文章编号:1671-4598(2018)03-0224-04 DOI:10.16526/j.cnki.11—4762/tp.2018.03.056 中图分类号:TP18 文献标识码:A

基于集成相关向量机的水质在线预测模型

谭承诚[1], 于广平[2], 邱志成[1]
(1. 华南理工大学 机械与汽车工程学院, 广东 广州 510640;
2. 广州中国科学院 沈阳自动化研究所分所, 广东 广州 511458)

摘要: 针对污水处理过程存在着强非线性和非稳态运行等特征, 传统传感器维护成本高昂且无法快速准确地测量生化需氧量 (BOD) 等水质指标的问题, 提出一种基于集成相关向量机的水质在线预测模型; 该模型首先采用相关向量机 (RVM) 为弱预测器, 利用改进的 AdaBoost.RT 算法将多个弱预测器集成为强预测器, 实现了污水处理过程中水质的在线预测; 仿真实验结果表明, 该水质在线预测模型预测精度高, 综合性能突出, 克服了单一预测器随着异常点增多, 模型泛化能力下降和鲁棒性不足的问题, 能较好地实现了污水处理过程中的水质在线预测。

关键词: 污水处理; 相关向量机; 集成; 在线预测; 鲁棒性

Online Prediction Model of Water Quality Based on Ensemble RVM

Tan Chengcheng[1], Yu Guangping[2], Qiu Zhicheng[1]
(1. School of Mechanical & Automotive Engineering, South China University of Technology, Guangzhou 510640, China;
2. Shenyang Institute of Automation in Guangzhou. Chinese Academy of Science, Guangzhou 511458, China)

Abstract: Wastewater treatment exists strong nonlinearity, unsteady operation and other characteristics, traditional hardware transducer are with huge maintenance problems and make it extremely difficult to obtain water—quality index quickly and accurately, such as BOD. Concerning the concert problems, an online prediction model of water quality based on ensemble RVM is proposed. Firstly, set RVM as weak predictor and then use improved AdaBoost. RT to embody several weak predictor into strong predictor. The simulation experiments demonstrated that this online prediction model has higher precision, better generalization ability, and overcomes the less effectiveness and robust problem of single predictor induced by increasing abnormal points. Therefore, the proposed model can meet the requirements of online prediction of water quality of wastewater treatment process.

Keywords: wastewater treatment; RVM; ensemble; online modeling; robustness

0 引言

污水处理过程工况复杂, 在线测量仪表维护成本高昂, 存在着如毒性物质浓度、化学需氧量 (chemical oxygen demand, COD)、氨氮、生化需氧量 (biological oxygen deman, BOD) 等难以在线测量的水质指标, 而污水处理中各设备间的自动化控制与调度依赖于水质的实时准确测量[1]。现有的常用水质测量方法有减压库伦法、离线分析法、活性污泥快速法、标准稀释法等, 水质的测定周期较长, 不能及时的提供污水处理的实时信息, 为污水处理的实际生产带来不便。预测水质的变化趋势对于掌握污水处理现状具有重要意义, 然而影响水质变化的因素众多, 各因素间又相互影响, 传统方法依托建立精确的生化机理反应模型的来预测水质变化, 难以克服各因素间的强非线性, 预测误差较大。面对这些特点, 国内外学者提出了大量基于数据驱动的污水水质指标软测量方法, 该方法无需了解污水生化反应过程中的复杂机制, 根植于数据本身, 极大的促进了污水处理水质预测的研究。文献 [2-3] 采用神经网络建立污水水质预测模型, 文献 [4-5] 采用支持向量机建立了污水处理过程出水水质的软测量模型, 文献 [6] 采用相关向量机建立了污水水质的在线软测量预测模型, 并引入快速似然边界算法加快了模型的更新速度。然而, 神经网络容易陷入局部最小值和过拟合, 存在着健忘和泛化性能较弱, 权值不易在线调整等缺点[7], 支持向量机随着样本量的增加, 训练时间会变长, 支持向量增多, 稀疏性降低, 且核函数的选择受到 Mercer 条件制约[8], 当样本中存在较少异常值时, 相关向量机可以得到鲁棒性良好的回归模型[9], 但随着异常值的增多, 其泛化能力会下降, 鲁棒性渐失, 同时这些缺陷均影响了污水水质在线预测的可靠性和实时性。

