

文章编号:1671-6833(2013)03-0112-05

## 基于 PSO-WSVR 的短期水质预测模型研究

徐龙琴<sup>1</sup>, 刘双印<sup>1,2,3</sup>

(1. 广东海洋大学 信息学院, 广东 湛江 524088; 2. 中国农业大学 现代精细农业系统集成研究教育部重点实验室, 北京 100083; 3. 中国农业大学 北京市农业物联网工程技术研究中心, 北京 100083)

**摘要:** 针对传统方法很难建立精确的非线性水质预测模型的情况, 提出了基于粒子群优化加权支持向量回归机(PSO-WSVR)的水质短期预测模型. 在建模过程中, 根据各样本重要性的差异, 给各个样本的惩罚系数赋予不同权重, 改进了标准支持向量回归机算法, 克服了标准支持向量回归算法因不同样本均采用相同权重造成预测精度低的问题, 并采用粒子群优化算法对加权支持向量回归机参数组合进行自适应优化, 模型收敛速度明显加快. 运用 PSO-WSVR 模型对江苏宜兴市集约化河蟹养殖池塘水质进行预测, 与标准支持向量回归机和 BP 神经网络对比分析. 结果表明, 该模型性能可靠、泛化能力强, 预测精度高, 为集约化水产养殖水质短期预测提供了一种新思路.

**关键词:** 水质预测; 加权支持向量回归机; 粒子群优化算法; 参数优化

**中图分类号:** X83 **文献标志码:** A **doi:**10.3969/j.issn.1671-6833.2013.03.027

### 0 引言

集约化水产养殖水体是水产品的栖息场所, 水质的好坏直接影响着水产品的质量和产量. 对养殖水质参数进行精准预测, 及时掌握水质动态变化趋势, 具有重要的理论价值和现实意义. 因养殖池塘水质受多方面因素的影响, 具有多变量、非线性、大时滞、模糊不确定性特点, 利用传统的方法很难建立精确的非线性预测模型<sup>[2]</sup>. 为此, 国内外学者采用回归分析<sup>[3]</sup>、灰色理论<sup>[4]</sup>、时间序列<sup>[5]</sup>、模糊推理<sup>[6]</sup>、神经网络<sup>[7]</sup>等方法, 对总磷、总氮、叶绿素、水温、COD、溶解氧、氨氮、PH 值等水质参数预测已做了大量的工作, 并取得了一定的成效. 但上述方法以基于样本无限大的假设或经验风险最小化原则为基础, 存在易陷入局部极小点、过学习与欠学习、维数灾难、泛化能力差等缺陷, 尤其对小样本数据进行预测时, 其预测结果很难令人满意.

支持向量回归机(SVR)以统计学习理论和结构风险最小化原理为基础, 较好地解决了传统预

测方法难以解决的问题<sup>[8-9]</sup>. 但是, 作为一种新型的机器学习方法, SVR 参数以及核函数参数选取的是否恰当严重影响着模型的预测精度和性能. 传统的 SVR 参数选取方法主要有试凑法、经验选择法、Bayesian 法、交叉验证法等, 但传统方法通常凭借专家经验进行, 计算量大, 时间复杂度高, 不易获取适宜的参数组合<sup>[10]</sup>. 为获取较优的参数组合, 提高 SVR 模型预测精度, 减少人为主观因素对预测性能的影响, 一些学者开始采用遗传算法(GA)<sup>[11]</sup>、蚁群算法(ACA)<sup>[12]</sup>和粒子群优化算法(PSO)<sup>[13-14]</sup>等算法对 SVR 参数进行优化组合, 以提高模型的性能. 与 GA 和 ACA 相比<sup>[14]</sup>, PSO 没有交叉和变异操作, 且以信息共享机制为基础, 每个粒子可以直接利用全局信息进行并行性全局搜索, 具有调节参数少、快速、实现简单等优点, 很适合于 SVR 参数优化.

针对集约化水产养殖水域环境的特点, 以及监测水质数据中距预测时刻的近期数据较远期历史数据对预测值影响较大, 提出了基于 PSO 和加权支持向量回归机(weighted support vector regres-

收稿日期:2012-11-20; 修订日期:2013-01-07

基金项目:国家科技支撑计划资助项目(2011BAD21B01); 广东省科技计划项目(2012A020200008); 湛江市科技计划项目(2010C3113011).

