基于时间相似数据的支持向量机水质溶解氧在线预测

刘双印[1,2,3]，徐龙琴[1]，李道亮[2,3,4※]，曾立华[3,4,5]

（1. 广东海洋大学信息学院，湛江 524025； 2. 农业部农业信息获取技术重点实验室，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； 4.中国农业大学 先进农业传感技术北京市工程研究中心，北京 100083； 5. 河北农业大学机电工程学院，保定 071001）

摘 要：为及时辨识集约化水产养殖水质变化趋势、动态调控水质，确保无应激环境下健康养殖，该文提出了基于时序列相似数据的最小二乘支持向量回归机（least squares support vector regression，LSSVR）水质溶解氧在线预测模型。采用特征点分段时间弯曲距离（feature points segmented time warping distance，FPSTWD）算法对在线采集的时间序列数据进行分段与相似度计算，以缩减规模的子序列数据集对 LSSVR 模型进行快速训练优化，实现了多个 LSSVR 子模型在线建模，将预测数据序列与 LSSVR 子模型的相似度匹配，自适应地选取最佳的子模型作为在线预测模型。应用该模型对集约化河蟹福利养殖水质参数溶解氧浓度进行在线预测，模型评价指标中最大相对误差、平均绝对百分比误差、相对均方根误差和运行时间分别为 4.76%、8.18%、5.23%、8.32 s。研究结果表明，与其他预测方法相比，该模型具有较好的综合预测性能，能够满足河蟹福利养殖水质在线预测的实际需求，并为集约化水产养殖水质精准调控提供研究基础。

关键词：水产养殖；水质；模型；支持向量机；在线预测；特征点分段时间弯曲距离；相似数据

doi：10.3969/j.issn.1002-6819.2014.03.021

中图分类号：TP391 **文献标志码**：A **文章编号**：1002-6819(2014)-03-0155-08

刘双印，徐龙琴，李道亮，等. 基于时间相似数据的支持向量机水质溶解氧在线预测[J]. 农业工程学报，2014，30(3)：155－162.

Liu Shuangyin, Xu Longqin, Li Daoliang, et al. Online prediction for dissolved oxygen of water quality based on support vector machine with time series similar data[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2014, 30(3): 155－162. (in Chinese with English abstract)

0 引 言

集约化水产养殖水质在线预测为及时辨识水质变化趋势、水质动态调控和应对突发水质事件等方面实现科学决策和养殖精细化管理的重要依据，从而确保水产品在无应激条件下健康生长，对其研究具有重要的研究价值和现实意义[1]。

近年来，已提出了各种在线预测模型和方法，如滑动时间窗方法[2]、动态神经网络方法[3-4]和基于增量训练的在线支持向量机（online support vector regression，Online SVR）算法[5-7]等。滑动时间窗方法采用时间窗滚动或滑动的方式进行在线预测，但模型自身不具备随时间序列在线更新和动态学习的能力，致使预测精度不理想。动态神经网络方法通过改进网络结构和调整网络参数，实现模型的动态更新与在线预测，但存在计算复杂度很高，不适于在线更新较快的时间序列预测等缺陷。基于增量训练的 Online SVR 是目前应用最广泛的在线时间序列预测模型，但由于在线训练过程中所有的样本数据都参与增量计算与迭代优化，涉及大规模矩阵的求逆，造成算法计算复杂度较高，执行效率低，难以满足在线预测的需要。而最小二乘支持向量回归机（least squares support vector regression，LSSVR）具有计算效率高、泛化性能强等优点，但若直接用于在线建模，其所需存储空间和计算量会随着时间序列获取或在线采集的样本数增加而增大，易产生数据过饱和、泛化能力差，甚至模型性能失效等问题[8-9]。为此，一些国内外学者以 LSSVR 为基础，从在线获取样本点组成训练样本集的方式上和减少计算复杂度等方面入手，采用剪枝算法、增减式学习、滑动窗和加权等改进策略分别提出了性能各异的 LSSVR 在线学习算法，并取得了较好的预测效果[8-11]。如：张浩然等[8]根据分块矩阵和

收稿日期：2013-07-17 修订日期：2013-12-06

基金项目：国家科技支撑计划项目（2011BAD21B01）；国家基金项目（61100115）；广东省省部产学研结合项目（2012B090500008）；广东省科技计划项目（2012A020200008，2012B091100431）；广东省自然基金项目（S2013010014629，S2012010008261）

作者简介：刘双印（1977－），男，山东单县人，副教授，博士生，CCF 会员，主要从事智能计算、智能信息系统、农业信息化技术等研究。湛江 广东海洋大学信息学院，524025。Email：hdlsyxlq@126.com

※通信作者：李道亮（1971－），男，山东垦利人，教授，博士，博士生导师，主要从事农业先进传感与智能信息处理研究。北京 中国农业大学信息与电气工程学院，100083。Email：dliangl@cau.edu.cn

核函数矩阵公式特点，提出了增量式 LSSVR 算法（incremental least square support vector machine，ILSSVR），仿真试验验证了算法的有效性。周欣然等提出了稀疏在线无偏置最小二乘支持向量回归机（sparse online non-bias least square support vector machine，SONB-LSSVR）算法[12]，采用递推地学习新样本并删除贡献最小样本有效提高样本集的多样性和代表性，在液位预测控制领域得到了成功的应用。虽然上述算法从不同的角度对在线最小二乘支持向量机性能进行改进，但在算法实施过程中，随着时间序列推进和在线获取使训练数据集不断更新，增、减量训练算法涉及大规模矩阵的求逆过程，常用的分块法也失去优势，导致算法效率明显降低。而采用以欧氏距离进行样本间相似度计算为基础的剪枝算法、加权方法，以及滑动窗口方法通过压缩训练样本的规模来提高 OLSSVR 在线建模的计算效率，但欧氏距离不具备在时间轴上对时间序列数据形状扭曲变形进行辨识的能力，使训练数据信息局限于在线数据邻域，遗忘了大量的历史信息，导致模型预测精度和泛化能力下降。

