划分为训练样本及测试样本。

步骤 2 按照 2.1 节提出的样本稀疏化方法实现训练样本的稀疏化。

1) 对训练样本 $X = \{x_i, i = 1 \cdots n\}$ 进行归一化处理,再进行 AP 聚类分析,得到每个类别的聚类中心。

2) 将每个聚类中心作为参考列,对属于本类别的训练样本进行灰色关联度分析,计算关联系数

$$\varepsilon_i(j) = \frac{\min_i \min_j \Delta_{ij} + \rho \max_i \max_j \Delta_{ij}}{\Delta_{ij} + \rho \max_i \max_j \Delta_{ij}} \quad (10)$$

式(10)中, Δ_{ij} 表示参考序列和第 i 个样本序列的第 j 个特征的绝对差,通常取 $\rho = 0.5$, $\min_i \min_j \Delta_{ij}$ 为两级最小差, $\max_i \max_j \Delta_{ij}$ 为两级最大差。计算样本序列和参考序列的关联度

$$r_i = \frac{1}{N} \sum_{j=1}^N \varepsilon_i(j) \quad (11)$$

式(11)中, r_i 表示样本序列和参考序列的关联度, N

表示特征向量的数目。经过多次实验分析的结果表明关联度大于 0.9 的样本序列可视为参考序列的相似样本,将其剔除并不影响数据的整体特征信息。

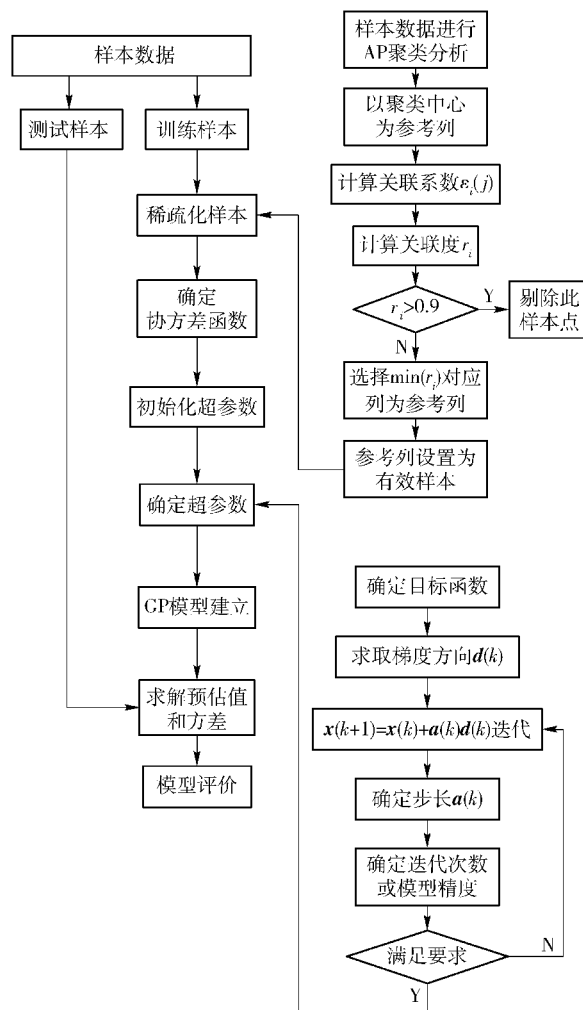


图 2 样本稀疏化 GP 实现流程

Fig. 2 The flow diagram for sample sparse GP implementation

3) 以步骤 2) 中关联度最小的数据列作为参考列,重复步骤 2) 和 3), 直至不存在关联度小于 0.9 的数据列,联合所有的参考列数据,将其作为稀疏样本集合。

