基于 *K*-means 聚类和 ELM 神经网络的养殖水质溶解氧预测

宦 娟[1,2]，刘星桥[1※]

（1. 江苏大学电气信息工程学院，镇江 212013； 2. 常州大学信息科学与工程学院，常州 213164）

摘 要：为解决养殖水质溶解氧预测传统方法引入不良样本、精度低等问题，该文以 2014、2015 年江苏常州养殖基地水质和气象数据为基础，提出了一种基于 *K*-means 聚类和 ELM 神经网络（extreme learning machine，ELM）的溶解氧预测模型。采用皮尔森相关系数法确定环境因素与溶解氧的相关系数，自定义相似日的统计量－相似度，通过 *K*-means 聚类方法将历史日样本划分为若干类，然后分类识别获得与预测日最相似的一类历史日样本集，将其与预测日的实测环境因素作为预测模型的输入样本建立 ELM 神经网络溶解氧预测模型。试验结果表明，该模型均具有较快的计算速度和较高的预测精度，在常规天气下，平均绝对百分误差和均方根误差分别达到 1.4%、10.8%；在突变天气下，平均绝对百分误差和均方根误差分别达到 2.6%和 11.6%，有利于水产养殖水质精准调控。

关键词：神经网络；模型；养殖；溶解氧预测；相似日；*K*-means 聚类；ELM 神经网络

doi：10.11975/j.issn.1002-6819.2016.17.024

中图分类号：TP391 **文献标志码**：A **文章编号**：1002-6819(2016)-17-0174-08

宦 娟，刘星桥. 基于 *K*-means 聚类和 ELM 神经网络的养殖水质溶解氧预测[J]. 农业工程学报，2016，32(17)：174－181. doi：10.11975/j.issn.1002-6819.2016.17.024 http://www.tcsae.org

Huan Juan, Liu Xingqiao. Dissolved oxygen prediction in water based on *K*-means clustering and ELM neural network for aquaculture[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2016, 32(17): 174－181. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2016.17.024 http://www.tcsae.org

0 引 言

随着水产养殖业的发展，加强养殖水质预测关键技术研究[1]，提升水产养殖减灾防灾能力，保障水产养殖安全生产已成为农业生产的关键内容之一。在集约化养殖池塘中，溶解氧（dissolved oxygen，DO）是养殖池塘水质的一项重要指标，氧气不足或过量都会对鱼类的生存环境产生不利影响。如何利用现有实时监测信息，准确预测未来水中溶解氧变化趋势，及时做出合理的决策，已成为国内外学者关注和研究的热点。

在预测方法上，国内外研究者们的方法可大致分为传统方法和现代方法 2 大类。传统方法主要是根据水质变化特性来开展研究，如成因分析法和水文统计法[2]。现代方法是依托计算机技术而拓展的新途径[3]，目前应用较广泛的现代方法包括人工神经网络方法[4]、模糊方法[5]、混沌方法[6]、小波分析法[7]、灰色系统方法[8]等。这些方法对溶解氧、pH 值、水温、氨氮、总氮、叶绿素、化学需氧量（chemical oxygen demand，COD）等水质参数预测中取得了较好的效果。国内学者刘双印[9-10]等利用支持向量机对养殖水体中的 pH 值的含量进行了预测分析。刘明等[11]提出基于小波分解的 ARMA 预测模型，预测了养殖水体中亚硝酸盐的变化。这些方法主要是利用现代智能算法自身优势展开的预测应用，但往往没有有效地分析环境变化对溶解氧变化过程的影响，未能揭示本质规律，从而影响变化环境下的水质模拟和预报精度。一方面，若对环境因素的处理不够精细，当环境条件发生剧烈变化时，如天气类型、气温、气压、风速等突变时，模型预测精度将受到影响。另一方面，在相似环境因素的影响下，溶解氧变化曲线呈现一定的规律，具备一定的相似性，因而综合考虑各种环境因素，优化选取样本，通过选取相似日建立溶解氧预测模型以提高预测精度具有较好的实用性和可行性。

以优化样本空间提高执行效率和预测精度为目的，本文提出了基于相似日模糊聚类的溶解氧预测模型。由于相似日的溶解氧曲线具有很高的关联度，通过 *K*-means 聚类方法将历史日样本划分为若干类，然后分类识别获得与预测日最相似的一类历史日样本集，建立（extreme learning machine，ELM）神经网络溶解氧预测模型。以中国江苏常州养殖基地采集到的数据为数据源，对模型进行训练和预测，预测曲线能有效反映溶解氧的变化趋势，预测值能够很好地拟合实际值。

1 数据和方法

1.1 试验区域

选取江苏省南部常州市武进名优水产引繁推广中心（常州市武进水产养殖场 31°48′～31°69′N，119°75′～

收稿日期：2016-02-23 修订日期：2016-06-02

基金项目：江苏高校优势学科建设工程资助项目（PAPD, NO.6-2014）；江苏省农业科技支撑项目（BE2013402）；常州市科技支撑计划（CJ20140057）

作者简介：宦娟，女，副教授，博士生，2015 年赴美国德克萨斯大学达拉斯分校研修，主要从事农业信息化研究。镇江 江苏大学电气工程学院，212013。Email：huanjuan@cczu.edu.cn

※通信作者：刘星桥，教授，博士生导师，主要从事农业设施智能控制系统研究。镇江 江苏大学电气工程学院，212013。E-mail: xqliu@ujs.edu.cn 中国农业工程学会会员：刘星桥（E041200581S）

119°89′E）为试验区域。区域内河流密布，水质清澈，共占地面积 150 hm^2，建有高标准现代化养殖池塘 69 hm^2，育苗孵化车间 916 m^2，区水生动物疫病防治实验室 343 m^2，3 000 m^2 的循环水养殖车间，池塘水循环养殖系统 133 hm^2。中心全面应用增氧机、全自动投饵机、智能溶氧检测仪、微孔管道增氧及水产养殖远程无线监控系统等现代渔业装备，进排水系统全面配套，水、电、路设施齐全。

