DOI:10.3969/j.issn.1673-4785.201310070
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20150113.1130.002.html

基于 FCM 与集成高斯过程回归的赖氨酸发酵软测量

嵇小辅,张翔
(江苏大学 电气信息工程学院,江苏 镇江 212013)

摘　要:为解决赖氨酸发酵过程中菌体浓度难以在线检测的难题,提出一种基于模糊 C 均值聚类(FCM)与集成高斯过程回归(GPR)的软测量建模方法。针对典型生物发酵过程可分为延滞期、指数生长期、稳定期、死亡期 4 个反应周期的特点,采用模糊 C 均值聚类算法对样本集进行聚类分析以形成若干子样本集;对每个子样本集分别采用高斯过程回归训练时,为提高 GPR 模型的泛化能力,利用 Adaboost 算法提升 GPR 模型,分别在各子集建立集成 GPR 软测量子模型;采用欧氏距离计算新样本点对应于每一子模型的隶属度;加权求和获得最终的软测量模型的预测输出。基于氨基酸类典型菌种 L-赖氨酸反应过程菌体浓度参数预测的试验研究表明:与全局单一 GPR 模型、集成 GPR 模型和基于 FCM 与多 GPR 模型相比,所建立的基于 FCM 与集成 GPR 软测量模型拟合精度高,泛化能力强,较好地满足了赖氨酸发酵过程的控制要求。

关键词:高斯过程回归(GPR);模糊 C 均值聚类(FCM);Adaboost 算法;L-赖氨酸;软测量;欧氏距离;隶属度;加权求和
中图分类号: TP274　　**文献标志码**:A　　**文章编号**:1673-4785(2015)01-0156-07

中文引用格式:嵇小辅,张翔.基于 FCM 与集成高斯过程回归的赖氨酸发酵软测量[J]. 智能系统学报, 2015, 10(1):156-162.
英文引用格式:JI Xiaofu, ZHANG Xiang. Soft measurement of lysine fermentation based on FCM and integrated Gaussian process regression [J]. CAAI Transactions on Intelligent Systems, 2015, 10(1): 156-162.

Soft measurement of lysine fermentation based on FCM and integrated Gaussian process regression

JI Xiaofu, ZHANG Xiang
(School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: In order to solve the problem that cell concentration is difficult to directly measure in the lysine fermentation process, a kind of soft measurement modeling method is proposed on the basis of fuzzy *C*-mean clustering (FCM) and integrated Gaussian process regression (GPR). The characteristics of typical biological fermentation process can be divided into 4 reaction cycles, including lag phase, exponential growth phase, stable phase, and dead phase. The cluster analysis is conducted for a sample set by applying fuzzy C-mean clustering algorithm, so as to form several sub-sample sets. In order to improve the generalization performance of the GPR, each group is trained through Gaussian Process Regression based on Adaboost and the corresponding integrated sub-models are established. The memberships between each new sample and each group are set as the weights through Euclidean distance and the predicted result is obtained by weighted sum by using typical bacterium of amino acid—L-lysine fermentation as an example. The simulation results showed that compared with the global single GPR model, integrated GPR model and the model based on FCM and multiple GPR, the soft measurement model based on integrated GPR and FCM has high fitting precision. It also had strong generalization ability, which meets the control requirements of lysine fermentation process

Keywords: Gaussian process regression(GPR); fuzzy *C*-mean clustering (FCM); Adaboost algorithm; L-lysine; soft measurement; Euclidean distance; membership; weighted sum

收稿日期:2013-10-27.　**网络出版日期**:2015-01-13.
基金项目:国家 863 计划资助项目(2011AA09070301);江苏高校建设优势学科工程资助项目(苏政办发[2011]6 号);江苏省科技支撑计划资助项目(BE2010354);江苏省自然科学基金资助项目(BK2011465).
通信作者:张翔.E-mail: zhangxiang_mail@126.com.

生物发酵过程是一个具有高度非线性与不确定性和兼有生物、化学、物理和热力变化的复杂生化反应过程。一些反映发酵过程品质的重要参量,如菌

体浓度、基质浓度、产物浓度等,目前还缺乏在线测量的仪器与手段,影响了在线优化控制的实施。对于这些重要过程参量的测量,目前主要采取在线取样、离线分析的方法,但这种方法时间间隔长、数据滞后大,难以满足生物反应过程在线控制的要求,同时在线取样容易造成反应过程菌体污染,降低反应过程的质量。为了解决这一难题,人们提出软测量的理论与方法。