步骤 3 按照协方差函数的选取要求选择合适的协方差函数。

步骤 4 设定协方差函数中超参数的初始值,利用共轭梯度等优化方法实现参数寻优,选取的目标函数如式(4)所示。

1) 利用共轭梯度法进行参数寻优时,梯度 $d(k)$ 选取的是目标函数对各个参数求导所得到的梯度的负方向。

2) 利用式(12)进行迭代,对迭代步长的求取采

用一维线性搜索方法。

$$\mathbf{x}(k+1) = \mathbf{x}(k) + \mathbf{a}(k)\mathbf{d}(k) \quad (12)$$

式(12)中 $\mathbf{a}(k)$ 表示迭代步长, $\mathbf{d}(k)$ 表示梯度方向。

3) 根据设定的迭代次数或者迭代精度判定是否结束,如果满足要求,转至4),如果不满足要求返回2)继续迭代,直至满足要求。

4) 根据参数寻优算法确定超参数的最优值。

步骤5 利用测试样本及式(1)、(2)分别求解高斯过程测试样本的预估值及预测方差。

3 实例分析

本文采用一个通用的青霉素发酵过程仿真模型,由 Pensim 仿真平台^[14]进行实验验证过程。以2 h 为采样周期,采集15批数据,每批数据含有400个时刻点的数据作为样本,其中10批数据作为训练样本,5批数据作为测试样本。通过对发酵过程进行机理分析,选取对输出影响较大的可在线测量变量中的溶解氧浓度(mol/L),二氧化碳浓度(mmol/L),发酵液体积(L)为辅助变量,作为软测量模型的输入变量;选取过程变量中的青霉素浓度为主导变量,作为软测量模型的输出变量。

3.1 样本稀疏化实验

选取同一批次的200组数据作为训练样本,不同于训练样本批次的50组数据作为测试样本。利用高斯过程构建软测量模型,选取式(3)作为协方差函数,超参数的初始值 $\theta = \{\sigma_n, \sigma_s, l_1, l_2, l_3\} = \{0.1, 1, 1, 1, 1\}$,利用共轭梯度法求取最优超参数。

以预测均方根误差(RMSE) R 、平均对数密度误差(LD) L 和最大绝对误差(MAE) M 作为模型性能评价标准。计算公式为

$$R = \frac{1}{n} \sum_{d=1}^l (\hat{y}_i - y_i)^2 \quad (13)$$

$$L = \frac{1}{l} \sum_{i=1}^l \left[\lg(2\pi) + \lg\sigma_i^2 + \frac{(\hat{y}_i - y_i)^2}{\sigma_i^2} \right] \quad (14)$$

$$M = \max(|\hat{y}_i - y_i|) \quad (15)$$

式(13)~(15)中, n 表示样本规模, \hat{y}_i 表示模型预测值, y_i 表示实际测量值, σ_i^2 表示高斯过程软测量模型的方差。RMSE 越小,LD 越小,模型的泛化能力越好,在给定的置信概率下,模型的不确定度越小,模型预测精度越高。

利用本文提出的稀疏化 GP 方法构建发酵过程的软测量模型。首先采用样本稀疏化的方法对发酵过程的训练样本进行稀疏化处理,对发酵过程的仿

真数据经过稀疏化之后,样本数目由给定的200组样本数据缩减为122组;然后利用得到的有效样本集合构建高斯过程软测量模型。通过理论分析可知当样本规模由 n 稀疏化至 m 后,GP 训练过程的计算复杂度由 $O(n^3)$ 降低至 $O(m^3)$ 。表1给出了基于样本稀疏化方法的改进 GP 和标准 GP 的训练时间及预测精度。

表1 不同模型预测效果对比

Table 1 Different ways of modeling prediction effect comparisons

方法	训练时间/s	RMSE	MAE	LD
标准 GP	17.8277	0.0028	0.0405	0.6924
稀疏化 GP	3.0237	0.0029	0.0398	0.7012

由表1可知经样本稀疏化改进后的 GP 训练时间明显缩减,虽然选取的样本规模有所降低,但是得到的软测量模型的预测精度并没有降低,表明利用本文提出的样本稀疏化方法能得到可表征原始训练样本的有效样本集合,证明了该方法是一种比较有效的样本稀疏化方法。