以确保在线时间序列预测算法执行效率又兼顾预测精度为目的，结合集约化水产养殖水质序列的相似性和连续变换的规律，在前人研究的基础上，以“特征相似输入产生相似输出”为原则，本文提出了基于时序相似数据的 LSSVR 水质溶解氧在线预测模型，给出了模型的推导过程，并以集约化水产养殖水质时间序列数据为例进行建模和在线预测试验验证。与其他模型相比，结果表明该在线预测模型综合性能较好，具有一定的应用前景。

1 研究区域与数据源

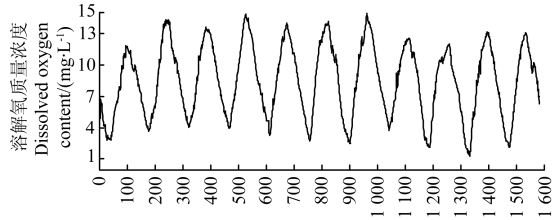
1.1 研究区域

江苏省宜兴市在江苏省南部（31°07′～31°37′N， 119°31′～120°03′E），属亚热带季风气候，全年温暖湿润，比较适合水产养殖。全市共有水产养殖面积 16 000 hm^2，其中河蟹养殖面积近 9 333 hm^2，河蟹养殖是广大养殖户及相关企业的重要经济来源。但河蟹养殖模式仍主要为靠天靠经验的传统养殖模式，超容量、饵料投喂和实施渔药方法不科学、不合理，易造成养殖水质污染，疾病时常爆发，不但给养殖户带来巨额损失，对当地生态环境安全构成严重威胁，还制约了河蟹养殖产业健康可持续发展。

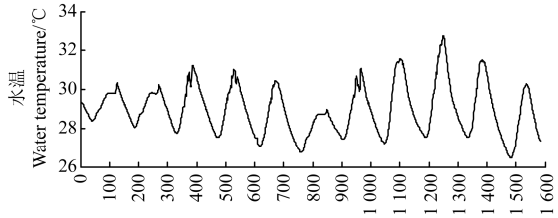
1.2 研究对象与数据源

本研究以江苏省宜兴市河蟹养殖某池塘为研究对象，采用中国农业大学研制的基于水产养殖物联网的集约化水产养殖在线监控系统获取河蟹福利养殖生态环境数据作为数据源[1]，其水产养殖监控管理系统网址 http://sc.agriot.net/caiotAqu/login_val.action。

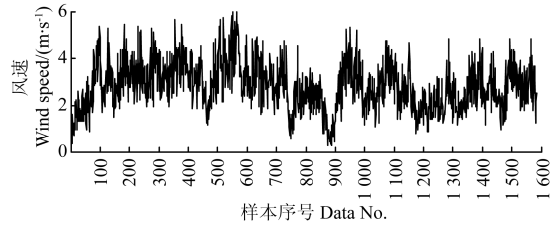
在线监控的样本特征属性由溶解氧、水温、太阳辐射、气压、风速、风向、空气湿度等 7 个指标组成。采样周期为 2012 年 7 月 21 日至 7 月 31 日，每 10 min 采样 1 次，共计 1 584 个样本，从中抽取前 9 天的 1 296 个样本为训练集，剩余的 288 个样本作为测试集，对河蟹福利养殖重要的水质参数溶解氧进行在线预测。其监测的生态环境原始数据变化曲线图如图 1 所示。



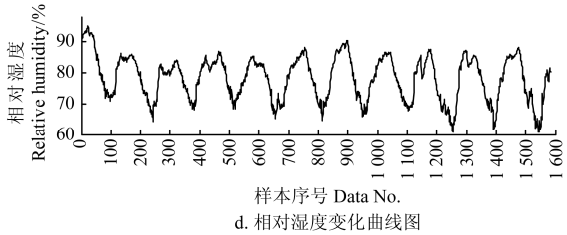
a. 溶解氧变化曲线图
a. Variation curves of dissolved oxygen



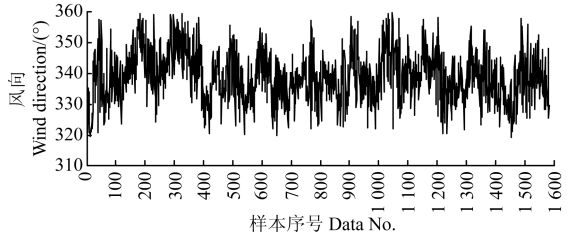
b. 水温变化曲线图
b. Variation cures of water temperature



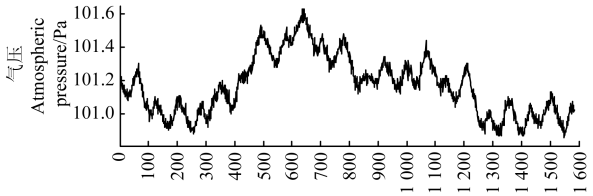
c. 风速变化曲线图
c. Variation curves of wind speed



d. 相对湿度变化曲线图
d. Variation curves of relative humidity

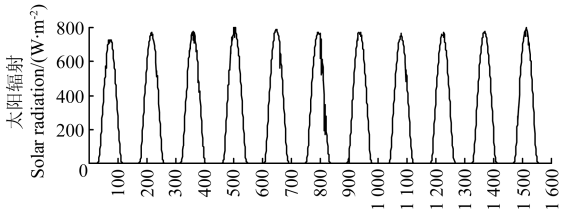


e.风向变化曲线图
e. Variation curves of wind direction



f. 气压变化曲线图

f. Variation curves of atmospheric pressure



g. 太阳辐射变化曲线图

g. Variation curves of solar radiation

图 1　原始水质数据变化曲线图

（2012 年 7 月 21 日－2012 年 7 月 31 日）

Fig.1 Variation curve of original water quality data

(2012-07-21－2012-07-31)