1.2　数据来源

本试验选取区域内某个 2.5 hm^2 的标准池塘河蟹养殖池塘作为试验池，养殖环境数据来源于江苏大学研制的水产养殖远程无线监控系统[12-13]。如图 1 所示，该系统每间隔 15min 对 pH 值、水温、溶解氧、浊度、叶绿素、天气类型、气温、气压、湿度等环境特征因素在线采样一次，将 2014 年 6 月 1 日－9 月 30 日、2015 年 6 月 1 日－9 月 30 日的 244 天 23424 个数据作为研究的数据源。由于监测数据具有日周期性质，故以每日数据集为一个标准单位定义为一个样本。在 244 个样本中，训练集和测试集比例为 7∶3，溶解氧预测模型经训练集训练后，进一步用测试集验证其性能。

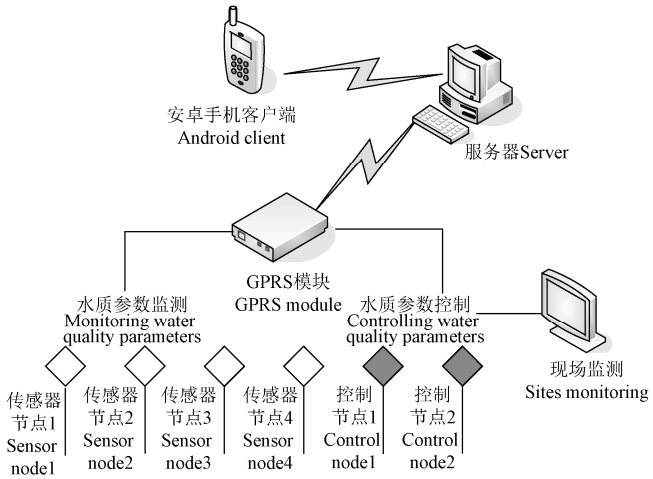


图 1　水产养殖无线水质远程监测系统

Fig.1　Wireless remote monitoring system of aquaculture water quality

1.3　建模步骤

养殖水体中的溶解氧变化复杂，合适的数据样本可提高预测精度，本文提出了基于相似日、*K*-means 聚类和 ELM（extreme learning machine）神经网络集合的组合预测模型，如图 2 为预测流程图。建模步骤如下：

1）相似度统计量构造。数据归一化后，利用皮尔森相关系数确定环境因子权重，构造相似日的统计量－相似度。

2）*K*-means 聚类[14]。根据相似度应用 *K*-means 聚类法对历史日数据样本聚类，找出合适样本，使得历史日样本被分为若干类。

3）预测日所属类别识别。以相似度最大的类别作为预测日的类别，形成训练样本。

4）ELM 神经网络建模与预测。利用训练样本建立 ELM 神经网络模型，利用测试样本对模型进行验证。最后，经过补偿及反归一化过程得出最后预测值。

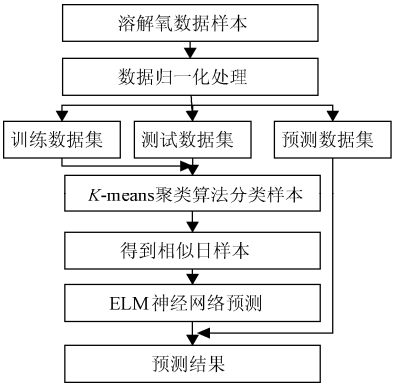


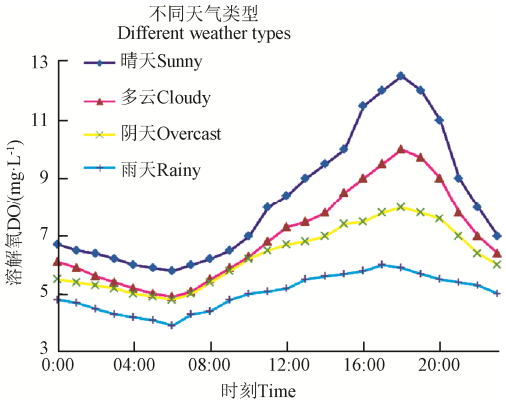
图 2　预测流程图

Fig.2　Flowchart of prediction algorithm

2　基于相似日聚类的预测模型设计

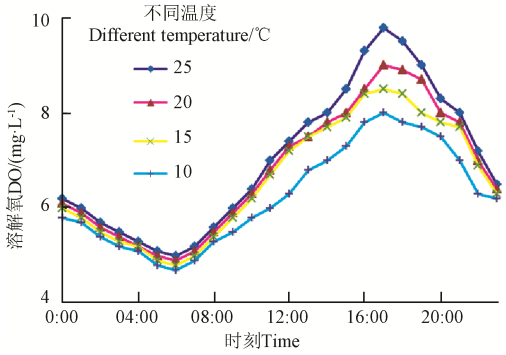
2.1　日特征相关因素权重确定

与溶解氧有关的因素有天气类型、风速、气温、水温、气压、pH 值等。很显然，天气类型因素对溶解氧的影响最大，原因是天气类型不同，进而太阳辐射强度不同引起植物光合作用的差异。为方便计算，本文将天气类型量化为光照强度。图 3a 是不同天气类型和不同温度条件下日平均溶解氧变化曲线。



a. 不同天气类型下的溶解氧曲线

a. DO curve under different weather types



b. 不同温度下的溶解氧曲线

b. DO curve under different temperatures

图 3　不同天气类型和不同温度下的溶解氧曲线

Fig.3　DO curve under different weather types and different temperatures

可以看出，溶解氧曲线随着天气类型的不同变化很大，这种差异不仅体现在溶解氧的变化趋势上，也体现在输出浓度值的大小上。图 3b 是相同天气类型、不同温度条件下溶解氧变化曲线，可以看出，在相同的天气类型条件下，溶解氧曲线具有相同的变化趋势，但是其浓度值大小随着温度的不同具有一定的差异，这主要是不同温度下太阳辐射强度差异所致。除此以外，风速、气温、水温、气压、pH 值等也会影响水中溶解氧浓度。