作者简介:徐龙琴(1977-), 女, 陕西汉中, 广东海洋大学讲师, 硕士, 主要从事数据库与人工智能、智能信息处理、数据库安全等研究, E-mail:xlqlw@126.com.

sion, WSVR)的短期水质预测模型.首先以观测数据对预测结果贡献不同应赋予不同权重为指导思想,将标准 SVR 改进为加权支持向量回归机,对近期数据的误差错误采取更高的惩罚强度,有效提高了算法精度;运用 PSO 算法对 WSVR 参数组合进行优化,实现 WSVR 模型参数的自适应选取;以江苏省水产养殖智能控制技术标准化示范区——宜兴市集约化河蟹养殖池塘水质数据为基础,将本文算法与标准 SVR 和 BP 神经网络进行对比分析.试验结果表明,本文算法具有较高的预测精度和泛化能力,非常适合于集约化水产养殖水质预测.

### 1 加权支持向量回归机

在水产养殖水质预测领域,输入与输出变量之间存在随时间发生周期性变化的关系,且距离预测时刻越近的历史数据对预测结果影响越大,故对近期观测数据应赋予更高的权重.基于此,在标准支持向量回归机的基础上,通过引入一个适当的权重函数  $\mu_i$ ,实现随着时间临近对数据的错误加以越来越严重的惩罚强度,以调整输入数据在回归模型中的作用<sup>[15]</sup>.根据结构风险最小化原则,其加权支持向量回归机的优化形式如下:

$$\min_{w, b, \xi_i^*} \frac{1}{2} \|w\|^2 + \sum_{i=1}^n \mu_i C(\xi_i + \xi_i^*), \quad (1)$$

$$\text{s. t.} \begin{cases} (w \cdot \theta(x_i)) + b - y_i \leq \varepsilon + \xi_i, \\ i = 1, 2, \dots, n; \\ y_i - (w \cdot \theta(x_i)) - b \leq \varepsilon + \xi_i^*, \\ i = 1, 2, \dots, n; \\ \xi_i, \xi_i^* \geq 0, i = 1, 2, \dots, n. \end{cases} \quad (2)$$

$$\mu_i = \frac{i}{n(n+1)/2}. \quad (3)$$

式中: $w$ 为权重, $b$ 为偏置量; $C$ 为大于0的惩罚因子, $\xi_i, \xi_i^*$ 为非负松弛变量; $\varepsilon$ 为不敏感损失函数参数; $x_i \in R^n$ 是输入向量; $y_i \in R$ 为相应的目标输出值.根据 Karush-Kuhn-Tucker (KKT) 条件,引入拉格朗日函数,回归问题可以转化为其相应的对偶形式:

$$\min_{a_i^* \in R^{2n}} \frac{1}{2} \sum_{i,j=1}^n (a_i^* - a_i)(a_j^* - a_j)K(x_i, x_j) + \varepsilon \sum_{i=1}^n (a_i^* + a_i) - \sum_{i=1}^n (a_i^* - a_i)y_i; \quad (4)$$

$$\text{s. t.} \begin{cases} \sum_{i=1}^n (a_i^* - a_i) = 0; \\ a_i, a_i^* \in [0, \mu_i C]. \end{cases} \quad (5)$$

式中: $K(x_i, x_j) = (\theta(x_i) \cdot \theta(x_j))$ 是核函数.将

WSVR 与标准 SVR<sup>[8]</sup>对比发现,只有  $a_i$  和  $a_i^*$  的约束发生改变,而目标函数并没有变化.常用核函数有多项式核、Sigmoid 核、线性核、Fourier 级数核与高斯径向基核等.与其他核函数相比,高斯径向基核函数具有模型参数少,计算难度低、效率高易于实现等优点.笔者在本模型中采用高斯径向基函数作为核函数  $K(x_i, x_j) = \theta(x_i)^T \theta(x_j) = \exp(-\|x_i - x_j\|^2 / \sigma^2)$ .由式(5)可求出:

$$b = \frac{1}{N_{NSV}} \left\{ \sum_{0 < a_i < \mu_i C} [y_i - \sum_{x_j \in SV} (a_j^* - a_j)K(x_j, x_i) - \varepsilon] + \sum_{0 < a_i^* < \mu_i C} [y_i - \sum_{x_j \in SV} (a_j - a_j^*)K(x_j, x_i) - \varepsilon] \right\}. \quad (6)$$