软测量技术起源于 20 世纪 70 年代 Brosillow 提出的推断控制思想[1],其基本原理是构造以直接可测的辅助变量为输入的数学模型来估计难以直接测量的主导变量。软测量建模方法可分为机理建模、数据驱动建模、混合建模 3 种,其中数据驱动建模是从大量过程数据中提取过程信息从而建立软测量模型,不需要获得对象过程的精确数学模型,它尤其适用于生物反应过程这一类内部机理尚不明确的复杂过程。数据驱动建模主要有神经网络、支持向量机、高斯过程回归(GPR)等方法,其中高斯过程回归将高斯先验分布运用于非参数回归函数空间,通过推导预测目标的后验分布建立统计模型,高斯过程回归可以同时获得预测输出与预测精度,显著提高软测量模型的性能。与神经网络、支持向量机等回归方法相比,高斯过程回归具有优化参数少、收敛速度快、模型精度高、泛化能力强等优点,在实际生物发酵工业中具有较好的应用效果[2-6]。

文中给出一种生物反应过程关键参量的基于模糊 C 均值聚类(FCM)与集成高斯过程回归建模方法。针对生物发酵过程可分为延滞期、指数生长期、稳定期、死亡期 4 个反应周期的特点,采用模糊 C 均值聚类方法将过程数据样本分成 4 个子类,分别对每个子样本集采用高斯过程回归训练,训练过程采用 Adaboost 算法以进一步提高软测量模型的泛化能力。对于新样本,通过计算样本点到各聚类中心的距离确定新样本点对每个 GPR 模型的隶属度,通过加权综合给出软测量模型的预测输出。针对氨基酸类典型菌种 L-赖氨酸反应过程的菌体浓度参量,建立软测量模型并开展相关的仿真实验研究。仿真结果表明,与单一高斯过程模型、集成 GPR 模型和基于 FCM 与多 GPR 模型方法相比,基于 FCM 与集成 GPR 的软测量模型具有更好地逼近精度与泛化能力,较好地满足了赖氨酸发酵过程软测量的要求。

1　高斯过程及其回归算法

1.1　高斯过程原理

高斯过程(GP)是在高斯随机过程与贝叶斯学习理论基础上发展起来的一种新型机器学习方法,它描述了一类随机过程:其任意有限变量集合的分布都是高斯过程,即对任意整数 $n \geqslant 1$ 及任意一族随机变量 $\boldsymbol{X}$,与其对应的 t 时刻过程状态 $f(X)$ 的联合概率分布服从 n 维高斯分布。GP 的全部统计特征完全由其均值 $m(t)$ 与协方差函数 $k(t,t')$ 确定,定义表示如下:

$$f(X) \sim GP(m(t),k(t,t'))$$

式中:t 为时间变量。为了符号描述方便起见,通常对数据作预处理,使其均值函数等于 0。

1.2　高斯过程回归

给定样本训练集 $D=\{(x_i,y_i) \mid i=1,2,\cdots,n\}$,其中 $x_i \in R^d$ 是 d 维输入向量,$y_i \in R$ 是相应的输出量。为了符号描述方便,用 $\boldsymbol{X}$ 表示输入向量构成的 $d \times n$ 维输入矩阵,$\boldsymbol{y}$ 表示输出标量构成的输入矢量,那么训练集可表示为 $\boldsymbol{D}=(\boldsymbol{X},\boldsymbol{y})$。对于新样本 x^*,GP 模型根据先验知识预测与 x^* 相对应的输出值 y^*。

假设观察目标值 y 被噪声腐蚀,它与真实输出值 t 相差 ε,即

$$y = t + \varepsilon$$

式中:ε 为独立的随机变量,满足均值为 0、方差为 σ_n^2 的高斯分布,即

$$\varepsilon \sim N(0,\sigma_n^2)$$

观测目标值 y 的先验分布为

$$y \sim N(0,\boldsymbol{K}+\sigma_n^2 I)$$

式中:$\boldsymbol{K}=\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})=(\boldsymbol{K}_{ij})_{n\times n}$ 为对称正定的协方差矩阵,元素 $\boldsymbol{K}_{ij}$ 度量了 x_i 和 x_j 的相关度。

由此,n 个训练样本输出 $\boldsymbol{y}$ 和 1 个新样本输出 y^* 构成的联合高斯先验分布为

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim N\left(0,\begin{bmatrix} \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\sigma_n^2\boldsymbol{I} & \boldsymbol{K}(\boldsymbol{X},x^*) \\ \boldsymbol{K}(\boldsymbol{X},x^*)^{\mathrm{T}} & k(x^*,x^*) \end{bmatrix}\right)$$

式中:$\boldsymbol{K}(\boldsymbol{X},x^*)$ 是新样本 x^* 与训练集样本 $\boldsymbol{X}$ 的 $n \times 1$ 阶协方差矩阵,$k(x^*,x^*)$ 是新样本 x^* 的自协方差。

GP 的协方差函数需要满足对任一点集都能够保证产生一个非负正定协方差矩阵。常用的协方差函数为

$$k_y(x_p,x_q)=\sigma_f^2\exp\left(-\frac{1}{2l^2}(x_p-x_q)^2\right)+\sigma_n^2\delta_{pq}$$

式中:超参数 l、σ_f、σ_n 对预测效果的影响很大。最优超参数可通过极大似然法获得,即通过建立训练样本条件概率的对数似然函数对超参数求偏导,再采用共轭梯度优化方法搜索出超参数的最优解。该