3.2 对比实验

采用于振亚等^[15]提出的 SVM 建模方法和本文提出的样本稀疏化后的高斯过程建模方法进行对比实验。利用 SVM 建模方法构建软测量模型选取高斯核函数,核宽度设置为0.5,惩罚因子 C 设置为200,不敏感系数 ε 设置为0.35。选取200组发酵过程数据作为训练样本,另外50组数据作为测试样本。分别利用 SVM 方法和改进 GP 构建发酵过程软测量模型的预估值如图3所示,模型误差值如图4所示。

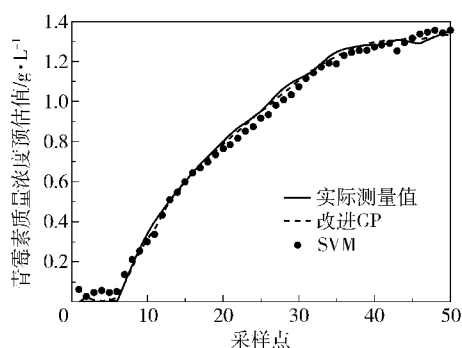


图3 模型预估值

Fig. 3 Soft-sensor model prediction

通过图4可以明显看出稀疏化后的 GP 建模方法得到的模型误差值比 SVM 模型的要小,模型的预

测能力比较好。

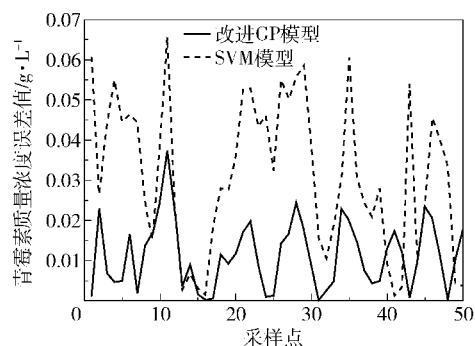


图4 模型误差值

Fig. 4 Soft-sensor model error

表2给出了样本稀疏化GP和SVM软测量建模方法预测效果的详细对比结果。因为SVM建模方法不涉及不确定度的问题,故其LD不存在。SVM软测量建模方法的RMSE和MAE的值明显大于基于稀疏化GP软测量建模方法相对应的模型评价指标的值。由此可见基于样本稀疏化方法改进的GP建模方法较SVM建模方法是一种更加有效的软测量建模方法,可以明显的提高模型的预测精度。

表2 不同建模方式预测效果对比

Table 2 Different ways of modeling prediction effect comparisons

方法	RMSE	MAE	LD
SVM	0.0274	0.0655	—
稀疏化 GP	0.0024	0.0374	0.7109

4 结论

本文提出的基于样本稀疏化GP的发酵过程软测量建模方法,采用聚类 and 灰色关联度分析相融合 of 样本稀疏化方法,能得到可表征原始数据的有效样本子集;利用该样本子集构建基于GP的发酵过程软测量模型,在保证模型预测精度的前提下,有效的降低了模型训练过程中的计算复杂度,减少了模型的训练时间。

参考文献:

[1] 陈如清, 俞金寿. 基于改进神经网络集成算法的软测量建模[J]. 仪器仪表学报, 2008, 29(6): 1240-1245.
Chen R Q, Yu J S. Soft sensing modeling based on improved neural network ensemble algorithm[J]. Chinese Journal of Scientific Instrument, 2008, 29(6): 1240-