根据现有成果与存在的问题, 本文提出一种基于集成相关向量机的污水水质在线预测模型。该模型以 RVM 为弱预测器, 利用改进的 AdaBoost.RT 算法将多个弱预测器集成为强预测器。RVM 是建立在贝叶斯理论 (Bayesian Principle) 下的稀疏核机, 不受 Mercer 定理的限制, 可以任意选择核函数, 具有较好的泛化能力, 但污水生化处理的过程, 存在着参数时变的现象, 随着异常值的增多, 会使 RVM 的预测精度下降, 针对这一问题, 该模型利用改进的 AdaBoost.RT 算法将 RVM 弱预测器分层组合, 使迭代的重点聚焦与少数异常样本上, 提高了模型的预测精度的鲁棒性, 同时也满足污水水质在线预测的实时性的要求, 并通过仿真实验得到了验证。

1 基于集成相关向量机的水质在线预测模型

1.1 相关向量机回归模型

相关向量机是由 Tipping M E[10] 在稀疏贝叶斯学习理论的

收稿日期:2017-10-21; **修回日期**:2017-11-21。
基金项目:广东省科技项目(2016A020221002)。
作者简介:谭承诚(1994-),男,四川自贡人,硕士研究生,主要从事水质预测与故障诊断方向的研究。

基础上，将最大似然估计、自动相关决策先验和马尔科夫等理论有机结合形成的一种有监督的学习算法。给定污水处理数据集 $\{x_n,t_n\}_{n=1}^N$，其中 x_n 为输入向量，t_n 为输出标量，N 为样本个数。假定输出标量 t_n 含附加噪声，即：

$$t_n = y(x_n,w)+\varepsilon_n \tag{1}$$

其中：权值向量 $w=(w_0,w_1,\cdots,w_n)$，ε_n 为期望为 0，方差为 σ_2 的附加高斯噪声，即 $\varepsilon_n \sim N(0,\sigma^2)$。因此 $p(t_n \mid x) = N(t_n \mid y(x_n),\sigma^2)$，式中 $y(x)$ 定义为：

$$y(x,w) = \sum_{i=1}^{N} w_i K(x,x_i)+w_0 \tag{2}$$

式中，$K(x,x_i)$ 为核函数。其中常用的全局性核函数有多项式核函数，如式（3）和常用的局部性核函数有径向基核函数，如式（4）。

$$K(x,y) = (x^T y+c)^q, q \in N, c \geqslant 0 \tag{3}$$

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\delta^2}\right), \delta > 0 \tag{4}$$

其中：c 为常数，q 表示多项式次数，δ 为核宽度。

由于 t_n 是相互独立的，因此训练样本集的似然函数为：

$$p(t \mid w,\sigma^2) = (2\pi\sigma^2)^{-N/2}\exp(-\frac{1}{2\sigma^2}\mid t-\Phi w\mid^2) \tag{5}$$

其中：$t=(t_1,t_2,\cdots,t_n)^T$，$\Phi=[\varphi(x_1),\varphi(x_2),\cdots,\varphi(x_n)]^T$，$\varphi(x_n)=[1,K(x_n,x_1),\cdots,K(x_n,x_T)]^T$。