对数似然函数可以表示为

$$L = \ln p(\boldsymbol{y} \mid \boldsymbol{X}) = -\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{K}+\boldsymbol{I})^{-1}\boldsymbol{y} - \frac{1}{2}\ln|\boldsymbol{K}+\sigma_n^2\boldsymbol{I}| - \frac{n}{2}\ln 2\pi$$

对于新的输入样本 x^*,根据贝叶斯原理可以获得对应的预测值 y^*,其预测分布是高斯型:

$$p(y^* \mid x^*, X, y) \sim N(\hat{y}(x^*), \hat{\sigma}(x^*))$$

式中:均值 $\hat{y}(x^*)$ 和方差 $\hat{\sigma}(x^*)$ 分别为

$$\hat{y}(x^*) = \boldsymbol{k}^{\mathrm{T}}(x^*)(\boldsymbol{K}+\sigma_n^2\boldsymbol{I})^{-1}\boldsymbol{y}$$

$$\hat{\sigma}(x^*) = \boldsymbol{k}(x^*, x^*) - \boldsymbol{k}^{\mathrm{T}}(x^*)(\boldsymbol{K}+\sigma_n^2\boldsymbol{I})^{-1}\boldsymbol{k}(x^*)$$

2 模糊 C 均值聚类算法

FCM 算法[7-11]是由 Dunn 提出,经由 Bezdek 应用发展起来的一种模糊聚类算法,其原理是通过最小化基于范数与聚类原型的目标函数将无标签的数据进行分类。对于给定的样本集合 $X = \{x_1, x_2, \cdots, x_n\} \subset R^s$,其中 s 是样本空间的维数,n 是样本个数,令 $c(c > 1)$ 是对 X 进行划分的聚类个数,则 FCM 算法的优化目标为

$$\min J_{\mathrm{fcm}}(\boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m d_{ij}^2$$

约束条件为

$$\sum_{i=1}^{c} u_{ij} = 1, 1 \leqslant j \leqslant n$$

$$\sum_{j=1}^{n} u_{ij} > 0,\ 1 \leqslant i \leqslant c$$

$$u_{ij} \geqslant 0,\ 1 \leqslant i \leqslant c,\ 1 \leqslant j \leqslant n$$

式中:$m > 1$ 是模糊系数,$\boldsymbol{U} = (u_{ij})_{c\times n}$ 是模糊划分矩阵,u_{ij} 是样本 x_j 属于第 i 类的隶属度值,$\boldsymbol{V} = [v_1 \quad v_2 \quad \cdots \quad v_c]$ 是由 c 个聚类中心向量构成的矩阵;$d_{ij} = \| x_j - v_i \|$ 表示样本点 x_j 到中心 v_i 的欧式距离。本质而言,FCM 算法是一个关于自变量 $(\boldsymbol{U}, \boldsymbol{V})$ 的约束优化问题,利用极值点的 KT 必要条件可以得到对应的迭代方程为

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}, i = 1, 2, \cdots, c \tag{1}$$

记 $I_j = \{(i,j) \mid x_j = v_i, 1 \leqslant i \leqslant c\}$,若 $I_j = \varnothing$,则:

$$u_{ij} = \left(\sum_{r=1}^{c}\left(\frac{d_{ij}}{d_{rj}}\right)^{\frac{2}{m-1}}\right)^{-1}, i = 1, 2, \cdots, c;\ j = 1, 2, \cdots, n$$

若 $I_j \neq \varnothing$,则 u_{ij} 是满足如下条件的任意非负实数:

$$\sum_{i=1}^{c} u_{ij} = 1, u_{ij} = 0, d_{ij} \neq 0$$

关于隶属度的迭代公式是一个从点到集合的映射,在实际计算中通常采用如下的隶属度更新公式:

$$u_{ij} = \begin{cases} \left(\sum_{r=1}^{c}\left|\frac{d_{ij}}{d_{rj}}\right|^{\frac{2}{m-1}}\right)^{-1}, & I_j = \varnothing \\ \frac{1}{|I_j|}, & I_j \neq \varnothing,\ i \in I_j \\ 0, & I_j \neq \varnothing,\ i \notin I_j \end{cases} \tag{2}$$

FCM 算法先初始化类中心或隶属度矩阵,然后利用式(1)和式(2)进行迭代直至满足设定的终止条件,其具体实现步骤如下:

1)设定聚类个数 c 和模糊指数 m;初始化类中心 $V^{(0)}$;设置收敛精度 $\varepsilon > 0$;令迭代次数 $k = 0$。

2)利用式(2)计算 $\boldsymbol{U}^{(k+1)}$。

3)利用式(1)计算 $\boldsymbol{V}^{(k+1)}$,令 $k = k + 1$。

4)重复 2)和 3),直到满足如下的终止条件:

$$\| \boldsymbol{V}^{(k)} - \boldsymbol{V}^{(k-1)} \| \leqslant \varepsilon, k \geqslant 1$$