2　在线预测模型构建

2.1　最小二乘支持向量回归机

对于非线性时间序列样本数据(x_1, y_1)，(x_2, y_2)，…，(x_i, y_i)，…，(x_n, y_n)，$x_i \in R^n$ 和 $y_i \in R$，采用最小二乘支持向量回归机进行函数估计，优化问题则变成：

$$\min J(\omega,\xi)=\frac{1}{2}\omega^T\omega+\frac{C}{2}\sum_{i=1}^{n}\xi^T\xi \qquad (1)$$

$$\text{s.t.}\ y_i=\omega^{\mathrm{T}}\varphi(x_i)+b+\xi_i \quad (i=1, 2, \ldots, n) \qquad (2)$$

式中：J 为损失函数，ω为权重向量，T 为向量转置符号，$\xi_i \in R$ 为经验误差，b 为偏置量，$C \in R^+$是正则化参数，$\varphi(\cdot)$为输入空间到特征空间的非线性映射[1]。为求解上述约束优化问题[1]，其对偶问题的 Lagrange 函数为：

$$L(\omega,b,\xi,\alpha)= J(\omega,\xi)-\sum_{i=1}^{n}\alpha_i(\omega^T\phi(x_i)+b+\xi_i-y_i) \qquad (3)$$

式中：α_i 为拉格朗日乘子。由 Karush-Kuhn-Tucher（KKT）条件，分别对ξ_i、ω、b 和α_i 参数求偏导数并令其分别等于 0[13-14]，则有：

$$\begin{cases}\dfrac{\partial L}{\partial \xi_i}=0\rightarrow \alpha_i=C\xi_i,\ i=1,...,n\\ \dfrac{\partial L}{\partial \omega}=0\rightarrow \omega=\sum_{i=1}^{n}\alpha_i\varphi(x_i)\\ \dfrac{\partial L}{\partial b}=0\rightarrow \sum_{i=1}^{n}\alpha_i=0\\ \dfrac{\partial L}{\partial \alpha_i}=0\rightarrow \omega^T\varphi(x_i)+b+\xi_i-y_i=0,\ i=1,...,n\end{cases} \qquad (4)$$

从上式消去ω，ξ_i后可得到如下线性方程组[15]：

$$\begin{bmatrix}0 & e^T\\ e & \Gamma+\psi^{-1}I\end{bmatrix}\cdot\begin{bmatrix}b\\ \alpha\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix} \qquad (5)$$

式中：$e=[1, 1, \ldots, 1]^T$，I 是 $n\times n$ 维的单位矩阵；$y=[y_1, y_2, \ldots, y_n]^T$，$\Gamma_{ij}=\varphi(x_i)\cdot\varphi(x_j)=K(x_i, x_j)$，$a=[a_1, a_2, \ldots, a_n]$，求解式（5）得到最小二乘支持向量回归机决策模型为：

$$\hat{y}=f(x,\alpha)=\sum_{i=1}^{n}\alpha_i k(x,x_i)+b \qquad (6)$$

从式（5）可以看出，LSSVR 的训练问题归结为求解线性方程组的问题，具有计算相对简单，快速的特点；但 LSSVR 的解丧失稀疏性，缺乏遗忘机制，随着时间序列样本不断增加，需要保持很多样本参与训练，致使矩阵维数剧增，严重制约着在线学习的效率，甚至因维数灾难导致训练失败。所以如何处理在线新增样本点，简化学习算法，提高求解速度，是 LSSVR 算法在线预测的关键。

2.2　时间序列相似度计算

鉴于特征点分段时间弯曲距离（feature points segmented time warping distance，FPSTWD）具有能提供一种全局趋势信息[15-16]，缩减经典时间弯曲距离计算数据维数[17]，时间序列相似度高、计算复杂度低等特点[18-19]。本文采用 FPSTWD 方法对时间序列数据进行相似度计算，即运用 FPSTWD 对历史数据库或在线获取的数据序列进行特征点分段，构建多个分段子序列簇（cluster）或分段子序列集合，以特征点分段时间弯曲距离作为相似测度，使同一个簇内的对象之间具有较高的相似度，而不同的簇中的对象差别比较大。而准确定义并获取特征点是基于 FPSTWD 的时间序列数据相似度计算执行过程中时间序列分段及时间序列数据相似度计算中的重要环节。

定义 1：时间序列 x 的特征点：给定阈值 Ψ 和时间序列$\{x_1=(a_1, \cdots, a_N)\}$，如果 x_i 是一个特征点（$1\leqslant i\leqslant N$），它必须满足 2 个条件：①x_i 必须是时间序列的极值点或拐点，其中序列的起点与终点均默认为特征点；②若 $x_i>x_{i-1}$，则 $x_i/x_{i-1}>\Psi$ 必须成立，否则，$x_i<x_{i-1}$，则 $x_{i-1}/x_i>\Psi$ 必须成立[15]。

阈值 Ψ 是极值点的影响因子取值的最小范围，取值与具体应用领域知识、序列长度及用户关注角度有关，一般情况下 $\Psi\in[0.01, 0.1]$。在得到时间序列的特征点后，对相邻的特征点间的点集进行直线拟合，即可得到时间序列的分段线性表示。

定义 2：假定时间序列 x 与 y 经线段化后分别为 x^S 与 y^S，其中 $x^S=<a_1, a_2, \ldots, a_m>$，$y^S=<b_1, b_2, \ldots, b_n>$，$m$ 和 n 分别是 x^S，y^S 的长度，构造 $m\times n$ 的时

间归整矩阵 $d=(d(i, j))_{m\times n}$，元素 $d(i, j)$表征序列点对(x_i, y_i)间的距离。则 x^S 与 y^S 之间的 FPSTWD 距离定义如下[16]：

$$D(x^S,y^S)=\begin{cases}0, & \text{if } x^S=y^S\\ \infty, & \text{if } x^S=0 \text{ or } y^S=0\\ D_{base}(x_1^S,y_1^S)+\min\{D(x^S,rest(y^S)),\\ D(rest(x^S),y^S),D(rest(x^S),rest(y^S))\}, & \text{otherwise}\end{cases} \quad (7)$$