为避免 RVM 回归模型出现过度拟合（over－fitting），RVM 定义一个高斯先验概率来约束权值向量 w：

$$p(w \mid \alpha) = \prod_{j=0}^{N} N(w_j \mid 0,a_j^{-1}) \tag{6}$$

其中：超参数 $\alpha=(a_0,a_1,\cdots,a_N)^T$，由式(6)可以看出每一个权值 w_i 对应唯一的超参数 α_i，参数受到先验分布的影响，通过对样本集的不断训练，大部分超参数 α_i 将趋于无穷大，其对应的权值 w_i 会趋于 0，从而确保了相关向量机的稀疏性。

根据贝叶斯准则，定义了先验概率，可得到后验概率：

$$p(w,\alpha,\sigma^2 \mid t) = \frac{p(t \mid w,\alpha,\sigma^2)}{p(t)} \tag{7}$$

当输入一个新样本 x_*，预测相应输出 t^* 预测分布为：

$$p(t^* \mid t) = \int p(t^* \mid w,\alpha,\sigma^2)\ p(w,\alpha,\sigma^2 \mid t)dwd\alpha d\sigma^2 \tag{8}$$

也可以得到权重 w 的后验概率分布为：

$$p(w \mid t,\alpha,\sigma^2) = \frac{p(t \mid w,\sigma^2)p(w \mid \alpha)}{p(t \mid \alpha,\sigma^2)} =$$

$$(2\pi)^{-\frac{N+1}{2}} \mid \sum \mid^{-\frac{1}{2}}\exp(-\frac{1}{2}(w-\mu)^T\sum{}^{-1}(w-\mu)) \tag{9}$$

其中后验协方差 $\sum$ 和均值 μ 分别为：

$$\sum = (\sigma^{-2}\Phi^T\Phi+A)^{-1} \tag{10}$$

$$\mu = \sigma^{-2}\sum\Phi^T t \tag{11}$$

式中，$A = diag(\alpha_0,\alpha_1,\cdots,\alpha_N)$。基于最大期望超参数估计，多次迭代计算后可得：

$$(\alpha_i)^{new} = \gamma_i/\mu_i{}^2 \tag{12}$$

$$(\sigma^2)^{new} = \frac{\mid t-\Phi\mu\mid^2}{N-\sum_i\gamma_i} \tag{13}$$

式中，μ_i 为第 i 个后验平均权值；γ_i 定义 $\gamma_i \equiv 1-\alpha_i\sum_{ii}$。

针对权值后验概率分布的预测而言，其限制条件 α_{MP}，σ^2_{MP} 均取最大值，根据正态分布的性质可知，$p(t^* \mid t)$ 服从正态分布，所以有：

$$p(t^* \mid t,\alpha_{MP},\sigma^2_{MP}) = N(t^* \mid y^*,\sigma^2_*) \tag{14}$$

其中：

$$\sigma^2_* = \sigma^2_{MP}+\Phi(x^*)^T\sum\Phi(x^*) \tag{15}$$

$$y^* = \mu^T\Phi(x^*) \tag{16}$$

式中，y^* 为 t^* 的预测值。

1.2　改进的 AdaBoost．RT 算法

集成算法[11]的基本思想是将多个仅比随机猜测略好的弱预测器进行迭代融合从而使得到的强预测器具有更高的预测精度和泛化能力。与传统的集成算法 AdaBoost 相比，AdaBoost．RT 算法的主要区别在于引进了固定阈值 φ，通过与训练误差的对比，确定权值更新的方式，它解决了最初 boosting 算法在训练集中损失函数超过 0.5 便停止和不能任意设置迭代次数的缺陷。AdaBoost．RT 算法的具体计算过程如下：

1）输入：m 个样本、弱预测器算法 f_t、迭代次数 T、阈值 $\varphi(0<\varphi<1)$。

2）初始化：训练样本的初始权重为 $D_t(i)=\frac{1}{m}$，误差率 $\varepsilon_t=0$。

3）训练：当 $t<T$ 时，利用若预测器算法进行预测，得到样本误差向量：

$$E_t = \mid (f_t(x_i)-y_i)/y_i \mid \tag{17}$$

根据式（18）计算弱预测器 f_t 的误差率：

$$\varepsilon_t = \sum_{i:E_t(i)>\varphi} D_t(i) \tag{18}$$

令 $\beta_t=\varepsilon_t^n$，其中 $n=1,2$ 或 3（在本文后续建模中 n 取 1）。根据式（18）更新权重：

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t}\times\begin{cases}\beta_t, & E_t(i)\leqslant\varphi\\ 1, & others\end{cases} \tag{19}$$