3 Adaboost 算法

Adaboost(Adaptive boost)算法[12-14]是 Boosting 算法的一种,这种思想源于 Valiant 提出的 PAC[15](probably approximately correct)学习模型。其主要内容是获取各学习样本的权重分布,所有权重被赋予相等的初始数值,但在训练过程中,这些样本权重被不断调整,预测精度低的样本权重得到加强,预测精度高的样本权重则被减弱。最终,弱预测器加强了对难以预测的样本的学习。这样以后,达到一定预测精度的弱预测器,经组合后形成的强预测器就具有很高的预测精度。文中利用此算法对高斯过程回归模型的学习能力进行提升,以提高模型的泛化能力和预测性能。

4 基于 FCM 与集成 GPR 的软测量模型构建

这里以氨基酸类典型菌种 L-赖氨酸反应过程的菌体浓度软测量为例,阐述基于 FCM 与集成 GPR 的软测量模型构建步骤。考虑到赖氨酸反应过程中微生物生长可分为延滞期、指数生长期、稳定期和死亡期 4 个阶段,将预处理后的样本集依据 FCM 算法分成 4 个子类,即 $c = 4$,根据经验取 $m = 2$[16],然后采用基于 Adaboost 算法的集成 GPR 方法对每个子集建立软测量模型。具体建模步骤为

1)根据赖氨酸反应实验过程的实测值建立训练样本 $(\boldsymbol{x}_i, \boldsymbol{y}_i)$,$i = 1, 2, \cdots, k$,其中 $\boldsymbol{x}_i \in \mathbf{R}^d$ 是 d 维输入矢量,表示反应过程中影响菌体浓度的主要因

素;相应的输出标量 $y_i \in \mathbf{R}$,为菌体浓度值。

2)为进一步提高建模精度,对所有数据进行噪声滤波和归一化预处理,得到样本矩阵 $\hat{\boldsymbol{U}} = (\hat{u}_{iq})$。

3)构造模糊关系矩阵 $\boldsymbol{R} = (r_{iq})$,其中 r_{iq} 为描述样本 i 和 q 之间相似程度的系数,采用欧式距离表示:

$$r_{iq} = 1 - c\sqrt{\sum_{k=1}^{m}(x_{iq} - x_{jk})^2} \tag{3}$$

4)计算模糊关系等价矩阵 $\boldsymbol{R}'$。利用平方法对模糊关系矩阵 $\boldsymbol{R}$ 进行 $2,4,\cdots,2^k$ 幂次计算,直到 $\boldsymbol{R}^{2k} = \boldsymbol{R}^{2k-1}$,则得到 $\boldsymbol{R}' = \boldsymbol{R}^{2k}$。

5)在模糊关系等价矩阵 $\boldsymbol{R}'$ 中,根据设定的分类数 $c = 4$,采用 λ-截距阵法得到样本的初始聚类,然后,以各子类平均值为初始聚类中心,记为 $\boldsymbol{V}_1^{(0)}$,$\boldsymbol{V}_2^{(0)}$,$\boldsymbol{V}_3^{(0)}$ 和 $\boldsymbol{V}_4^{(0)}$。

6)用式(3)计算样本与初始聚类中心的近似程度。

7)计算归一化后的样本矩阵 $\hat{\boldsymbol{U}}$ 的隶属度矩阵 $\boldsymbol{U}^{(0)}$。$\boldsymbol{U}^{(0)} = (u_{ij}^{(0)})$,$u_{ij}^{(0)} = r_{ij} / \sum_{j=1}^{4} r_{ij}$。

8)利用 FCM 聚类算法进行迭代,直至达到收敛精度。

9)样本聚类完成后,针对每个子类的样本数据分别训练子模型,结构如图 1 所示。

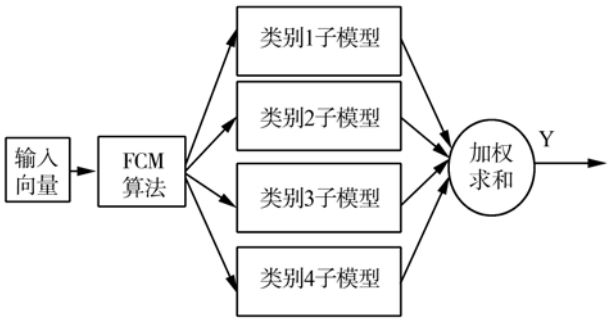


图 1　基于 FCM 与多集成 GPR 软测量模型结构

Fig. 1　Model structure of multiple integrated GPR based on FCM

10)针对训练子模型,采用基于 Adaboost 算法的集成高斯过程回归模型,给定子学习样本为 $(x_1,y_1),\cdots,(x_p,y_p)$,$x_u \in x_i$,$y_u \in y_i$,$u = 1,2,\cdots,p$ 从子类的样本空间中随机选择 d 组训练数据,初始化测试数据的分布权值 $D_t(u) = 1/d$、GPR 模型的最大迭代次数 $t = 1,2,\cdots,T$。