式中，$D_{base}(x_1^S,y_1^S)=(x_1^S[1]-y_1^S[1])^2+(x_1^S[m]-y_1^S[n])^2$，$rest(x^S)=<a_1, a_2, \ldots, a_{m-1}>$，$rest(y^S)=<b_1, b_2, \ldots, b_{n-1}>$。利用递归公式（7）以行或列的顺序填充矩阵 $\boldsymbol{d}$，最后矩阵 $\boldsymbol{d}$ 中(m, n)元素中的值即为两序列的 FPSTWD 值。与经典 TWD 距离的时间复杂度 $O(|x||y|)$相比，FPSTWD 的时间复杂度 $O(\min(m, n))$缩减若干常数倍，有效提高在线计算效率。

根据时间序列数据分段子序列的长度不同，其相似度匹配搜索主要分为 2 类[16]：

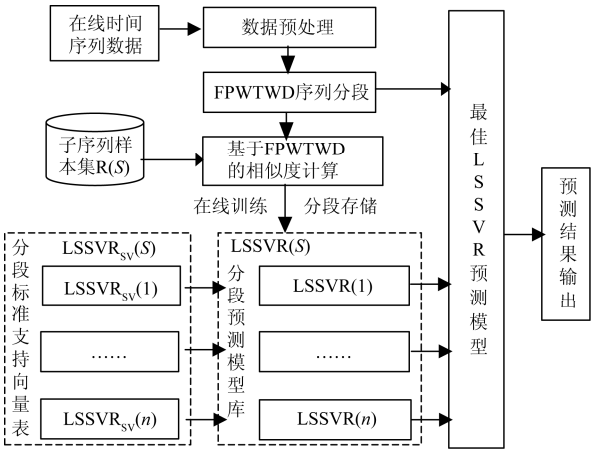
①全序列匹配，给定搜索序列 x_1，需要在相同长度的分段子序列簇中找到与 x_1 最相似的序列或者与 x_1 的特征点分段时间弯曲距离小于某个阈值 μ 的所有序列。

②子序列匹配，给定搜索序列 x_1 小于分段子序列簇中序列的长度时，需要在序列簇中搜索到子序列片段与 x_1 最相似或与 x_1 的特征点分段时间弯曲距离小于某个阈值 μ。

2.3 基于 FPSTWD 和 LSSVR 的在线预测模型

2.3.1 在线预测设计思路

为解决现有 LSSVR 在线预测存在的问题，并兼顾预测精度和计算复杂度为指导思想，以“特征相似输入产生相似输出”为原则，即把同预测时间段具有相似特征的生态环境数据构建的参考预测模型作为桥梁，计算预测时间段与其相似历史时间段的生态环境数据特征的相似度，选择适宜的参考预测模型进行拟合预测，实现对研究对象的准确预测[20-23]。为此，本文将特征点分段时间弯曲距离（FPSTWD）与最小二乘支持向量回归相结合构建在线预测模型。建模过程中采用 FPSTWD 算法对样本序列数据进行分段与相似度计算，组成特征相似的子序列集合，然后应用特征子序列样本集对 LSSVR 进行在线训练优化，构建相应的分段 LSSVR 子模型，并获得相应的支持向量，实现了多个 LSSVR 子模型在线建模；对新增样本序列，运用 FPSTWD 方法对新增样本序列进行分段和相似度计算，获得与新增样本最相似的子序列样本所对应的分段预测模型，将新增样本子序列输入到该模型进行在线预测，提高了预测模型随时间序列数据变化的自适应能力。其模型设计原理如图 2 所示。



注：$LSSVR_{SV}(S)$为子分段的支持向量、LSSVR(*S*)为子分段模型。
Note: $LSSVR_{SV}(S)$ is the support vector of subsegments, LSSVR(*S*) is the model of subsegments.

图 2 基于特征点分段时间弯曲距离的最小二乘支持向量机在线预测模型原理

Fig.2 Schematics of online prediction with least squares support vector regression based on feature points segmented time warping distance

2.3.2 基于 FPSTWD 的 LSSVR 在线预测模型构建

以时间序列水质预报为例，基于 FPSTWD 的 LSSVR 在线预测模型描述如下：

步骤 1：FPSTWD-LSSVR 模型参数初始化

需设置的参数有惩罚参数 C、核函数参数σ^2、精度阈值ϑ、阈值 Ψ 和相似度阈值 μ；LSSVR 核函数类型，子分段模型 LSSVR(*S*)，S=1，2，3，…，训练集初始长度 TL。

步骤 2：历史数据训练样本集与当前第 q 批次时间窗训练样本点的表示

因预测模型的需要，在构造训练样本集时应将输入输出样本错位结合，那么第 p 个历史批次的训练样本集可用 $U_p=\{(X_1, Y_1), \ldots, (X_p, Y_p)\}$来表示，其中训练样本集中任一个样本点表示为$(x_i, y_{i+1})$，$l<i<n-1$。若当前第 q 批次时间窗内数据训练样本点用(x_q, y_{q+1})表示，那么就得到第 q 时刻窗长为 l 的训练样本集 $U_q=\{X_{\text{now}}(q), Y_{\text{now}}(q)\}$，其中 $X_{\text{now}}(q)=[x_{q-l}, x_{q-l+1}, \ldots, x_q]$，$Y_{\text{now}}(q)=[y_{q-l+1}, y_{q-l+2}, \ldots, y_{q+1}]$，$l+1\leqslant q\leqslant n$，$y_q\in R$，$x_q\in R^n$。

步骤 3：基于 FPSTWD 相似度计算的子分段预测模型库与标准支持向量表的构建

采用 2.2 节基于 FPSTWD 的时间序列数据相似度计算方法对历史样本集 U_S 中的时间序列数据进行分段处理，组成特征相似的子序列样本集 $R(S)$，（S=1，2，…）。应用 $R(S)$对 LSSVR 进行训练优化，对每个子序列样本集 $R(S)$构建相应的子分段预测模型 LSSVR(*S*)，获得子分段的支持向量 $LSSVR_{SV}(S)$，