在多环境因素并存的情况下，溶解氧受天气类型、风速、气温、水温、气压、pH 值、氨氮等诸多因素影响。本文采用距离分析法中皮尔森相似度识别变量之间的相依关系。

首先，为避免因环境因素量纲和数量级的不同造成数量级大的特性指标左右分类结果，本文采用“极差化”方法对原始数据无量纲化处理，将其归一映射至[0,1]区间。

$$x = \frac{x' - x_{\min}}{x_{\max} - x_{\min}}。 \tag{1}$$

式中 x'、$x_{\min}$、$x_{\max}$ 分别为原始输入数据、原始输入数据中的最小值、最大值，x 为归一化后的值。接着将 23424 个数据代入皮尔森相关系数计算公式（2），计算溶解氧与各环境因素之间的相关系数。结果见表 1。

$$\sigma_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{m}(y_i - \bar{y})^2}}。 \tag{2}$$

式中 x_i、y_i 分别为向量 x 和 y 中的第 i 个元素；$\bar{x}$、$\bar{y}$ 分别为向量 x 和 y 中元素的平均值。

从表 1 可以看出溶解氧与光照 S、温度 T、气压 H、风速、湿度和 pH 值相关性较大，因此选取这 6 个环境因素体现每日特征。

表 1 各环境因素与溶解氧的相关系数

Table 1 Correlation coefficients between environment factors and dissolved oxygen

指标 Indicator	相关系数 Correlation coefficients							
	温度 Temperature	光照 Illumination	气压 Pressure	氨氮 Ammonia nitrogen	湿度 Humidity	亚硝酸盐 Nitrite	风速 Wind speed	pH 值 pH value
溶解氧 Dissolved oxygen	0.3423	0.2937	0.2702	-0.028	-02386	-0.03	0.2621	0.2815

2.2 相似日 *K*-means 聚类

为避免因引入不良样本而导致精度低、收敛慢的问题，本文提出在预测前先用 *K*-means 聚类方法选取与预测日具有高度相似特征的相似日数据，然后以其为样本建立 ELM 神经网络模型[14]进行溶解氧预测。

2.2.1 *K*-means 聚类

当前数据挖掘技术中，*K*-means 聚类是最常用的聚类算法之一。其基本原理是按指定聚类数 s，将 n 个样本分为 s 个簇。通过计算当前样本到簇中心的距离，把当前样本分到距离最近的簇，反复迭代直至同一簇中的样本相似度尽可能的高，不同簇中相似度尽可能的小，最终得到最佳聚类。典型步骤如下:

1）任意指定 n 个样本中的 s 个对象作为簇中心；

2）计算当前和所有簇中心之间的相似度（距离）；

3）确定样本归属，当相似度大于阈值则形成新簇；

4）重新计算聚类后所得簇的簇中心，重复步骤 2～3 直到每个簇不再发生变化为止。

本文中，设论域 $\boldsymbol{U}$ 为 n 个待分类样本，$\boldsymbol{x_1}$ 为第 1 天样本，$\boldsymbol{x_n}$ 为第 n 天样本，即

$$\boldsymbol{U} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}。 \tag{3}$$

且一天 96 个时间段的监测数据集构成一个样本单位，同时每个样本都有 m 个指标表示其性状，由此可得到论域的每个样本的数据矩阵为

$$\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{x}_{11i} & \boldsymbol{x}_{12i} & \cdots & \boldsymbol{x}_{1m,i} \\ \boldsymbol{x}_{21i} & \boldsymbol{x}_{22i} & \cdots & \boldsymbol{x}_{2mi} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{x}_{t1i} & \boldsymbol{x}_{t2i} & \cdots & \boldsymbol{x}_{tmi} \end{bmatrix}。 \tag{4}$$

式中 $\boldsymbol{x_i}$ 为第 i 天的样本，m 为特征因素，t 为监测时刻，t=96。计算待分类样本与中心之间的相似度，按不同的相似度值进行分类，得到不同的聚类结果。

2.2.2 聚类相似度统计量的改进

在聚类分析中通常用欧式距离和夹角余弦来计算相似度类，两种方法的定义如图 4a，可以看出，欧式距离体现样本值相似程度，而夹角余弦体现样本形相程度，两者都不能完全准确的反映样本间的真正相似程度，有一定的局限性。正如图 4b，在二维平面上，A 点与 B 点、C 点的距离分别用 dist(A,B)、dist(A,C)表示，向量 $\overrightarrow{\boldsymbol{OB}}$、$\overrightarrow{\boldsymbol{OC}}$ 与向量 $\overrightarrow{\boldsymbol{OA}}$ 的夹角分别为 β、γ，以 A 为圆心，r 为半径做圆，显然，dist(A,B)=dist(A,C)，但欧式距离无法进一步判断临近距离，根据 cosβ>cosγ，可以判断 C 点到 A 点更接近些。另外，这 2 种方法均未考虑样本特征因素之间的重要度差异。