式中: $SV$ 为标准支持向量集合; $N_{NSV}$ 为标准向量的数量.加权支持向量回归机的决策函数表达式为

$$y = f(x) = \sum_{i=1}^n (a_i - a_i^*)K(x_i, x) + b. \quad (7)$$

### 2 基于 PSO-WSVR 的水质预测建模

#### 2.1 粒子群优化算法

粒子群优化算法是由 Kenney 和 Eberhart<sup>[16]</sup>于1995年提出的一种基于种群并行搜索全局最优的算法,通过群体中粒子之间的合作与竞争使群体达到最优. PSO 的数学描述为:设种群规模为  $n$ ,第  $i$  个粒子在  $m$  维搜索空间的位置表示为  $x_i = (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{im})$ ,飞行速度  $v_i = (v_{i1}, v_{i2}, \dots, v_{ij}, \dots, v_{im})$ ,迄今为止搜索到的个体最优位置为  $p_i = (p_{i1}, p_{i2}, \dots, p_{ij}, \dots, p_{im})$ ,整个粒子群全局最优位置为  $pg_{besti} = (pg_{best1}, pg_{best2}, \dots, pg_{bestm})$ .可根据式(8)和(9)来更新粒子的速度和位置:

$$v_{ij}^{t+1} = \omega \cdot v_{ij}^t + c_1 \cdot r_1 \cdot (p_{ij} - x_{ij}^t) + c_2 \cdot r_2 \cdot (pg_{bestj} - x_{ij}^t). \quad (8)$$

$$x_{ij}^{t+1} = x_{ij}^t + v_{ij}^{t+1}. \quad (9)$$

其中, $i = 1, 2, \dots, n; j = 1, 2, \dots, m; c_1, c_2 > 0$ ,分别为个体学习因子和社会学习因子; $t$ 为当前迭代次数, $r_1$ 和 $r_2$ 为在 $[0, 1]$ 范围内均匀分布的随机数. $\omega$ 是惯性权重系数,用来控制历史速度对当前速度的影响.为了平衡全局和局部搜索能力,使 $\omega$ 随迭代次数的增加线性递减,能够显著改善 PSO 算法的收敛性能.

$$\omega = \omega_{\max} - t \times \frac{\omega_{\max} - \omega_{\min}}{t_{\max}}. \quad (10)$$

式中: $\omega_{\max}$ 为初始惯性权重; $\omega_{\min}$ 为最后惯性权重;

$t_{max}$ 为最大迭代次数. 飞行速度  $v_i \in [-v_{max}, v_{max}]$ , 该限制条件可有效防止粒子速度过快错过最优解, 通过对算法的改进进一步提高了粒子群算法的全局搜索能力.

### 2.2 基于 PSO 的 WSVR 参数组合优化

研究发现惩罚因子  $C$ 、不敏感损失函数  $\epsilon$  以及核函数的参数  $\sigma$  决定着 WSVR 回归模型的性能<sup>[10]</sup>. 为了提高预测性能, 获得最佳的  $C$ 、 $\epsilon$  和  $\sigma$  参数组合是关键. 关于 SVRM 模型参数优化组合, 目前尚未有有效的方法, 常通过交叉验证试算或梯度下降法求解, 耗时且人为影响较大. 为此, 笔者采用 PSO 算法对 WSVR 的参数进行自动优化选择, 不仅克服人为选择的随机性, 还可以通过粒子适应度函数的设置, 实现参数组合自动选择的目的. 笔者选择能直接反映 WSVR 模型性能的均方误差(MSE)的倒数作为 PSO 算法的适应度函数  $Fitness()$ , 其表达式如下所示:

$$Fitness(C, \sigma^2, \epsilon) = \frac{1}{\sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}} \quad (11)$$

其中,  $y_i$ 、 $\hat{y}_i$  分别为真实值和预测值.

基于改进 PSO 的 WSVR 参数组合优化算法步骤描述如下:

Step1: 粒子群( $C, \sigma^2, \epsilon$ )初始化. 设置粒子数规模  $n$ 、循环迭代最大次数  $t_{max}$ 、惯性权重  $\omega$  的范围, 粒子速度  $v$  的限定范围, 学习因子  $c_1$  和  $c_2$  等参数, 并随机产生一组粒子的初始速度和位置.

Step2: 用训练集来训练 WSVR, 用式(11)计算每一个粒子的适应度值  $Fitness(C, \sigma^2, \epsilon)$ , 然后根据粒子的适应度值更新个体极值  $p_i$  和全局极值  $pg_{best}$ .

Step3: 由公式(8)、(9)和(10), 更新每个粒子的速度和位置.

Step4: 检查算法终止条件, 如迭代次数等于  $t_{max}$  或最优解不再发生变化, 则算法结束, 输出最优的参数组合. 否则返回 Step3 继续寻优.