式中，Z_t 是归一化因子。

4）输出预测结果：结果 T 轮的训练后的得到的强预测器 F 由 T 个弱预测器根据误差率加权平均得到，如式（20）。

$$F(x) = \sum_t\left(\log\frac{1}{\beta_t}\right)f_t(x)/\sum_t\left(\log\frac{1}{\beta_t}\right) \tag{20}$$

由上述计算步骤可以看出，AdaBoost．RT 算法性能对于阈值 φ 的选择比较依赖，若阈值 φ 的取值过大将难以实现对异常数据的重点学习，若取值过小则无法将足够的样本进行充分的学习，实际应用中通常需要反复实验才能最终确定阈值 φ 的取值，阈值 φ 的取值对于模型的性能至关重要。针对这一问题，对 AdaBoost．RT 进行改进：将固定的阈值 φ 改为自适应型阈值，使阈值 φ 在集成学习的过程中根据误差的变化而变化，即当误差 ε_t 大于误差 ε_{t-1} 时，增大阈值 φ，当误差 ε_t 小于误差 ε_{t-1} 时，减小阈值 φ，加强了对难以预测的样本的学习，最终提高了模型最终的预测精度和泛化能力。具体做法是在每次迭代计算后，计算弱预测器的均方根误差（root mean square error，RMSE），如式（21）：

$$e = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i-y_i)^2} \tag{21}$$

然后按式（22）对阈值 φ 进行调整：

$$\varphi_t = \begin{cases}\varphi_t(1-\alpha), & e_t<e_{t-1}\\ \varphi_t(1+\alpha), & e_t\geqslant e_{t-1}\end{cases} \tag{22}$$

其中：α 的计算公式为：

$$\alpha = \frac{1}{2}\left|\frac{e_t - e_{t-1}}{e_t}\right| \tag{23}$$

所以，样本权重的更新公式在改进 AdaBoost. RT 算法中被修改为：

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \beta_t, \\ 1/\beta_t, \end{cases} \tag{24}$$

文献［12］指出，当改进的 AdaBoost.RT 算法的阈值 φ 初始设置大于 0.4 时，模型对阈值 φ 的变化十分敏感，预测误差变化较大，模型的鲁棒性不足，而当初始阈值 φ 设置在（0，0.4）之间时，模型的预测误差输出较为稳定，也较好地避免了模型训练过程中阈值 φ 陷入局部最优。

1.3 建模步骤

污水处理具有复杂的生化反应过程和非稳态运行等特征，不同的水质指标受到不同的特征的影响，本文以预测生化需氧量 BOD 为例，建立水质在线预测模型。BOD 定义为 20℃时，污水中好氧微生物将有机污染物氧化分解时所需氧气量，是反应污水中可降解有机物的指标，BOD 越大则污染情况就越严重[13]。根据 BOD 的特性选择与之相关联的水质特征组成数据集，为消除不同特征间幅值对预测精度的影响，对污水水质数据进行归一化处理。在线仿真实验中模型更新时，为保持训练样本集的容量不变，选择限定记忆法来确保每次都有新的样本增加，同时删除最早的样本数据。基于集成 RVM 的水质预测模型流程简图如图 1 所示。