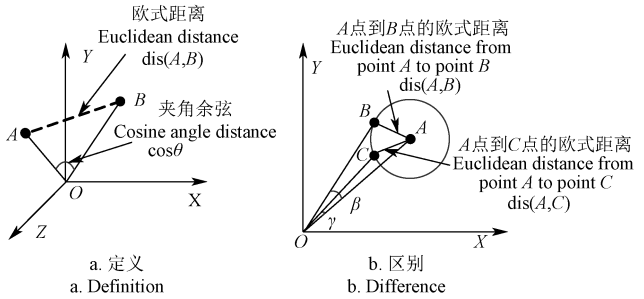
为了综合考虑历史数据中的数值信息和趋势信息，综合欧式距离和夹角余弦计算法，本文提出了一种改进的相似度统计量 D_{xy}。D_{xy} 表示样本 $\boldsymbol{x}$ 和样本 y 之间的相似度，越接近 0 越相似。

$$D_{xy} = \alpha d_{xy} + \beta D_{\cos xy}。 \tag{5}$$

$$d_{xy} = 1 - \frac{1}{t}\sum_{i=1}^{t}\sqrt{\frac{1}{m}\sum_{j=1}^{m}\sigma_j\left(x_{ij} - y_{ij}\right)^2}。 \tag{6}$$

$$D_{\cos xy} = \frac{1}{t}\sum_{i=1}^{t}\frac{\sum_{j=1}^{m}\sigma_j x_{ij} y_{ij}}{\sqrt{\sum_{j=1}^{m}x_{ij}^{\,2}\sum_{j=1}^{m}y_{ij}^{\,2}}}。 \tag{7}$$

式中样本 $\boldsymbol{x}$ 和样本 $\boldsymbol{y}$ 在 i 时刻、j 特征因素的数值分别用 x_{ij}、y_{ij} 表示，且在[0,1]之间；样本中第 j 个特征因素的权值用 σ_j 表示；α 和 β 为欧式距离 d_{xy} 和夹角余弦 $D_{\cos xy}$ 权重系数，相似度统计量由值系数和形系数两项共同决定。α+β=1，不同天气情况取值会不同，若环境因素在一天内有明显变化，α 应取接近 1，否则接近 0。



a. 定义　　b. 区别

a. Definition　　b. Difference

注：图 a 中，*X*、*Y*、*Z* 分别为三维坐标系中的 *X* 轴、*Y* 轴和 *Z* 轴；*O* 为坐标原点；*A*、*B* 分别为坐标系中的两点。图 b 中，*X*、*Y* 分别为二维坐标系中的 *X* 轴、*Y* 轴；*O* 为坐标原点；*A*、*B*、*C* 分别为坐标系中的三点。β、γ 分别为向量 $\overrightarrow{OA}$ 与向量 $\overrightarrow{OB}$ 之间的夹角、向量 $\overrightarrow{OA}$ 与向量 $\overrightarrow{OC}$ 之间的夹角。

Note: In Figure a, *X*, *Y*, *Z* are the *X* axis, *Y* axis and *Z* axis respectively of the three-dimensional coordinates. *O* is origin of coordinate. *A* and *B* are the two points in the coordinate system. In Figure b, *X*, *Y* are the *X* axis and *Y* axis of the two-dimensional coordinate system. *O* is origin of coordinate. *A*, *B* and *C* are the three points in the coordinate syste. β is the angle between the vector $\overrightarrow{OA}$ and the vector $\overrightarrow{OB}$. γ is the angle between the vector $\overrightarrow{OA}$ and the vector $\overrightarrow{OC}$.

图 4　欧式距离及夹角余弦原理示意图

Fig.4　Schematic diagram on theory of Euclidean distance and cosine angle distance

2.2.3　新样本分类识别

原始样本经过 *K*-means 聚类后得到 s 个簇，待预测的新样本需进一步识别所属簇。首先根据聚类情况，各簇的中心按公式（8）计算。

$$\overline{x}_i = \frac{1}{p}\sum_{j=1}^{p} x_j \text{。} \tag{8}$$

式中第 i 类包括 p 个样本。

然后，待预测时刻点之前的 95 个时刻点监测数据组成一个预测日样本，根据公式（5）计算得出预测日样本与各聚类中心的相似度，相似度最大的簇作为预测日样本的所属簇。

2.3　ELM 神经网络的预测

黄广斌[15]等人提出的极限学习机器（extreme learning machine，ELM）是一种常用的学习算法，它属单隐层前馈神经网络，隐藏层的权重和阈值随机生成[16]。与传统的 BP 神经网络（back propagation，BP）和支持向量机（support vector machine，SVM)相比，其参数设置容易、收敛速度快、泛化能力强，且效果也很精确，同时可以克服局部极小过拟合问题。ELM 三层网络结构能逼近任何非线性函数。网络模型通常分为 3 层：输入层，隐含层和输出层，网络的拓扑结构如图 5 所示。

网络中输入层神经元 p 个，输出层神经元 1 个，隐含层神经元 L 个，输入层和隐含层之间的激活函数[16]为 f，对于 N 个样本总数，$\{x_i,y_i\}N_i$=1 为数据样本集，其中 $\{x_{1i}, x_{2i},\ldots, x_{pi}\} \in R^p$。那么网络的输入和输出可表示为

$$\sum_{t}^{L} \beta_t f(W_t, b_t, X_i) = o_i \qquad i = 1,2,\cdots,N \tag{9}$$

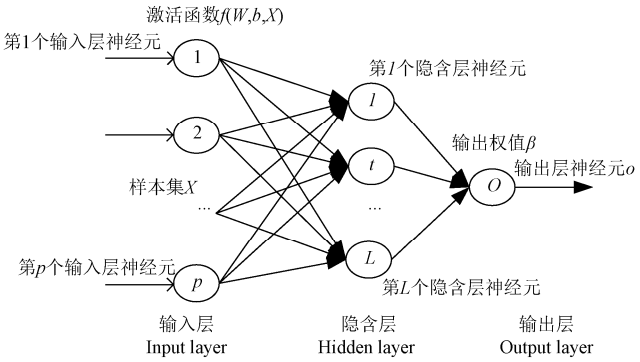


图 5　ELM 神经网络拓扑结构

Fig.5　Topological structure of ELM neural network

其中，$\boldsymbol{W}_t=[\boldsymbol{W}_{t1}\ldots\boldsymbol{W}_{td}]'$ 为第 t 个隐含层神经元与输入向量 $\boldsymbol{X}_i$ 间的权值向量；b_t 为第 t 个隐含层神经元的阈值。本文以公式（10）sigmoid 函数为激活函数，用激活函数以 0 误差逼近这 N 个样本，存在 $\boldsymbol{\beta}$，$\boldsymbol{W}$，b 使得公式（12）成立。