以分段方式将 LSSVR(S)模型和 $\text{LSSVR}_{\text{SV}}(S)$分别保存分段预测模型库及相应的支持向量样本表 $\text{Tb}_{\text{SV}}(S)$。

步骤 4：新增时间序列数据 U'分段与相似度计算

1）搜索时间序列数据 U'的特征点，由特征点对 U'进行分段处理；采用公式（7）计算 U'子序列与所有子序列样本集 $R(S)$的特征点分段时间弯曲距离，FPSTWD 距离越小相似度越大，找出与 U'子序列相似度最大的 $R(j)$，$j\in[1，S]$。

2）在 $R(S)$中若找不到与 U'子序列相匹配的子序列，或者 FPSTWD 距离大于指定相似度阈值 μ，将 U'作为新的子序列样本集或做删除处理，并记下该时间序列 U'对应的时刻以及相应的状态信息。

步骤 5：在线预测

1）把 $R(j)$所对应的子分段预测模型 LSSVR(j)，$j\in[1，S]$，作为最佳的 LSSVR 预测模型，将数据 U'子序列输入到 LSSVR(j)模型进行在线预测，输出预测结果；

2）若新增数据序列 U'的子序列因奇异值或噪声找不到相匹配的子序列 $R(j)$，此时采用与 U'的子序列相似日期相似时刻的历史数据替代之进行预测，并输出预测结果；

3）若预测精度小于指定的精度阈值 ϑ，则将新增样本序列 U'与所有支持向量 $\text{LSSVR}_{\text{SV}}(S)$一起训练 LSSVR，并将符合预测精度要求的预测模型和支持向量保存下来，这样不断完善分段预测模型库和各分段的支持向量。

步骤 6：子分段预测模型库与标准支持向量表更新策略[24]

随着时间序列数据预测的不断推进，子分段模型数量与标准支持向量也与之剧增，为提高存储效率和分段模型预测的性能，采取更新策略为：1）丢弃应用次数较少的子分段模型及相应的支持向量；2）根据 FPSTWD 子分段序列相似度计算删减或合并相似度较高的 LSSVR(S)子模型，与其相对应的支持向量也随之删除或合并，有效节约系统存储空间。

步骤 7：时间序列数据在线更新，重复执行步骤 4～6。

3　仿真结果与分析

3.1　算法实现与测试

本文以径向基函数（radial basis function，RBF）为 LSSVR 模型的核函数，参照 LIBSVM 算法，采用 Visual C#.NET 语言对 FPSTWD-LSSVR 在线算法进行编程。运用 FPSTWD 方法对训练样本序列进行分段与相似度计算，最终形成 10 个子序列集合 S_1，S_2，S_3，S_4，S_5，S_6，S_7，S_8，S_9，S_{10}，每个子集的样本个数分别为 128、166、174、158、138、181、116、202、169、152，以这 10 个子序列集合样本分别对 LSSVR 离线训练，得到对应 10 个 LSSVR(S)预测子模型和支持向量 $\text{LSSVR}_{\text{SV}}(S)$，$S\in[1，\ldots，10]$。采用 FPSTWD-LSSVR 算法对 2012 年 7 月 30 日至 7 月 31 日 24h 池塘的 288 个时间序列数据进行单步预测，其预测结果如图 3 所示，预测误差曲线如图 4 所示。

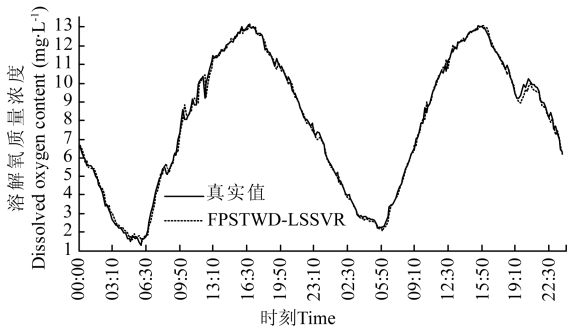


图 3　FPSTWD-LSSVR 预测值和真实值曲线
（2012 年 7 月 30 日－2012 年 7 月 31 日）
Fig.3　Curves of forecast value and actual value for FPSTWD-LSSVR (2012-07-30－2012-07-31)

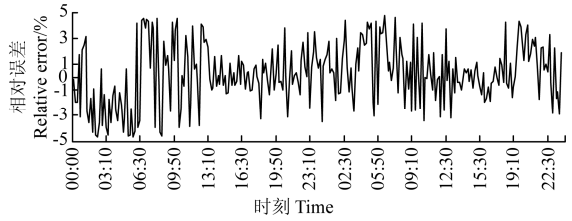


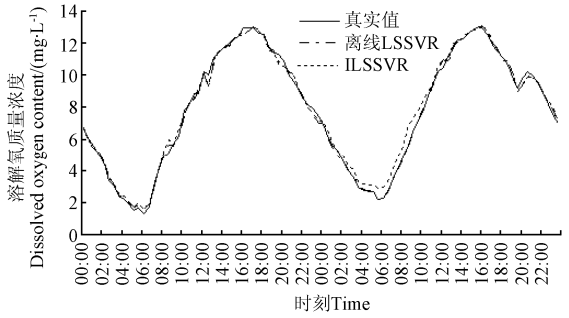
图 4　FPSTWD-LSSVR 预测误差曲线
（2012 年 7 月 30 日－2012 年 7 月 31 日）
Fig.4　Forecast error curves of FPSTWD-LSSVR (2012-07-30－2012-07-31)

由图 3 知，本文提出的模型预测曲线能够与养殖池塘溶解氧真实值曲线拟合较好。从图 4 可以看出，该算法输出结果在曲线下拐点处与真实值误差较大，最大相对误差为 4.76%，但能够满足集约化河蟹福利养殖池塘溶解氧在线预测的需要。