11)利用样本权重 D_t 训练弱 GPR 学习器。

12)获取弱 GPR 学习器的预测函数 $h_t: X \to Y$,并用 $\varepsilon_t = \Pr_{u\sim D_t}[h_t(x_u) \neq y_u]$ 用来表示对应的预测误差。

13)选取 $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$,更新样本权重:

$$D_{t+1}(u) = \frac{D_t(u)}{Z_t} \times \begin{cases} e^{-\alpha_t}, h_t(x_u) = y_u \\ e^{\alpha_t}, h_t(x_u) \neq y_u \end{cases} = \frac{D_t(u)\exp[-\alpha_t y_u h_t(x_u)]}{Z_t}$$

式中:Z_t 为归一化因子,以使 $\sum_{u=1}^{d} D_{t+1}(u) = 1$。

14)输出最终的预测函数为

$$f_{\text{fin}}(x) = \text{sign}\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right)$$

其结构如图 2 所示。

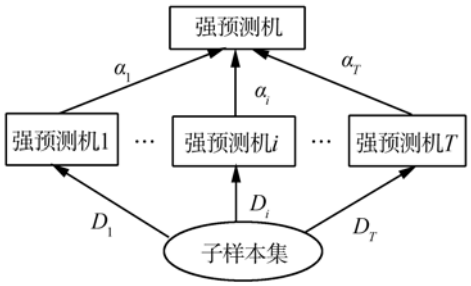


图 2　基于 Adaboost 的 GPR 软测量模型结构

Fig. 2　Model structure of GPR based on Adaboost

15)确定输入数据对于每个子模型的隶属度 u_i,根据隶属度将每个子模型的输出为 $f_{\text{fin}}(x)$,最后得到赖氨酸发酵过程的菌体浓度预测值 Y 为

$$Y = \sum_{i=1}^{c} u_i f_{\text{fin}}(x)$$

5　实验与仿真分析

5.1　实验数据说明

实验采用 WKT-30L 型液态发酵装备及其控制系统。根据赖氨酸发酵过程的工艺要求,发酵过程中的罐压保持在 0.11 MPa,温度控制在 31 ℃,搅拌电机转速为 330 r/min。根据前期对赖氨酸反应过程的机理和数据分析[17-20],赖氨酸反应过程的菌体浓度值与溶解氧值、发酵液 pH 值、空气流等参量紧密相关。因此,本实验选取发酵液 pH 值、溶解氧 Do 与空气流量 F 3 个参量作为辅助变量;关键生物参量菌体浓度 Y 作为主导变量,则赖氨酸发酵过程中菌体浓度软测量模型可描述为

$$Y = f(\text{pH},\text{Do},\text{F})$$

式中:$f(\cdot)$ 表示主导变量与辅助变量之间的复杂非线性关系。

每批次反应周期为 72 h,采样周期为 15 min,通

过测试仪器对发酵液 pH 值、溶解氧值、空气流量参量进行实时采集,每 2 min 取样并离线化验得到菌体浓度。菌体浓度采用细胞干重法计算得到,即取 10 mL 发酵液于离心管中,在 3 000 r · min^{-1}下离心 5 min,弃上清,蒸馏水洗涤 2 次,在 105 ℃干燥至恒质量后称量。

考虑 5 个批次培养数据以检验赖氨酸反应过程的基于 FCM 与集成 GPR 软测量模型。为增强各批次间的差异性,每批次间初始条件不同,补料策略亦有相应变化。采用 WKT-30 L 自动控制系统,将发酵罐温度控制在(0~50±0.5) ℃。获得批次数据的前 4 批次作为训练样本集,离线训练获得基于 FCM 与集成 GPR 软测量模型,第 5 个批次用于测试样本集,检验基于 FCM 与集成 GPR 软测量模型的泛化能力。

数据样本的 3D 聚类效果如图 3 所示,训练样本集被分成 4 类,分别用 4 个椭圆标记。目标函数值的变化曲线如图 4 所示。结果分析可知,整体分类经过 44 次迭代,达到预设收敛精度,得到最终的目标函数值。