### 2.3 PSO-WSVR 水质预测模型构建

PSO-WSVR 水质预测模型的基本思想就是充分利用其强大的处理非线性多参数系统能力, 建立集约化河蟹养殖水域中水质变化与各个影响因素之间复杂的非线性关系, 挖掘水质内部变化规律, 从而实现利用历史影响因素的变化对未来水质变化做出精确预测, 有效提高了预测模型的精度以及泛化能力. 其水质预测模型构建步骤如图 1 所示.

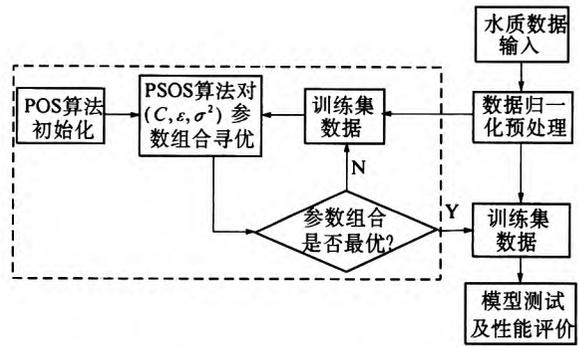


图 1 PSO-WSVR 水质预测模型构建步骤

Fig. 1 Implementation process of predict the water quality based on PSO-WSVR

## 3 实验结果与分析

### 3.1 研究对象及数据源

采用江苏省水产养殖智能控制技术标准化示范区——宜兴市河蟹养殖池塘生态环境在线监测的数据作为样本数据, 每个样本包括溶解氧、水温、PH 值、空气湿度、风速、太阳辐射等 6 项指标. 采样周期为 2010 年 7 月 18 至 7 月 20 日, 每 30 min 采样一次, 共计 144 个样本, 抽取前 2 d 的 96 个样本为训练集, 剩余的 48 个样本作为测试集, 以对池塘溶解氧浓度进行定量预测. 其监测的生态环境原始数据变化曲线图如图 2 所示.

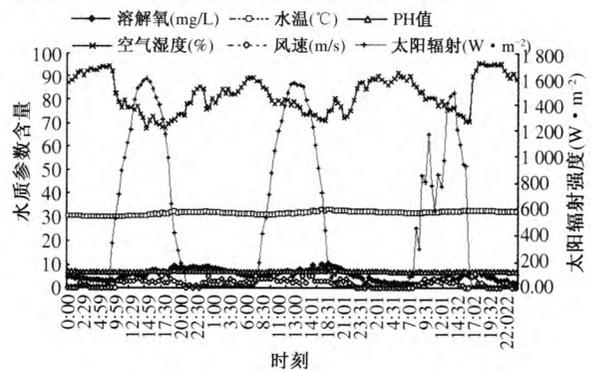


图 2 原始数据变化曲线图

Fig. 2 The variation curve of the original water quality data

### 3.2 数据归一化预处理

在实际应用中, 为消除由于样本数据精度和量纲不同对预测模型性能的影响, 为此有必要对样本数据进行归一化预处理. 采用公式(12)将图 2 所示的河蟹养殖生态环境数据分别归一化到  $[0, 1]$  之间, 既保留了数据的原始特征又提高算法性能.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (12)$$

式中:  $x_{min}$ 、 $x_{max}$  分别为原始数据的最小值和最大值;  $x$  和  $x'$  为归一化前后的数据.

3.2 算法实现及性能对比分析

算法利用 Matlab7.13 语言编程实现, PSO 算法初始化为:种群规模为  $n = 50$ ,  $c_1 = c_2 = 1.49$ , 最大迭代次数  $t_{max} = 100$ ,  $\omega$  的范围设置为  $[0.9, 0.35]$ , 粒子速度  $v$  的限定范围  $[0.3, 8]$ . 将前半小时的溶解氧、水温、PH 值、空气湿度、风速、太阳辐射作为输入项,下一时刻的溶解氧预测值作为输出项,预测值与下一时刻的溶解氧实际值的均方误差作为粒子适应度函数.按照基于改进 PSO 的 WSVR 参数组合优化算法步骤对 PSO - WSVR 进行训练,在训练次数  $t = 80$  时,获得 WSVR 的最佳参数组合: $\sigma = 0.0075$ ,  $C = 78.36$ ,  $\varepsilon = 1.6859$ .将组合参数带入 PSO - SVRM 模型中对溶解氧浓度进行预测.