3.2　结果对比分析

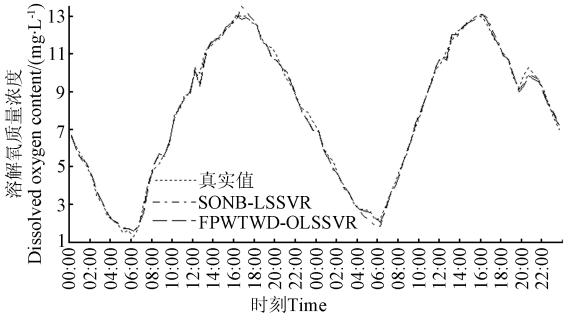
为了验证 FPSTWD-LSSVR 预测模型性能，以 Matlab 为试验环境，选择增量式 LSSVR 算法（incremental least square support vector machine，ILSSVR）[8]，稀疏在线无偏置最小二乘支持向量回归机（sparse online non-bias least square support vector machine，SONB-LSSVR）算法[12]和离线 LSSVR 算法进行对比分析；分别采用最大相对误差（$\text{error}_{\max}$，$E_{\max}$）、平均绝对百分比误差（mean absolute percentage error，MAPE）、相对均方根误

差（relative root mean square error，RRMSE）和运行时间 t 作为上述算法性能评价指标。以生态环境数据和模型参数都相同的条件下，综合比较各种算法的预测精度与效率，其预测结果对比图见图 5，其性能评价结果统计见表 1。



a. 离线 LSSVR 和 ILSSVR 预测方法的溶解氧预测结果对比图（2012 年 7 月 30 日－2012 年 7 月 31）

a. Prediction comparison of dissolved oxygen with off-line LSSVR and ILSSVR prediction methods (2012-07-30－2012-07-31)



b. SONB-LSSVR 和 FPWTWD-OLSSVR 预测方法的溶解氧预测结果对比图（2012 年 7 月 30 日－2012 年 7 月 31）

b. Prediction comparison of dissolved oxygen with SONB-LSSVR and FPWTWD-OLSSVR prediction methods (2012-07-30－2012-07-31)

图 5 不同预测方法的溶解氧预测结果对比图

Fig.5 Prediction comparison of dissolved oxygen with different prediction methods

表 1 各模型预测结果对比

Table 1 Comparison of various predicted results

方法 Method	E_{max}/%	MAPE/%	RRMSE/%	T/s
离线 LSSVR	4.33	6.86	4.79	19.59
ILSSVR	8.42	15.71	7.57	13.48
SONB-LSSVR	7.15	13.63	6.74	11.06
FPSTWD-LSSVR	4.76	8.18	5.23	8.32

注：E_{max} 为最大相对误差；MAPE 为平均绝对百分比误差；RRMSE 为相对均方根误差；T 为运行时间。

Note: E_{max} is the maximum relative error; MAPE is the mean absolute percentage error; RRMSE is the relative root mean square error; T is the run time.

从图 5 和表 1 可看出，采用 FPSTWD-LSSVR 算法可以较好的在线拟合河蟹养殖生态环境因子与溶解氧浓度之间的复杂非线性关系，且预测曲线与实际监测值拟合效果明显好于 ILSSVR 和 SONB-LSSVRS 预测模型，精度虽略低于离线 LSSVR 拟合效果，但运行速度得到明显提高。

由表 1 统计结果知，在相同条件下，FPSTWD-LSSVR 算法与 ILSSVR 算法相比，评价指标 E_{max}、MAPE、RRMSE 和运行时间分别下降了 43.47%、47.93%、30.91%和 5.16s；FPSTWD-LSSVR 算法与稀疏在线无偏置最小二乘支持向量机 SONB-LSSVR 算法相比，评价指标 E_{max}、MAPE、RRMSE 和运行时间分别下降了 33.43%、39.99%、22.40%和 2.74 s；FPSTWD-LSSVR 算法与离线 LSSVR 算法相比，评价指标 E_{max}、MAPE、RRMSE 分别上升了 9.03%、16.14%、8.41%；运行时间下降 11.36 s。由图 5 和表 1 对比分析可知：1）在相同前提条件下，离线 LSSVR 预测精度最好，这与参与训练模型数据较多，获取的模型参数较优有关，但算法耗费的时间较多，不适于在线预测的实际需要；2）本文所提算法采用基于特征点的分段策略很好地保留了时间序列的历史知识特征信息，有效缩减在线建模和预测数据的规模，并且能够根据新增样本的序列特征，采用 FPSTWD 距离相似度计算实现预测子模型的自适应筛选，能在保证算法预测精度的同时降低时间复杂度；3)对于所有性能评价指标，由于本文提出的 FPSTWD-LSSVR 算法采用在线训练数据建模长度较小，与离线 LSSVR、ILSSVR 和 SONB-LSSVR 算法相比具有较好的综合性能，能够满足河蟹福利养殖水质在线预测的实际需求。

4 结论与讨论

1）针对以往在线预测方法存在动态学习和在线更新能力差、计算复杂度高、预测精度不理想等问题，该文以特征相似输入产生相似数据输出为指导思想，构建了基于时间序列相似性度量和 LSSVR 的河蟹养殖溶解氧在线预测模型，并对宜兴市集约化河蟹养殖池塘溶解氧进行在线预测，取得了较好的预测效果。在相同条件下与离线 LSSVR、ILSSVR 和 SONB-LSSVRS 预测模型对比分析，仿真结果表明，该文提出的 FPSTWD-LSSVR 在线预测模型各项性能评价指标最大相对误差为 4.76%、平均绝对百分比误差为 8.18%、相对均方根误差为 5.23%和运行时间为 8.32 s，远远优于 SONB-LSSVR 和 ILSSVR，略低于离线 LSSVR，这可能与子序列样本集 $R(S)$样本个数少、不能涵盖所有类型的时间序列数据特征有关；但随着不同特征类型的子序列样本集 $R(S)$的增加，其预测性能将会进一步改善。从在线预测的角度出发，该文提出的预测算法不仅降低了计算复杂度，还具有较高的在线预测精度，在综合性能上优于其他时间序列预测方法，为集约化河蟹福利养殖生态环境因子快速、在线预测提供了一种有效的解决方案，具有一定的理论指导意义及工程应用价值。