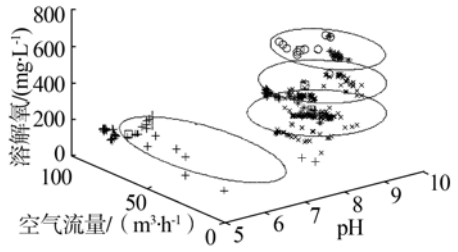


图 3 样本的 3D 聚类结果

Fig.3 3D cluster results in training set

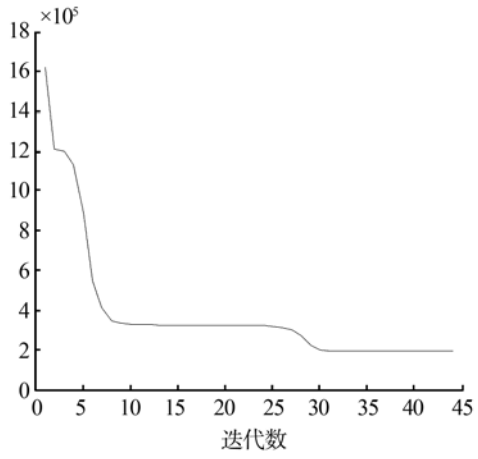


图 4 目标函数值变化曲线

Fig.4 Variation curves of the objective function value

5.2 实验结果分析及讨论

基于 FCM 与集成 GPR 的软测量算法是在 GPstuff- 4.1 基础上用 MATLAB 语言编写。为了验证该种方法的性能,将其与传统的单一 GPR、集成 GPR 软测量模型和基于 FCM 与多 GPR 软测量模型相比,其中选取菌体浓度参量为预测目标,样本数据经过 FCM 聚类以后,分别建立采用 Adaboost 算法的集成 GPR 子模型,式中:设定算法的最大迭代次数 t 为 10,初始误差为 $\varepsilon_t = 0$,针对测试样本集检验软测量模型的泛化能力。图 5 所示为基于 FCM 与集成 GPR 软测量模型的预测值与样本真实值。

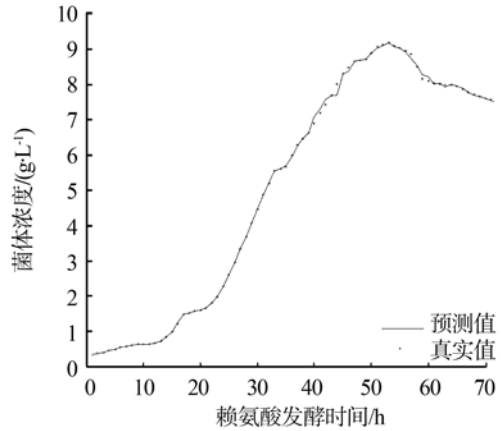


图 5 基于 FCM-Adaboost-GPR 的软测量模型预测值与真实值

Fig.5 Comparison between the predictive value and true value using integrated GPR soft sensor model based on FCM

图 6 显示了弱高斯过程回归模型(传统的单一高斯过程回归模型)与强高斯过程回归模型(集成高斯过程回归模型)针对测试样本集的泛化能力,经过 Adaboost 算法,弱高斯过程回归模型提升为强高斯过程回归模型,误差明显减小,泛化能力也大幅度提高。

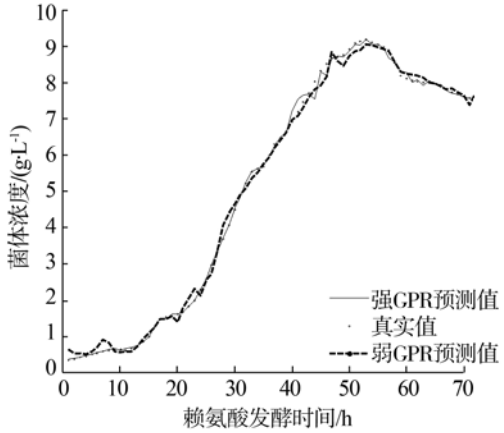


图 6 强 GPR 预测器与弱 GPR 预测器的预测值与真实值

Fig.6 Comparison between the predictive value and true value using strong GPR predictor and weak GPR predictor

为了与其他软测量方法作对比研究,同时建立基于 FCM 与多 GPR 模型和单一 GPR 模型,模型对测试样本集的预测效果如图 7 所示。单一 GPR 模型的预测输出误差大,基于 FCM 与多 GPR 模型的预测输出误差比单一 GPR 模型的预测输出误差小。

通过图 5~7 仿真结果对比发现,基于 FCM 与集成 GPR 模型输出的预测值误差较小,预测精度高且较为平滑。

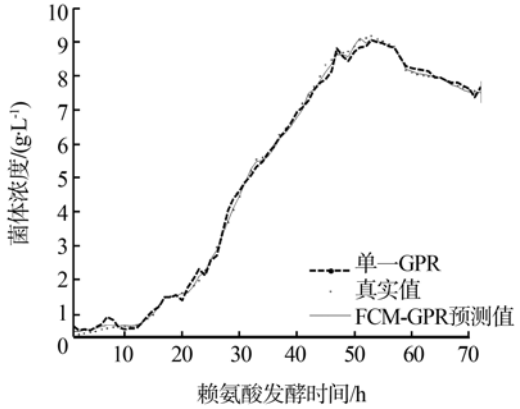


图 7　多 GPR 和单一 GPR 软测量模型预测值与真实值

Fig.7　Comparison between the predictive value and true value using multiple GPR soft sensor model based on FCM and single GPR soft sensor model