为了检验 PSO - WSVR 预测模型的性能,采用均方根误差 RMSE、均方差 MSE、平均绝对误差 MAE、平均绝对百分比误差 MAPE、相关系数  $R^2$  分别作为算法性能的评价指标.将本文算法、标准 SVR 算法和三层 S 型激励函数的 BP 神经网络进行对比分析,其拟合预测曲线对比图如图 3 所示,预测结果误差统计见表 1.

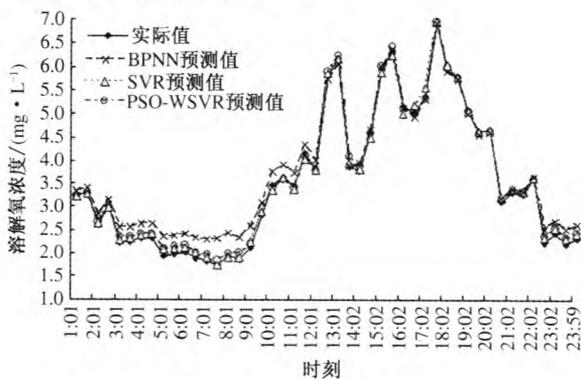


图3 PSO - WSVR、标准的 SVR 和 BPNN 算法预测结果比较

Fig.3 Comparison of results forecasted by BPNN, standard SVR and PSO - WSVR

表1 BPNN、标准 SVR 和 PSO - WSVR 的预测结果误差比较

Tab.1 The forecasting error result comparison of BPNN, standard SVR and PSO-WSVR

方法	RMSE	MAE	MAPE	MSE	$R^2$
BPNN	0.102 2	0.032 0	0.013 3	0.010 50	0.894 5
标准 SVR	0.065 9	0.011 2	0.009 4	0.008 21	0.948 8
PSO-WSVR	0.035 1	0.009 2	0.005 3	0.002 64	0.997 5

由图 3 可知,基于 PSO - WSVR 的水质预测算法拟合度较好,预测结果比较符合实际情况;标准 SVR 预测曲线拟合效果次之;而 BP 神经网络预测曲线拟合的相对较差,这与 BP 神经网络

太适合小样本数据预测有关.

由表 1 不难看出, PSO - WSVR 模型与 BP 神经网络相比,性能评价指标 RMSE、MAE、MAPE、MSE 分别下降了 65.66%、71.25%、60.15%、74.86%,  $R^2$  上升了 11.51%. 这表明,在小样本的情况下 PSO-WSVR 比标准 BP 神经网络具有更好的性能和预测精度. PSO - WSVR 模型与标准 SVR 相比,性能评价指标 RMSE、MAE、MAPE、MSE 分别下降了 46.74%、17.86%、43.62%、67.84%,  $R^2$  上升了 5.13%. 试验表明, PSO - WSVR 参数选择避免了交叉验证试算确定 WSVR 参数组合对用户经验的依赖,采用 PSO 算法获得参数组合有效提高预测模型的精度和性能.总的来说,笔者提出的 PSO-WSVR 模型不仅预测精度高、泛化性能好.

4 结论

(1) 针对集约化水产养殖领域水质预测中不同时刻的水质数据对预测模型的贡献程度不同,提出了加权支持向量回归机克服了标准支持向量机预测的不足,较好的体现了近期数据较远期数据对未来预测更具有影响力,应赋予更高的权重的基本思想,从而可以提高预测精度和可靠性,能够满足河蟹养殖水质管理的需要.

(2) 基于 PSO 优化算法对 WSVR 的参数组合进行寻优,克服了交叉验证试算确定 WSVR 参数组合的人为主观因素的影响,并将 PSO-WSVR 混合预测模型应用于江苏省水产养殖智能控制技术标准化示范区——宜兴市集约化河蟹养殖池塘水质变化趋势中.结果表明,基于 PSO-WSVR 预测模型较标准 SVR 和 BP 神经网络预测模型具有更好的预测精度和泛化能力.

(3) 通过对集约化水产养殖池塘的水质监测,提前预知未来短期池塘水质变化趋势,给养殖户和水利管理部门不仅提供充足水质污染尤其是突发水质恶化事件的应对时间,还对有效降低人力劳动强度,提高生产效益,有一定的应用推广价值.

参考文献:

[1] 申安华. 云南高原湖泊与池塘的河蟹养殖技术研究 [J]. 水生态学杂志, 2010, 3(2): 129 - 132.