2）在 FPSTWD-LSSVR 建模过程中，采用特征点分段相似性度量策略，使得特征相似、规模适中、邻域信息宽泛的历史时间序列样本参与 LSSVR 模型的快速训练优化，实现了多个 LSSVR 子模型在线建模；通过 FPSTWD 距离算法对待预测的时间序列与 LSSVR 子模型相似性匹配，自适应选择特征相似和性能较佳的LSSVR子模型进行在线预测。该文所提出的在线预测算法在一定程度上解决了以往在线预测算法存在的动态学习和在线更新能力差、计算复杂度高、预测精度不理想等问题。

通过对集约化河蟹养殖池塘溶解氧在线预测，及时辨识河蟹养殖水环境的状况，为增氧机、水泵等设备智能化调控、养殖精细化管理提供决策支持，确保河蟹在最适宜的环境下生长。而子分段预测模型库规模影响着 FPSTWD-LSSVR 模型在线预测精度和性能，今后在研究中需要进一步探讨子分段预测模型库和标准支持向量表的更新策略，完善子分段预测模型库中子模型的种类，以提高最小二乘支持向量回归机在线预测的性能。

[参　考　文　献]

[1] 刘双印，徐龙琴，李道亮，等. 基于蚁群优化最小二乘支持向量回归机的河蟹养殖溶解氧预测模型[J]. 农业工程学报，2012，28(23)：167－175.
Liu Shuangyin, Xu Longqin, Li Daoliang, et al. Dissolved oxygen prediction model of eriocheir sinensis culture based on least squares support vector regression optimized by ant colony algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2012, 28(23): 167－175. (in Chinese with English abstract)

[2] 王军，彭喜元，彭宇. 一种新型复杂时间序列实时预测模型研究[J]. 电子学报，2006，12A：2391－2394.
Wang Jun, Peng Xiyuan, Peng Yu. A novel real time predictor for complex time series[J]. ACTA Electronica Sinica, 2006, (12A): 2391－2394. (in Chinese with English abstract)

[3] Vairappan C, Tamura H, Gao S C, et al. Batch type local search-based adaptive neuro-fuzzy inference system (ANFIS) with self-feedbacks for time-series prediction[J]. Neurocomputing, 2009, 72(7/8/9): 1870－1877.

[4] Chen Y M, Lin C T. Dynamic parameter optimization of evolutionary computation for on-line prediction of time series with changing dynamics[J]. Applied Soft Computing, 2007, 7(4): 1170－1176.

[5] Wang W J, Men C Q, Lu W Z. Online prediction model based on support vector machine[J]. Neurocomputing, 2008, 71(4/5/6): 550－558.

[6] Wen Y, Li X O. On-line fuzzy modeling via clustering and support vector machines[J]. Information Sciences, 2008, 178(22): 4264－4279.

[7] Gu B, Wang J D, Yu Y C, et al. Accurate on-line *v*-support vector learning[J]. Neural Networks, 2012, 27: 51－59.

[8] 张浩然，汪晓东.回归最小二乘支持向量机的增量和在线式学习算法[J].计算机学报，2006，29(3)：399－406.
Zhang Haoran, Wang Xiaodong. Incremental and online learning algorithm for regression least squares support vector machine[J]. Chinese Journal of Computers, 2006, 29(3): 399－406. (in Chinese with English abstract)

[9] Zhao Y P, Sun J G, Du Z H, et al. Online independent reduced least squares support vector regression[J]. Information Sciences, 2012, 201: 37－52.

[10] Zhang W P, Niu P F, Li G Q, et al. Forecasting of turbine heat rate with online least squares support vector machine based on gravitational search algorithm[J]. Knowledge-Based Systems, 2013, 39: 34－44.

[11] 张淑宁，王福利，何大阔，等. 在线鲁棒最小二乘支持向量机回归建模[J]. 控制理论与应用，2011，28(11)：1601－1606.
Zhang Shuning, Wang Fuli, He Dakuo, et al. Modeling method of online robust least-squares-support-vector regression[J]. Control Theory and Applications, 2011, 28(11): 1601－1606. (in Chinese with English abstract)

[12] 周欣然，滕召胜，蒋星军. 稀疏在线无偏置最小二乘支持向量机的预测控制[J]. 电子测量与仪器学报，2011，25(4)：331－337.
Zhou Xinran, Teng Zhaosheng, Jiang Xingjun. Predictive control using sparse online non-bias LSSVM[J]. Journal of Electronic Measurement and Instrument, 2011, 25(4): 331－337. (in Chinese with English abstract)

[13] 陈磊. 遗传最小二乘支持向量机法预测时用水量[J]. 浙江大学学报：工学版，2011，45(6)：1100－1103.
Chen Lei. Genetic least squares support vector machine approach to hourly water consumption prediction[J]. Journal of Zhejiang University: Engineering Science, 2011, 45(6): 1100－1103. (in Chinese with English abstract)

[14] Liu Shuangyin, Xu Longqin, Li Daoliang, et al. Prediction of dissolved oxygen content in river crab culture based on least squares support vector regression optimized by improved particle swarm optimization[J]. Computers and Electronics in Agriculture, 2013, 95: 82－91.

[15] 肖辉，胡运发. 基于分段时间弯曲距离的时间序列挖掘[J]. 计算机研究与发展，2005，42(1)：72－78.
Xiao Hui, Hu Yunfa. Data mining based on segmented time war ping distance in time series database[J]. Journal of Computer Research and Development, 2005, 42(1): 72－78. (in Chinese with English abstract)

[16] 程文聪，邹鹏，贾焰. 多维时序数据中的相似子序列搜索研究[J]. 计算机研究与发展，2010，47(3)：416－425.
Chen Wencong, Zou Peng, Jia Yan. Similar sub-sequences search over multi-dimensional time series data[J]. Journal of Computer Research and Development, 2010, 47(3): 416－425. (in Chinese with English abstract)

[17] Park S, Wesley W C, Yoon J, et al. Similarity search of time-wraped subsequences via a suffix tree[J]. Information Systems, 2003, 28(7): 867－883.