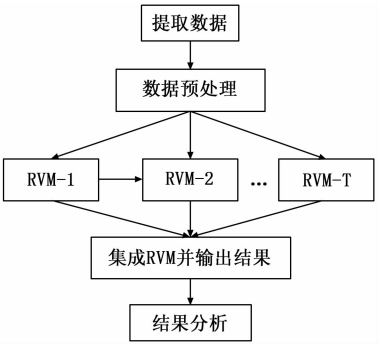

图 1 集成 RVM 模型建模流程图

2 实验仿真与结果分析

2.1 实验数据

本文所用污水处理过程数据来源于加州大学欧文分校机器学习数据库，数据为某城市污水厂 2 年间实际生产所测得数据，采样时间最大间隔为 2 天，共包含 527 个样本，每个样本具有 38 个特征，其中不含缺失值的样本有 380 个，具有 13 种运行状态。进行数据预处理时，将数据集中不满足 3σ 原则的异常值所在样本和具有缺失值的相邻样本删除，然后对缺失值进行拉格朗日插值填补。选取与 BOD 相关的输入特征，分别为整个污水厂的 BOD 输入、COD 输入、悬浮固体浓度和可降解固体浓度，初沉池的 BOD 和固体悬浮物浓度 Ss，二沉池的 BOD 和 COD，二级沉降器 BOD、COD、pH、固体悬浮物浓度和可降解固体浓度以及输出的 COD、BOD、悬浮固体浓度与可降解固体浓度，最终得到 420 组样本数据。

2.2 性能评价指标

污水处理过程的水质预测本质是一个回归问题，为评价各水质预测模型的综合性能，采用均方根误差 RMSE，平均相对误差 mean－re，最大相对误差 max－re 和每次模型更新的平均用时 time 四个指标作为模型综合性能的评价标准。其中：

$$mean - re = \frac{1}{N}\left(\sum_{i=1}^{N}\left|\frac{F(x_i) - y_i}{y_i}\right|\right) \times 100\%$$

$$\max - re = \max\left(\left|\frac{F(x_i) - y_i}{y_i}\right|\right) \times 100\%$$

2.3 在线仿真实验

仿真实验中计算机环境为：Windows 10 操作系统，Intel Core i5 处理器，主频为 3.2 GHz，8 G 内存，240 G 固态硬盘，采用 Matlab 2016b 软件编程实现。选取最终得到的 420 组污水水质的样本数据，其中 320 组用于模型的训练，余下的 100 组作为测试集来验证模型的综合性能。分别建立基于支持向量机 SVM、相关向量机 RVM 和集成相关向量机 RVM 的水质指标 BOD 的在线预测模型。所建模型均采用径向基核函数，如式（4）。SVM 的惩罚参数 c 和核参数 g_1 采用五折交叉验证和遗传算法寻优确定，RVM 与集成 RVM 的核参数 g_2 与 g_3 亦采用五折交叉验证法和遗传算法寻优确定，g_3 由首次寻优确定后，在后续弱预测器的迭代中不变，集成 RVM 模型的弱预测器个数设置为 5 个。其中，SVM 模型的预设迭代次数为 300 次，RVM 模型的预设迭代次数为 500 次，原因在于 SVM 的目标函数是一个凸二次规划求解问题，而 RVM 模型的目标函数非凸需要更多的迭代次数，考虑到集成 RVM 的迭代所需为弱预测器，其迭代次数预设为 300 次。

进行后续水质在线预测时，采用同上述一样的步骤方法以及限定记忆法对模型进行更新。本文中各模型的仿真实验均进行 10 次，得到的数据为 10 次实验结果的平均值，3 种模型的水质在线预测性能评价指标如表 1 所示，图 2～图 4 分别为 BOD 的实际数据与各模型的预测结果对比图。

表 1 3 种模型的水质预测结果

Model	RMSE	mean－re	max－re	time/s
SVM	2.306	0.110	0.579	8.37
RVM	2.347	0.103	0.359	1.58
RVM－A	1.857	0.076	0.278	4.72