基于 FCM 与集成 GPR 软测量模型、集成 GPR 软测量模型、单一 GPR 软测量模型和多 GPR 软测量模型的相对误差对比如图 8 所示。分析表明,多 GPR 软测量模型、集成 GPR 模型和基于 FCM 与集成 GPR 软测量模型总体上相对误差比较小,单一 GPR 软测量模型的相对误差则比较大,同时基于 FCM 与集成 GPR 模型误差明显小于集成 GPR 模型和多 GPR 模型。

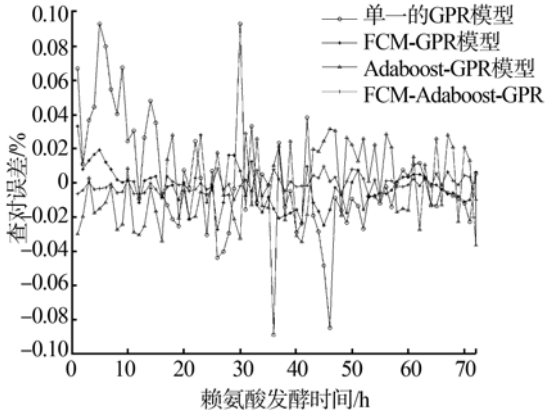


图 8　多 GPR 软测量模型、基于 Adaboost-GPR 软测量模型、单一 GPR 软测量模型和 FCM-Adaboost-GPR 软测量模型的相对误差

Fig.8　Comparison of relative error between multiple GPR soft sensor model based on FCM , GPR soft sensor model based on Adaboost , single GPR soft sensor model and integrated GPR soft sensor model based on FCM.

5.3　算法评价

为了进一步比较分析 4 种模型的预测效果,分别定义平均相对误差 σ_{ARE}、均方根误差 σ_{RMSE} 与最大绝对误差 σ_{MAXE} 如下:

$$\sigma_{\mathrm{ARE}} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

$$\sigma_{\mathrm{RMSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

$$\sigma_{\mathrm{MAXE}} = \max(|\hat{y}_i - y_i|), i = 1,2,\cdots,n$$

式中:n 为预测样本数;$\hat{y}_i$ 为预测结果;y_i 为真实值。

ARE 用来表示回归值符合真实值的平均程度,RMSE 用来表示回归值和真实值的均方差大小,MAXE 用来表示回归值偏离真实值的最大幅度。这 3 个性能指标参数的值越小,回归估计的效果就越好。表 1 所示为 4 种模型分别对应的 3 种误差值,误差值表明,基于 FCM 与集成 GPR 的软测量模型预测效果优于单一 GPR 软测量模型、集成 GPR 软测量模型和多 GPR 软测量模型。对于 3 个性能指标,基于 FCM 与集成 GPR 的软测量模型比单一 GPR 软测量模型、集成 GPR 软测量模型和多 GPR 软测量模型都要小,由此可见,基于 FCM 与集成 GPR 软测量模型预测精度更高,泛化能力更好。

表 1　软测量模型误差对比

Table 1　The error comparison of soft measurement model

模型	ARE	RMSE/g · L^{-1}	MAXE/g · L^{-1}
GPR	0.076 7	0.020 8	0.721 8
FCM-GPR	0.036 5	0.009 5	0.404 9
Adaboost-GPR	0.015 1	0.012 1	0.484 2
FCM-Adaboost-GPR	0.005 1	0.006 7	0.310 8

6　结束语

针对氨基酸类典型菌种 L-赖氨酸发酵过程中菌体浓度参量难以在线测量的难题,提出一种基于 FCM 与集成 GPR 的生物发酵关键参量软测量方法,该方法既有效克服了全局建模模型方法学习时间长,过程特性匹配不佳、预测精度低、外推和自适应能力差等缺点,又利用了 Adaboost 算法提升了高斯过程回归机的学习能力,增强了泛化能力。仿真结果表明,与多 GPR 软测量模型、集成 GPR 软测量模型和单一 GPR 软测量相比,基于 FCM 与集成 GPR 软测量模型具有较高的测量精度和较强的泛化能力,可应用于赖氨酸发酵过程菌体浓度关键参量的在线估计。

参考文献:

[1] BROSILLOW C B. Inferential control of process[J]. Automatica, 1978, 24(3): 485-509.

[2] RASMUSSEN C E, WILLIAMS C K. Gaussian processes for machine learning[M]. Boston: MIT Press, 2006: 7-31.

[3] KOCIJAN J. Control algorithms based on Gaussian process models: a state-of-the-art survey[C]//Proc of the Special International Conference on Complex Systems: Synergy of Control, Communications and Computing. Ohrid, 2011: 69-80.

[4] PETELIN D, KOCIJAN J, GRANCHAROVA A. Online Gaussian process model for the prediction of the ozone concentration in the air[J]. Comptes Rendus del Academie Bulgare des Sciences, 2011, 64(1): 117-124.

[5] HE Z K, LIU G B, ZHAO X J, et al. Temperature model for FOG zero-bias using Gaussian process regression[J]. Advances in Intelligent Systems and Computing, 2012, 180: 37-45.