[18] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series[C]// Proc of the AAAI-94 Workshop on Knowledge Discovery in Databases. Washington: AAAI Presss, 1994: 359－370.

[19] 高中灵，徐新刚，王纪华，等. 基于时间序列 NDVI 相似性分析的棉花估产[J]. 农业工程学报，2012，28(2)：148－153.
Gao Zhongling, Xu Xingang, Wang Jihua, et al. Cotton yield estimation based on similarity analysis of time-series NDVI[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2012, 28(2): 148－153. (in Chinese with English abstract)

[20] 金立兵，金伟良，王海龙，等. 多重环境时间相似理论及其应用[J]. 浙江大学学报：工学版，2010，44(4)：789－797.
Jin Libing, Jin Weiliang, Wang Hailong, et al.

Multi-environmental time similarity theory and its application[J]. Journal of Zhejiang University: Engineering Science, 2010, 44(4): 789－797. (in Chinese with English abstract)

[21] 万安平，陈坚红，盛德仁，等. 基于多重环境时间相似理论的燃气轮机热通道部件剩余寿命预测方法[J]. 中国电机工程学报，2013，33(5)：95－100.
Wan Anping, Chen Jianhong, Sheng Deren, et al. Residual life prediction method for gas turbine HGP component based on multi-environmental time similarity theory[J]. Proceedings of the CSEE, 2013, 33(5): 95－100. (in Chinese with English abstract)

[22] 杨锡运，孙宝君，张新房，等. 基于相似数据的支持向量机短期风速预测仿真研究[J]. 中国电机工程学报，2012，33(5)：35－40.
Yang Xiyun, Sun Baojun, Zhang Xinfang, et al. Residual life prediction method for gas turbine HGP component based on multi-environmental time similarity theory[J]. Proceedings of the CSEE, 2012, 33(5): 35－40. (in Chinese with English abstract)

[23] 刘晶. 基于相似日和支持向量机的短期负荷预测研究[D]. 广州：华南理工大学，2010.
Liu Jing. Short-term Power Load Forecasting Based on Similar Day and Support Vector Machine[D]. Guangzhou: South China University of Technology, 2010. (in Chinese with English abstract)

[24] 刘大同，彭宇，彭喜元，等. 一种分段在线支持向量回归算法[J]. 仪器仪表学报，2010，31(8)：1732－1737.
Liu Datong, Peng Yu, Peng Xiyuan, et al. Segmental online support vector regression algorithm[J]. Chinese Journal of Scientific Instrument, 2010, 31(8): 1732－1737. (in Chinese with English abstract)

Online prediction for dissolved oxygen of water quality based on support vector machine with time series similar data

Liu Shuangyin[1,2,3], Xu Longqin[1], Li Daoliang[2,3,4※], Zeng Lihua[3,4,5]

(1. *College of Information, Guangdong Ocean University, Zhanjiang* 524025, *China*;
2. *Key Laboratory of Agricultural Information Acquisition Technology*(*Beijing*), *Ministry of Agriculture, Beijing* 100083, *China;*
3. *Beijing ERC for Internet of Things in Agriculture, China Agricultural University, Beijing* 100083, *China*;
4. *Beijing ERC for Advanced Sensor Technology in Agriculture, China Agricultural University, Beijing* 100083, *China*
5. *College of Mechanical and Electrical Engineering, Agricultural University of Hebei, Baoding* 071001, *China*)

Abstract: Water quality regulation is one of the most important tasks in intensive aquaculture management. Grasping the trend of the dissolved oxygen concentration timely and accurately and regulating water quality dynamics are the key for healthy growth in the non-stress environment of aquatic products in order to solve the low prediction accuracy, inferior capability of dynamic learning, online updates, and high computational complexity of the traditional online forecasting methods for water quality in intensive aquaculture. The online prediction model of dissolved oxygen content in intensive aquaculture eriocheir sinensis cultures was introduced, which was based on the least squares support vector machine (LSSVR) with time series similar data. The time series data collected online was segmented clustered using a feature points segmented time warping distance algorithm. The subsequence data sets reduced the size and optimized the LSSVR models training process, achieving multiple LSSVR models online modeling, and segmented memory and storage. According to the forecast data sequence and LSSVR sub-model similarity, it adaptively chose the optimal sub-model to get the predicted output. The online model was used for the prediction of the dissolved oxygen changing in high-density eriocheir sinensis culture ponds during July 21, 2012 to July 31, 2012 in Yixing City, Jiangsu Province, China. Experimental results showed that the proposed prediction model of FPSTWD–LSSVR had a better prediction effect than the FPSTWD–LSSVR, ILSSVR, SONB-LSSVR, or off-line LSSVR algorithms. Under the same experimental conditions, the relative mean absolute percentage error (MAPE), maximum relative error (E_{max}), relative root mean square error (RRMSE), and the running time differences between the FPSTWD–LSSVR and ILSSVR models were 47.93%, 43.47%%, 30.91%, and 5.16 s in the test period respectively. The relative MAPE, E_{max}, RRMSE, and the running time differences between the FPSTWD–LSSVR and SONB-LSSVR models were 39.99%, 33.43%, 22.40%, and 2.74 s in the test period respectively. It is obvious that FPSTWD–LSSVR is more accurate than ILSSVR and SONB-LSSVR. The relative MAPE, E_{max}, RRMSE and the running time differences between the FPSTWD–LSSVR and off-line LSSVR models were 16.14%, 9.03%, 8.41%, and 11.36 s in the test period respectively. The lower sample number, which cannot cover all types of characteristic in time series data, probably caused the prediction performance of FPSTWD-LSSVR to be slightly lower than the off-line LSSVR model. Overall, the online prediction model has a low computational complexity, fast convergence rate, high online prediction accuracy, and strong generalization ability. It is an effective online prediction method for the dissolved oxygen controlling in the high density eriocheir sinensis culture, and provides the basis of decisions for controlling water quality, setting the aquaculture water plan, and reducing the risk of cultivation.

Key words: aquaculture; water quality; models; support vector machine; online prediction; feature points segmented time warping distance; similar data

（责任编辑：刘丽英）