注：RVM－A 表示集成 RVM 模型

从表 1 中可以看出，在具有强非线性和非稳态和参数时变的污水处理环境中，集成 RVM 模型的预测精度最高，均方根误差 RMSE 为 1.857，平均相对误差 mean－re 为 0.076，最大相对误差 max－re 为 0.278，相较于单一预测器的 RVM 和 SVM 模型，均方根误差 RMSE 分别下降了 19.5％和 20.9％，平均相对误差 mean－re 分别下降了 30.9％和 26.2％，最大相对误差 max－re 分别下降了 52.0％和 22.6％。

在 3 种水质预测模型中，SVM 的训练时间最长，原因在于 SVM 有惩罚系数 c 和核宽度 g 两个参数需要寻优确定，而 RVM 仅需寻优确定核宽度一个参数，RVM 的惩罚系数 c 会在训练过程中自动生成，而在 SVM 中惩罚系数 c 是平衡经验风险和置信区间的重要参数，需要人工设置，显然一维参数比二维参数寻优的时间复杂度要小很多[14]。由于 RVM 比 SVM 的稀疏性更好，在训练与预测中，RVM 极大地降低了核函数的计算量，模型训练完成后，对于新样本的预测所需要时间也更短。

从图 2～图 4 中各模型的预测对比图可以看出，集成 RVM 模型对于 BOD 的变化追踪效果最好，能更好的适应新的工况，

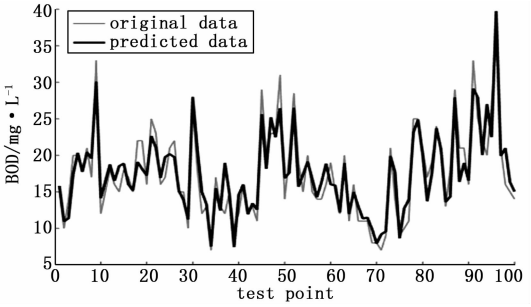

图 2　SVM 在线预测结果

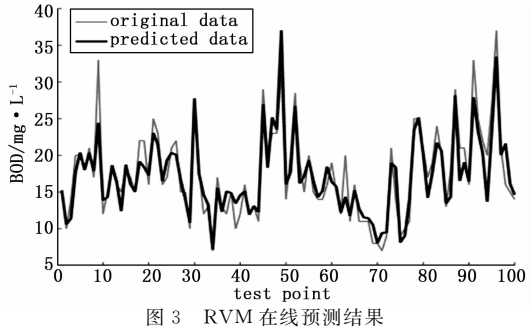

图 3　RVM 在线预测结果

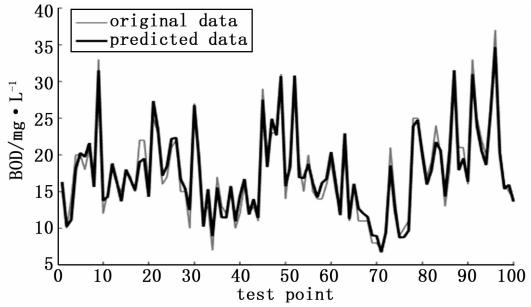

图 4　集成 RVM 在线预测结果

具有最佳的模型鲁棒性和泛化性能。图 5 为集成 RVM 模型和 RVM 模型的预测绝对误差对比图，可以看出：集成 RVM 模型对于绝大多数测试样本的预测误差低于单一 RVM 模型，特别是对异常样本的预测，集成 RVM 在多轮的迭代过程中针对异常值的重点学习在此情况下体现出了极大的优势。从表 1 可以看出，集成 RVM 的模型更新速度比 RVM 模型慢，原因在于多出了集成多个弱预测器的过程，从实际污水处理过程中对水质在线预测的要求考虑，模型的预测精度和泛化能力的优先级要略高于模型对时间的要求。