1245. (in Chinese)
[2] 刘国海, 周大伟, 徐海霞, 等. 基于SVM的微生物发酵过程软测量建模研究[J]. 仪器仪表学报, 2009, 30(6): 1229-1232.
Liu G H, Zhou D W, Xu H X, et al. Soft sensor modeling using SVM in fermentation process[J]. Chinese Journal of Scientific Instrument, 2009, 30(6): 1229-1232. (in Chinese)
[3] 王华忠. 高斯过程及其在软测量建模中的应用[J]. 化工学报, 2007, 58(11): 2840-2845.
Wang H Z. Gaussian process and its application to soft sensor modeling[J]. Journal of Chemical Industry and Engineering, 2007, 58(11): 2840-2845. (in Chinese)
[4] Gregorčič G, Lightbody G. Gaussian process approach for modelling of nonlinear systems[J]. Engineering Applications of Artificial Intelligence, 2009, 22(3): 522-533.
[5] Chen T, Morris J, Martin E. Gaussian process regression for multivariate spectroscopic calibration[J]. Chemometrics and Intelligent Laboratory Systems, 2007, 87: 59-71.
[6] di Sciascio F, Amicarelli A N. Biomass estimation in batch biotechnological processes by Bayesian Gaussian process regression[J]. Computers & Chemical Engineering, 2008, 32: 3264-3273.
[7] Azman K, Kocijan J. Application of Gaussian processes for black-box modelling of biosystems[J]. ISA Transactions, 2007, 46: 443-457.
[8] Wu Q, Law R, Xu X. A sparse Gaussian process regression model for tourism demand forecasting in Hong Kong[J]. Expert Systems with Applications, 2012, 39: 4769-4774.
[9] Mackay D J C. Introduction to Gaussian processes[J]. NATO ASI series. Series F, Computer and Systems Sciences, 1998, 168: 133-166.
[10] Seeger M, Williams C K I, Lawrence N D. Fast forward selection to speed up sparse Gaussian process regression [C] // The Ninth Workshop on AI and Statistics, Key West, USA, 2003: 2003-2010.
[11] Luo N, Qian F. On line estimation of color values (B^*) in pet process using gaussian process regression [C] // Proceedings of the 8th World Congress on Intelligent Control and Automation, Jinan, 2010: 5842-5845.
[12] 李雅芹, 杨慧中. 基于仿射传播聚类和高斯过程的多模型建模方法[J]. 计算机与应用化学, 2010, 27(1): 51-54.
Li Y Q, Yang H Z. Multi-model modeling method based on affinity propagation clustering and Gaussian processes

- [J]. Computers and Applied Chemistry, 2010, 27(1): 51-54. (in Chinese)
- [13] 林云, 郜丽鹏. 基于灰色关联和证据理论的故障诊断方法[J]. 电子测量与仪器学报, 2009, 23(7): 68-73.
- Lin Y, Gao L P. Fault diagnosis method based on gray correlation and evidence theory[J]. Journal of Electronic Measurement and Instrument, 2009, 23(7): 68-73. (in Chinese)
- [14] Birol G, Ündey C, Çinar A. A modular simulation package for fed-batch fermentation: penicillin production[J]. Computers & Chemical Engineering, 2002, 26: 1553-1565.
- [15] 于振亚, 王闻侠, 潘丰. 模糊支持向量机在青霉素发酵中的应用[J]. 微计算机信息, 2007, 23(7-1): 300-302.
- Yu Z Y, Wang W X, Pan F. Application of fuzzy support machine in penicillin fermentation[J]. Control & Automation, 2007, 23(7-1): 300-302. (in Chinese)

Soft-sensor modeling method in a fermentation process based on the samples of a sparse Gaussian process

HE Kun ZHAO LiQiang WANG JianLin YU Tao

(College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: A soft-sensor model method has been proposed for a fermentation process based on the samples of a sparse Gaussian process (GP). A method based on clustering and grey correlation analysis to select effective sample subsets was employed. This incorporated the Euclidean distance between sample points and the feature vector similarity between the sample points, whilst eliminating the sample points which had larger similarity. The method can give not only the forecast value of the fermentation, but also the forecast precision of the model. The experimental results show that the soft-sensor model method of a sample sparse GP can not only guarantee the accuracy of predictions but also save on training time of the model.

Key words: Gaussian process; sparse samples; affinity propagation clustering; grey relational analysis