[2] DELLANA S, WEST D. Predictive modeling for wastewater applications: Linear and nonlinear approaches [J]. Environmental Modelling and Software. 2009, 24: 96 - 106.

[3] ABAURREA J, ASIN J, CEBRIAN A C, et al. Trend

- analysis of water quality series based on regression models with correlated errors [J]. *Journal of Hydrology*, 2011, 400: 341 – 352.
- [4] 王小艺,赵晓平,刘载文,等. 基于灰色理论的湖库水体富营养化预测方法研究[J]. *计算机仿真*, 2011, 28(1): 17 – 19.
- [5] TAN Guo-hua, YAN Jian-zhou, GAO Chen, et al. Prediction of water quality time series data based on least squares support vector machine[J]. *Procedia Engineering*, 2012, 31: 1194 – 1199.
- [6] MAHAPATRA S S, NANDA S K, PANIGRAH B K. A cascaded fuzzy inference system for indian river water quality prediction [J]. *Advances in Engineering Software*, 2011, 42: 787 – 796.
- [7] EVNARD J, GRIEU S, POLIT M. Wavelet-based multi-resolution analysis and artificial neural networks for forecasting temperature and thermal power consumption [J]. *Engineering Applications of Artificial Intelligence*, 2011, 24(3): 501 – 516.
- [8] VAPNIK V. *Statistical learning theory* [M]. New York: Wiley, 1998.
- [9] SINGH K P, BASANT N, GUPTA S. Support vector machines in water quality management [J]. *Analytica Chimica Acta*, 2011, 703: 152 – 162.
- [10] 王雪松,程玉虎,郝名林. 一种支持向量机参数选择的改进分布估计算法[J]. *山东大学学报:工学版*, 2009, 39(3): 7 – 10.
- [11] AVCI E. Selecting of the optimal feature subset and kernel parameters in digital modulation classification by using hybrid genetic algorithm-support vector machines: HGASVM [J]. *Expert Systems with Applications*, 2009, 36(2): 1391 – 1402.
- [12] 张俊,程春田,申建建,等. 基于蚁群算法的支持向量机中长期水文预报模型 [J]. *水力发电学报*, 2010, 29(6): 34 – 40.
- [13] HSIEH T J, HSIAO H F, YEH W C. Mining financial distress trend data using penalty guided support vector machines based on hybrid of particle swarm optimization and artificial bee colony algorithm [J]. *Neurocomputing*, 2012, 82: 196 – 206.
- [14] BABAOGU I, FINDIK O, Ü LKER E. A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine [J]. *Expert Systems with Applications*, 2010, 37(4): 3177 – 3183.
- [15] 陈懿冰,张玲玲,聂广礼,等. 基于改进的支持向量回归机的金融时序预测 [J]. *数学的实践与认识*, 2012, 42(4): 38 – 44.
- [16] KENNEDY J, EBERHART R C, SHI Y. *Swarm Intelligence* [M]. San Francisco, CA: Morgan Kaufmann Publishers Inc, 2001.

## Study of Short-term Water Quality Prediction Model Based on PSO-WSVR

XU Long-qin<sup>1</sup>, LIU Shuang-yin<sup>1,2,3</sup>

(1. School of Information, Guangdong Ocean University, Zhanjiang 524025, China; 2. Key Laboratory of Modern Precision Agriculture System Integration Research of Ministry of Education, China Agricultural University, Beijing 100083, China; 3. Beijing Engineering Research Center for Agricultural Internet of Things, China Agricultural University, Beijing 100083, China)

**Abstract:** In view of the difficulty in establishing precise nonlinear water quality forecast model using the traditional method, the paper proposed the short-term forecast model of the water quality using the weighted support vector regression machine based on the particle swarm optimization (PSO-WSVR). According to the importance of the samples are significantly different, the authors proposed the penalty coefficient for the every sample to different weighted values and improved the standard support vector regression algorithm to avoid different samples using the same weight, which may cause the low prediction accuracy. The parameters combinations of the weighted support vector regression machine were adaptively optimized using particle swarm optimization algorithm, which the convergence rate could be sped up significantly. The water quality of crab intensive aquaculture in Yixing, Jiangsu was predicted using PSO-WSVR model. Compared with forecasting result of the standard support vector regression and BP neural network, the forecasting result of PSO-WSVR have reliable performance, generalization ability, and high forecast precision, so it provides a new way for short-term forecasting of the water quality in intensive aquaculture.

**Key words:** water quality prediction; weighted support vector regression; particle swarm optimization; parameters optimization