$$f(u) = \frac{1}{1+e^{-u}} \text{，} \tag{10}$$

$$\sum_{i=0}^{N} \left\| o_i - y_i \right\| = 0 \text{，} \tag{11}$$

$$\sum_{t}^{L} \beta_t f(W_t, b_t, X_i) = y_i \qquad i = 1,2,\cdots,N \text{。} \tag{12}$$

上式可简写为

$$H \cdot \beta = Y \text{。} \tag{13}$$

其中

$$\boldsymbol{H} = \begin{bmatrix} f(W_1,b_1,\boldsymbol{X}_1) & \cdots & f(W_L,b_L,\boldsymbol{X}_1) \\ & \vdots & \\ f(W_1,b_1,\boldsymbol{X}_N) & \cdots & f(W_L,b_L,\boldsymbol{X}_N) \end{bmatrix}_{N\times L} \text{。} \tag{14}$$

$$\beta = [\beta_1', \cdots, \beta_L']_L' \text{。} \tag{15}$$

$$Y = [Y_1', \cdots, Y_N']_N' \text{。} \tag{16}$$

式中 $\boldsymbol{H}$ 是网络的隐含层输出矩阵，$\boldsymbol{H}$ 的第 i 列是第 i 个隐节点关于输入 $\boldsymbol{X_1}$,$\boldsymbol{X_2}$,...,$\boldsymbol{X_N}$ 的输出向量，$\boldsymbol{H}$ 的第 j 行是隐含层关于输入 $\boldsymbol{X_j}$ 的输出向量。

$$\boldsymbol{\beta} = \boldsymbol{H}^{+} Y \text{。} \tag{17}$$

式中 $\boldsymbol{H}^{+}$ 是 $\boldsymbol{H}$ 的逆矩阵，权值 $\boldsymbol{\beta}$ 的范数越小，泛华能力越强。

ELM 网络算法[15]包含 3 个步骤：

1）随机生成隐单元的输出权值 W_t 和偏置 b_t,t=1,2...,L；

2）计算隐单元的输出矩阵 $\boldsymbol{H}$；

3）计算输出权值 $\boldsymbol{\beta}$：$\boldsymbol{\beta}$=$\boldsymbol{H}^{+}\boldsymbol{Y}$。

3　结果和分析

3.1　*K*-means 聚类算法的有效性评估

评判聚类结果的有效性是一个困难而复杂的问题。Halkidi 等[17]提出的有效性指数法，是一种基于聚类平均散布性和聚类间总体分离性的相对度量方法。有效性指数计算公式如下

$$\mathrm{SD}=\frac{Scatt\_Comp}{d}。\qquad(18)$$

式中，*Scatt_Comp*、*d* 分别为聚类的平均散布性和聚类间总体分离性。SD 为有效性指数，SD 值越小越好。

为了建立最佳的簇群，我们选择一个具有最小 SD 的分簇。计算结果如表 2 所示。

表 2 不同分簇下的 SD 值

Table 2 Clusters with different SD values

簇数目 Number of clusters(*K*)	*K*=2	*K*=3	*K*=4	*K*=5	*K*=6
有效性指数 SD	0.87	0.52	0.41	0.38	0.40

从表 2 不难看出，分簇数由 2 到 6，产生的 SD 不同。当分簇数为 5 时，SD 值最小为 0.38。因此，可以确定，5 是最适当的簇数目。

3.2 ELM 神经网络的性能

由 3.1 小节得出的结论，将 244 天样本集聚类到 5 个簇群，然后被用作 ELM 神经网络的输入。优化后的网络结构如表 3 所示，不同的训练和测试集影响隐层节点数。显然易见，随着隐含层节点数的增加，从 18 到 77，均方根误差逐渐增大。

表 3 5 个簇下神经网络结构

Table 3 Neural network structure with five clusters

簇序号 Serial number of clusters	样本集 Number of day sample	训练集 Number of training records	测试集 Number of testing records	网络结构 Structure of neural network	均方根误差 RMSE/%
1	38	2592	1056	9-33-1	11.15
2	57	3840	1632	9-48-1	14.58
3	81	5472	2304	9-77-1	15.02
4	47	3168	1344	9-51-1	14.77
5	21	1440	576	9-18-1	9.86

为进一步检验测 ELM 神经网络的性能，预测结果由运算时间、平均绝对百分误差（MAPE）和均方根误差（RMSE）[18-19]进行评估。由于该算法中每个解的维数较大，需要更多的训练时间来达到较小的误差，因此预测速度也是衡量的一个指标。运算时间由 tic /toc 指令测试。以表 3 中的簇 2 为例，将 ELM 神经网络与传统的 BP 神经网络、支持向量机（SVM）进行比较。预测结果见表 4，可以看出，BP 神经网络可以通过遗传算法找到最优初始参数提高预测精度，但较耗时。支持向量机耗时且预测精度低。虽然 BP 神经网络具有最小预测误差，但其计算时间比 ELM 网络的计算时间长。假设每个群簇的计算时间是相同的，运算较快的 BP 神经网络的总计算时间是 2.05s×5 = 10.25s，而 ELM 神经网络只需要 0.02s×5＝0.1s。可见，ELM 神经网络相比于其他预测算法具有较高的预测精度和运算速度，优势明显。

$$\mathrm{MAPE}=\frac{1}{T}\sum_{t=1}^{T}\left|\frac{Y_t-f_t}{Y_t}\right|。\qquad(19)$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t-f_t)^2}\quad。\qquad(20)$$