[6] 孙斌,姚海涛,刘婷.基于高斯过程回归的短期风速预测[J]. 中国电机工程学报, 2012, 32(29): 104-108.
SUN Bin, YAO Haitao, LIU Ting. Short-term wind speed forecasting based on gaussian process regression[J]. Proceedings of the CSEE, 2012, 32(29): 104-108.

[7] BEZDEK J C. Pattern recognition with fuzzy objective function algorithms[M]. New York: Plenum Press, 1981: 15-28.

[8] 曲福恒,胡雅婷,马驷良,等.基于核的模糊 C 均值聚类算法的收敛性定理[J].吉林大学学报:理学版, 2011, 49(6): 1079-1086.
QU Fuheng, HU Yating, MA Siliang, et al. A convergence theorem of kernel based fuzzy C-means clustering algorithm [J]. Journal of Jilin University: Science Edition, 2011, 49(6): 1079-1086.

[9] MOHAMMAD T K, MOHAMMAD H B. A powerful hybrid clustering method based on modified stem cells and fuzzy C-means algorithms[J]. Engineering Applications of Artificial Intelligence, 2013, 26, 1493-1502.

[10] XIAO Fengyin, LI P K, YIH T C. A fuzzy C-means based hybrid evolutionary approach to the clusering of supply chain[J]. Computers & Industrial Engineering, 2013, 66(4): 768-780.

[11] BENAICHOUCHE A N, OULHADJ H, SIARRY P. Improved spatial fuzzy C-means clustering for image segmentation using PSO initialization, Mahalanobis distance and post-segmentation correction[J]. Digital Signal Processing, 2013, 23(5): 1390-1400.

[12] PENG X J, SETLUR S, GOVINDARAJU V, et al. Using a boosted tree classifier for text segmentation in hand-annotated documents[J]. Pattern Recognition Letters, 2012, 33(7): 943-950.

[13] 曹莹,苗启广,刘家辰,等. Adaboost 算法研究进展与展望[J]. 自动化学报, 2013, 39(6): 745-758.
CAO Ying, MIAO Qiguang, LIU Jiachen, et al. Advance and Prospects of AdaBoost Algorithm[J]. Acta Automatica Sinica, 2013, 39(6): 745-758.

[14] CHENG Wenchang, DING M J. A self-constructing cascade classifier with AdaBoost and SVM for pedestrian detection[J]. Engineering Applications of Artificial Intelligence, 2013, 26(3): 1016-1028.

[15] VALIANT L G. A theory of the learnable[J]. Communications of the ACM, 1984, 27(11): 1134-1142.

[16] 仲蔚,俞金寿.基于模糊 C 均值聚类的多模型软测量[J]. 华东理工大学学报, 2000, 26(1): 83-87.
ZHONG Wei, YU Jinshou. Study on soft sensing modeling via FCM based multiple models[J]. Journal of East China University of Science and Technology, 2000, 26(1): 83-87.

[17] 黄永红,孙玉坤,王博,等. 赖氨酸发酵过程关键参数的模糊神经网络逆软测量研究[J]. 仪器仪表学报, 2010, 31(4): 862-867.
HUANG Yonghong, SUN Yukun, WANG Bo, et al. Research of soft sensor based on fuzzy neural network inverse system for lysine fermentation process[J]. Chinese Journal of Scientific Instrument, 2010, 31(4): 862-867.

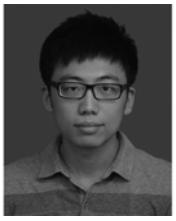
[18] 王博,孙玉坤,嵇小辅. 基于 PSO-SVM 逆的赖氨酸发酵过程软测量[J].化工学报, 2012, 63(9): 3000-3007.
WANG Bo, SUN Yukun, Ji Xiaofu, et al. Soft-sensor modeling for lysine fermentation processes based on PSO-SVM inversion[J]. CIESC Journal, 2012, 63(9): 3000-3007.

[19] 孙玉坤,王博,黄永红.基于聚类动态 LS-SVM 的 L-赖氨酸发酵过程软测量方法[J].仪器仪表学报, 2010, 31(2): 404-409.
SUN Yukun, WANG Bo, HUANG Yonghong, et al. Soft-sensing method for L-lysine fermentation process based on FDLS-SVM[J]. Chinese Journal of Scientific Instrument, 2010, 31(2): 404-409.

[20] 黄丽,孙玉坤,嵇小辅,等. 基于 tPSO-BPNN 的赖氨酸发酵软测量[J]. 仪器仪表学报, 2010, 31(10): 2317-2321.
HUANG Li, SUN Yukun, JI Xiaofu, et al. Soft sensor of lysine fermentation based on tPSO-BPNN[J]. Chinese Journal of Scientific Instrument, 2010, 31(10): 2317-2321.

作者简介:

嵇小辅,男,1979 年生,副教授,博士,主要研究方向为生物反应过程软测量与优化控制、鲁棒控制。



张翔,男,1988 年生,硕士研究生,主要研究方向为生物反应过程软测量与优化控制。