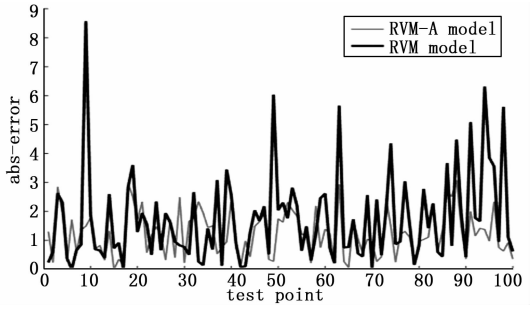

图 5　绝对误差对比图

结合以上分析，基于集成 RVM 的污水处理水质在线预测模型的综合性能优于其他模型，满足实际污水处理过程中对于水质预测准确性和实时性的要求。

3　结论

污水处理是一个复杂的生化反应过程，具有强非线性和参数时变等特征，传统的传感器难以测量 BOD 等反应水质情况的关键指标，并考虑到单一弱预测器难以在异常值增加的情况下获得良好的预测精度和鲁棒性，因此本文提出一种基于改进 AdaBoost. RT 算法集成相关向量机 RVM 的污水处理水质在线预测模型。通过与 SVM 和 RVM 预测模型的对比，仿真实验结果表明：集成 RVM 模型具有预测精度高，能够很好的适应 BOD 的变化，同时模型更新速度快，在异常值增多的情况下依然具有良好的鲁棒性等优点，较好地克服了复杂的污水处理环境带来的困难，完全满足水质在线预测的实际要求。

参考文献：

[1] 黄道平，刘乙奇，李　艳. 软测量在污水处理过程中的研究与应用[J]. 化工学报，2011，62(1)：1－9.

[2] 张瑞成，王　宇，李　冲. 基于 NW 型小世界人工神经网络的污水出水水质预测[J]. 计算机测量与控制，2016，24(1)：61－63.

[3] Pai T Y，Wan T J，Hsu S T，et al. Using fuzzy inference system to improve neural network for predicting hospital wastewater treatment plant effluent[J]. Computers & Chemical Engineering，2009，33(7)：1272－1278.

[4] 黄银蓉，张绍德. MIMO 最小二乘支持向量机污水处理在线软测量研究[J]. 自动化与仪器仪表，2010，30(4)：15－17.

[5] Chen Z M，Hu J. Wastewater treatment prediction based on chaos－GA optimized LS－SVM[C]. /Proceedings of the 2011 Chinese Control and Decision Conference. Mianyang：China Academic Journal Electronic Publishing House，2011：4021－4024.

[6] 许玉格，刘　莉，曹　涛. 基于 Fast－RVM 的在线软测量预测模型[J]. 化工学报，2015，66(11)：4540－4545.

[7] 冉维丽，乔俊飞. 基于 PCA 时间延迟神经网络的 BOD 在线预测软测量方法[J]. 电工技术学报，2004，19(12)：78－82.

[8] Pani A K，Mohanta H K. Application of support vector regression，fuzzy inference and adaptive neural fuzzy inference techniques for online monitoring of cement fitness[J]. Powder Technology，2014，264：484－497.

[9] 杨　飚，周　阳. 一种改进的相关向量机回归方法[J]. 科学技术与工程，2015，15(2)：241－245.

[10] TIPPING M E. Sparse Bayesian learning and the relevance vector machine[J]. Journal of Machine Learning Research，2001，1(3)：211－244.

[11] 毛志忠，田慧欣，王　琰. 基于 AdaBoost 混合模型的 LF 炉钢水终点温度软测量[J]. 仪器仪表学报，2008，29(3)：662－667.

[12] Solomatine D P，Shrestha D L. AdaBoost. RT：A boosting algorithm for regression problems[C]. IEEE International Joint Conference on Neural Networks，Piscataway，2004：1163－1168.

[13] Yu G P，Yuan M Z，Wang H. On simplified model for activated sludge wastewater treatment process and simulation based on benchmark[C]. Proceedings of the 26th Chinese Control Conference，2007：182－186.

[14] Masuda K. Global optimization of point search by equilibrium search of gradient dynamical system [J]. Electronic and Communication in Japan，2008，91(1)：19－31.