式中 MAPE 为平均绝对百分误差；RMSE 为均方根误差，Y_t和f_t分别为真实观测值和预测值；T为预测时间点的个数。

表 4 不同预测算法的性能对比

Table 4 Performance of different forecasting methods

预测模型 Prediction model	运算时间 Running time/s	平均绝对百分误差 MAPE/%	均方根误差 RMSE/%
神经网络 BP	2.05	1.47	13.93
支持向量机 SVM	4.75	7.41	27.83
极限学习机 ELM	0.02	1.5	14.58

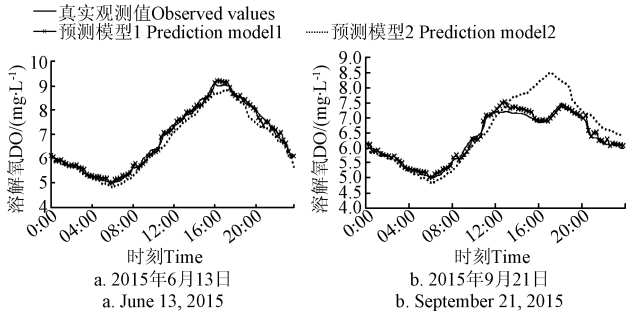
3.3 基于 *K*-means 聚类和 ELM 神经网络模型的预测精度

在这一节中，为了说明本文提出的构建相似日序列和 *K*-means 聚类与 ELM 结合对提升预测精度的作用，将 *K*-means 聚类和 ELM 组合预测模型（简称模型 1）、传统的 ELM 神经网络预测模型（简称模型 2，不做相似日分类）与真实观测值作以对比。

模型 1 根据自定义的相似度统计量，利用 *K*-means 聚类识别了预测日的相似日样本，以此构建 ELM 预测模型，该过程为单步滚动预测。而模型 2 直接选择预测日前 28 天的数据作为 ELM 神经网络的输入，不做任何分类处理。

为了验证预测模型在不同天气状况下的预测效果，分别选取 2015 年 6 月 13 日和 2015 年 9 月 21 日为验证日。2015 年 6 月 13 日是常规天气，晴天，最高、最低和平均温度分别为 30、21 和 27 ℃，归一化后的光照强度为 0.85。模型 1、模型 2 的溶解氧预测值与真实观测值对比如图 6a 所示，整体而言，模型 1 的结果接近真实值，明显优于模型 2，说明因溶解氧受到天气、温度、光照条件等多因素影响，进行相似日聚类能找到相似样本，剔除较大差异样本，从而有效提升了溶解氧的预测精度。而模型 2 直接将近期数据作为输入数据，形成的训练样本包括阴雨天和阴天，这与预测日的晴天相背，从而引起模型 2 的预测值比真实值低。从表 5，模型 1 和模型 2 平均绝对百分误差（MAPE）分别为 1.4%和 3.1%，均方根误差（RMSE）分别为 10.8%和 24.9%。可见只依靠网络自身训练的方式，让 ELM 神经网络建立溶解氧和环境因素的关系，预测效果并不理想。

2015 年 9 月 21 日这天是特殊天气，期间有突变情况，多云，最高、最低和平均温度分别为 26、17、19℃，归一化后的光照强度为 0.55，且在中午 12 点－4 点下雨，然后放晴。图 6b 中，模型 1 可以构造有效最优相似样本集，能够识别预测日是突变天气，在前几小时预测误差很大时，及时缩短实时监测时段，加大模型更新实时数据的频率，快速得到最新训练样本，从而降低后面时刻预测的误差，其结果接近真实值。而模型 2 的预测误差较大，存在相对较大的波动，其预测值比真实值大，尤其是在下午，这是因为模型 2 未能识别突变天气，输入样本不够优。从表 5，模型 1 和模型 2 平均绝对百分误差（MAPE）分别为 2.6%和 6.8%，均方根误差（RMSE）分别为 11.6%和 55.1%。因此总体而言，*K*-means 聚类和 ELM 组合预测模型的预测效率优于单纯的 ELM 算法。



注：模型 1 为 *K*-means 聚类与 ELM 结合预测模型。模型 2 为传统 ELM 预测模型。

Note: Mode l is *K*-means clustering and ELM prediction model, while Model 2 is traditional ELM prediction model.

图 6　2 种模型溶解氧的预测值与真实观测值对比

Fig.6　Comparison between observed and predicted DO of two prediction model

然而，从表 5 的 2 种模型预测结果误差分析看出，模型 1 和模型 2 在 2015 年 9 月 21 日的预测精度比 2015 年 6 月 13 日均有所下降，这是由于天气突变下数据的相似度不高，影响构建相似日序列，从而影响预测精度，还有待进一步优化。

表 5　2 种模型预测结果误差分析

Table 5　Error analysis for two prediction models

日期 Date	平均绝对百分误差 MAPE/%		均方根误差 RMSE/%	
	预测模型 1 Prediction model 1	预测模型 2 Prediction model 2	预测模型 1 Prediction model 1	预测模型 2 Prediction model 2
2015-06-13	1.4	3.1	10.8	24.9
2015-09-21	2.6	6.8	11.6	55.1

注：预测模型 1 为 *K*-means 聚类与 ELM 结合预测模型。预测模型 2 为传统 ELM 预测模型。

Note: Prediction mode l is *K*-means clustering and ELM prediction model,while Prediction mode 2 isTraditional ELM prediction model.

4　结　论

本文针对具有自相似性特征的养殖溶解氧预测精度低的问题，利用 *K*-means 聚类、ELM 神经网络对溶解氧预测进行了研究，得出以下结论：

1）准确的溶解氧预测对于水产养殖生产具有重要意义。本文将皮尔森相关系数法应用到确定环境因素与溶解氧的相关系数中，并将相似日的统计量-相似度作了改进，克服了欧式距离和夹角余弦计算法的局限性。

2）采用 *K*-means 聚类对溶解氧样本数据进行分类，找出合适样本，降低了不同趋势样本间的干扰，能够挖掘出溶解氧数据的固有规律，提高预测数据源的准确性。

3）ELM 预测网络与其他传统的 BP 神经网络、SVM 相比，预测精度尚可，其在运算速度上 ELM 神经网络只需要 0.1 s，而 BP 神经网络需 10.25 s，SVM 更慢，可见 ELM 预测网络具有很大优势。

4）与未经过 *K*-means 聚类的 ELM 预测网络相比，经过相似日聚类的预测模型，能够很好的反映溶解氧固有的规律，无论是常规天气还是突变天气，均能快速找到最优样本集，模型具有较好的鲁棒性。在常规天气下，平均绝对百分误差（MAPE）和均方根误差（RMSE）分别达到 1.4%、10.8%；在突变天气下，MAPE 和 RMSE 分别达到 2.6%和 11.6%。拟合效果比较理想，一定程度上提高了预测精度。

由于天气突变影响光照和气压，从而导致水生物降低产氧、好氧菌降低溶氧、养殖动物增加耗氧，影响鱼体健康。水产管理者应时刻关注天气变化，积极做好养殖生产管理。未来进一步研究工作有：

1）可以定购突变天气气象预报，在收到突变天气预报后，在原来 *K*-means 聚类和 ELM 组合预测模型的基础上，加入改进的修正算法，从而达到提高突变天气情况下溶解氧预测精度的效果。并集成智能控制系统，及时开动增氧机增氧曝气。

2）根据预测结果，给出针对天气突变前和突变时的具体应急预案，为管理者提供决策支持。如突变前，可通过加注新水、使用生石灰、减少喂食、开动增氧机等措施；突变时，可提高水位，开动增氧机，使用化学增氧剂、吸附剂和有益菌，使用抗应激药物拌料投喂，做到恒氧、恒温、防应激。

［参　考　文　献］

[1] 徐龙琴，李乾川，刘双印，等. 基于集合经验模态分解和人工蜂群算法的工厂化养殖pH值预测[J]. 农业工程学报，2016，32(3)：202－208.

Xu Longqin, Li Qianchuan, Liu Shuangyin, et al. Prediction of pH value in industrialized aquaculture based on ensemble empirical mode decomposition and improved artificial bee colony algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2016, 32(3): 202－208. (in Chinese with English abstract)

[2] 金光炎. 水文统计理论与实践[M]. 南京：东南大学出版社，2012.

[3] 胡金有，王靖杰，张小栓，等. 水产养殖信息化关键技术研究现状与趋势[J]. 农业机械学报，2015，46(7)：251－261.

Hu Jinyou, Wang Jingjie, Zhang Xiaoshuan, et al. Research status and development trends of information Technologies in aquacultures[J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(7): 251－261. (in Chinese with English abstract)

[4] 刘东君，邹志红. 灰色和神经网络组合模型在水质预测中的应用[J]. 系统工程，2011，29(9)：105－109.

Liu Dongjun, Zou Zhihong. Applications of gray forecast model combined with artificial neural networks model to water quality forecast[J]. Systems Engineering, 2011, 29(9): 105－109. (in Chinese with English abstract)

[5] 岳遥，李天宏. 基于模糊集理论的马尔可夫模型在水质定量预测中的应用[J]. 应用基础与工程科学学报，2011，19(2)：231－242.

Yue Yao, Li Tianhong. The application of a fuzzy-set-theory

based markov model in the quantitative prediction of water quality[J]. Journal of Basic Science and Engineering, 2011, 19(2): 231－242. (in Chinese with English abstract)

[6] 黄廷林，韩晓刚，卢金锁. 基于 Lyapunov 指数的混沌预测方法及在水质预测中的应用[J]. 西安建筑科技大学学报：自然科学版，2008，40(6)：546－581.
Huang Tinglin, Han Xiaogang , Lu Jinsuo. Chaos predication method based on Lyapunov exponent and its application in water quality forecast[J]. Journal of Xi'an University of Architecture & Technology: Natural Science Edition, 2008, 40(6): 546－581. (in Chinese with English abstract)

[7] 石子泊，邹志红. 基于小波变换的 ARIMA 模型在水质预测中的应用研究[J]. 环境工程学报，2014，8(10)：4550－4554.
Shi Zibo, Zou Zhihong. Applied study of ARIMA model base d on wavelet analysis on water quality prediction[J]. Chinese Journal of Environmental Engineering, 2014, 8(10): 4550－4554. (in Chinese with English abstract)

[8] 张颖，高倩倩. 基于灰色模型和模糊神经网络的综合水质预测模型研究[J]. 环境工程学报，2015，9(2)：537－545.
Zhang Ying, Gao Qianqian. Comprehensive predict ion model of water quality based on grey model and fuzzy neural network[J]. Chinese Journal of Environmental Engineering, 2015, 9(2): 537－545. (in Chinese with English abstract)

[9] 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. 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)

[10] 刘双印，徐龙琴，李振波，等. 基于 PCA-MCAFA-LSSVM 的养殖水质 pH 值预测模型[J]. 农业机械学报，2014，45(5)：239－246.
Liu Shuangyin, Xu Longqin, Li Zhenbo, et al. Forecasting model for pH value of aquaculture water quality based on PCA-MCAFA-LSSVM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2014, 45(5): 239－246. (in Chinese with English abstract)

[11] 刘明，李由明，王平，等. 基于小波分解的凡纳滨对虾养殖水体水质的仿真研究[J]. 广东农业科学，2015，9(2)：537－545.
Liu Ming, Li Youming, Wang Ping. Phantom study on water quality dynamic in Litopenaeus vannameiis's ponds based on wavelet analysis[J]. Guangdong Agricultural Sciences, 2015, 9(2): 537－545. (in Chinese with English abstract)

[12] Huan Juan, Liu Xingqiao, Cheng Qinfeng. Design of an aquaculture monitoring system based on android and GPRS[J]. Applied Engineering in Agriculture. 2014, 30(4): 681－687.

[13] 马从国，赵德安，王建国，等. 基于无线传感器网络的水产养殖池塘溶解氧智能监控系统[J]. 农业工程学报，2015，31(7)：193－200.
Ma Congguo, Zhao Dean, Wang Jianguo, et al. Intelligent monitoring system for aquaculture dissolved oxygen in pond based on wireless sensor network[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2015, 31(7): 193－200. (in Chinese with English abstract)

[14] 白俊良，梅华威. 改进相似度的模糊聚类算法在光伏阵列短期功率预测中的应用[J]. 电力系统保护与控制，2014，42(6)：84－90.
Bai Junliang, Mei Huawei. Improved similarity based fuzzy clustering algorithm and its application in the PV array power short-term forecasting[J]. Power System Protection and Control, 2014, 42(6): 84－90. (in Chinese with English abstract)

[15] Huang Guangbin, Chen Lei, Siew Chee-Kheong. Universal approximation using incremental constructive feedforward networks with random hidden nodes[J]. IEEE Transactions on Neural Networks, 2006, 17(4): 879－892.

[16] 邵庆言. ELM 网络结构选择研究[D]. 保定：河北大学，2013.
Shao Qingyan. Research on Architecture Selection of ELM Networks[D]. Baoding: Hebei University, 2013. (in Chinese with English abstract)

[17] Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques[J]. Journal of Intelligent Information Systems. 2001, 17: 107－145.

[18] 王守相，王亚旻，刘岩，等. 基于经验模态分解和 ELM 神经网络的逐时太阳能辐照量预测[J]. 电力自动化设备，2014，34(8)：7－12.
Wang Shouxiang, Wang Yamin, Liu Yan. Hourly solar radiation forecasting based on EMD and ELM neural network[J]. Electric Power Automation Equipment, 2014, 34(8): 7－12. (in Chinese with English abstract)

[19] 张娜. 间歇性能源输出功率预测与储能系统规划[D]. 天津：天津大学，2013.
Zhang Na. Intermittent Energy Output Power Forecasting and Storage System Planning[D]. Tianjin: Tianjin University, 2013. (in Chinese with English abstract)

Dissolved oxygen prediction in water based on *K*-means clustering and ELM neural network for aquaculture

Huan Juan[1,2], Liu Xingqiao[1※]

(1. *School of Electrical and Information Engineering, Jiangsu University, Zhenjiang* 212013, *China*;
2. *School of Information Science and Engineering, Changzhou University*, *Changzhou* 213164, *China*)

Abstract: Dissolved oxygen plays a vital role in water management as it is an important factor that determines the growth status of the fish. Either inadequate or excessive level of dissolved oxygen will be harmful to the survivability of the fish in their respective habitats. The accurate analysis of the data collected from the aquaculture ponds and the prediction for the anticipated level of dissolved oxygen are helpful for both water quality management and aquaculture production. Current studies reveal and understand the complex features of the water quality process mainly from the perspective of mathematical statistics. However, they cannot analyze the effects of changes in the environment on water quality, and cannot do well in dissolved oxygen prediction under the changing environment either. This paper proposed a new strategy to predict dissolved oxygen based on K-means clustering and ELM (extreme learning machine) neural networks. As the curves of similar days showed high correlation of dissolved oxygen, the history samples were divided into several classes to optimize sample space and improve prediction accuracy. After data normalization, the weights of the environmental factors on the dissolved oxygen were determined by Pearson correlation coefficient. The similarity statistics of similar days were improved and defined, which overcame the limitation of Euclidean distance and cosine calculation method. According to the similarity statistics, K-means clustering method was employed to divide the historical samples into several clusters with different daily samples. When the most similar cluster to the forecasting day was identified, the way could reduce the interference between samples and mine the inherent law of the dissolved oxygen data. Then, the ELM neural network of the identified cluster was constructed with the training samples and test data set, and the future amount of dissolved oxygen was predicted with the similar sample set and the real-time environmental factors of the forecasting day as the input data. A total of 23 424 data records of the aquaculture ponds in Wujin, Changzhou, China, were collected and used in the experiments. Taking 5 clusters as the example, ELM neural network was compared with other traditional BP (back propagation) neural networks and SVM (support vector machine). Its prediction accuracy was acceptable, and the running time was only 0.1 s, while that of BP neural network was 10.25 s and that of SVM was slower. It is visible ELM prediction network has a great advantage. Additionally, the caculation speed and prediction efficiency of the model are better than others in terms of the root mean square error and the mean absolute percentage error. Experiment results showed that MAPE and RMSE of our prediction method reached 1.4% and 10.8% respectively under normal climate condition. In case of a sudden change of weather, the MAPE and RMSE were 2.6% and 11.6%, respectively. It has higher forecasting accuracy and faster computation speed, which is beneficial to water quality control in aquaculture.

Keywords: neural networks; models; aquaculture; dissolved oxygen prediction; similar day; *K*-means clustering; ELM neural network