(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115456245 A
(43)申请公布日 2022.12.09

(21)申请号 202210967488.X

(22)申请日 2022.08.12

(71)申请人 生态环境部华南环境科学研究所(生态环境部生态环境应急研究所)
地址 510535 广东省广州市黄埔区瑞和路18号

(72)发明人 赵长进 叶颖欣 范中亚 杨汉杰 王文才 房怀阳 罗千里 曾凡棠 胡艳芳

(74)专利代理机构 北京栈桥知识产权代理事务所(普通合伙) 11670
专利代理师 张建生

(51)Int.Cl.
G06Q 10/04(2012.01)
G06Q 50/02(2012.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)
G06K 9/62(2022.01)

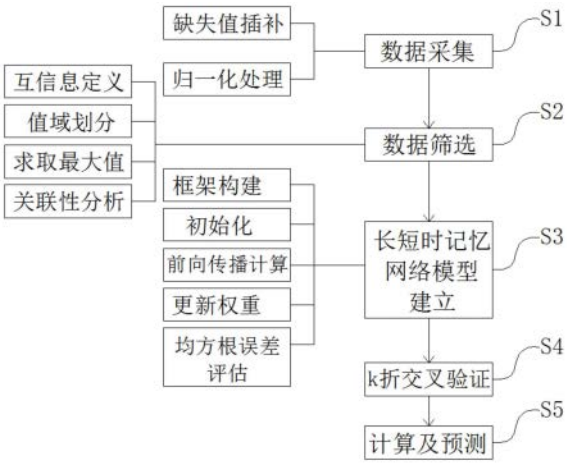
权利要求书4页 说明书10页 附图2页

(54)发明名称

一种感潮河网区溶解氧预测方法

(57)摘要

本发明公开了一种感潮河网区溶解氧预测方法，包括以下步骤：S1、数据采集；S2、数据筛选：S2-1、互信息定义；S2-2、值域划分；S2-3、求取最大值；S2-4、关联性分析；S3、长短时记忆网络模型建立：S3-1、框架构建；S3-2、初始化；S3-3、前向传播计算；S3-4、更新权重；S3-5、均方根误差评估；S4、k折交叉验证；S5、计算及预测。本发明充分考虑了感潮河网区受潮汐影响、溶解氧呈现周期性变化的特点，选取时间滞后的溶解氧数据作为输入变量，并通过最大互信息系数方法识别出影响溶解氧变化的关键因素作为输入变量，使用深度机器学习模型建立长短时记忆网络有效解决了传统循环网络中的梯度消失的问题。



CN 115456245 A

1.一种感潮河网区溶解氧预测方法,其特征在于,包括以下步骤:

S1、数据采集:在需要进行溶解氧预测的感潮河网区建立水质自动站,通过水质自动站采集水质时间序列数据,并对采集到的水质时间序列数据进行预处理,水质时间序列数据包括溶解氧和其他环境变量;

S2、数据筛选:计算步骤S1得到的水质时间序列数据中溶解氧和其他环境变量的最大互信息系数,筛选出与溶解氧相关性较大的其他环境变量,作为长短时记忆网络的输入变量;

S2-1、互信息定义:互信息是衡量其他环境变量与溶解氧之间相关程度的指标,给定变量A＝{x_i,i＝1,2,...,n}和B＝{y_i,i＝1,2,...,n},其中,n为样本数量,A和B的互信息I(A;B)定义公式为:

$$I(A;B)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

式中,p(x,y)为A和B的联合概率密度,p(x)为A的边缘概率密度,p(y)为B的边缘概率密度;

S2-2、值域划分:假设D＝{(a_i,b),i＝1,2,...,n}为一个有限集合,同时,将变量A和变量B的值域分别划分为x段和y段,得到x×y的网格G,再在得到的每一种网格划分的内部计算互信息MI(A,B),得到互信息MI(A,B)的最大值G,则定义最大值G条件下有限集合D的最大归一化值公式为:

MI*(D,x,y)＝maxMI(D|G)

式中,D|G为有限集合D在使用网格G进行划分,MI*(D,x,y)为最大归一化值;

S2-3、求取最大值:对每一种网格划分下得到的最大归一化值组成的特征矩阵求取最大值,即得最大信息系数的公式为:

$$MIC(D)=\max_{x\times y<B(n)}\left\{\frac{MI*(D,x,y)}{\log\min\{x,y\}}\right\}$$

式中,MIC(D)为最大信息系数;

S2-4、关联性分析:将溶解氧作为变量A,其他环境变量作为变量B,计算溶解氧与其他环境变量的最大信息系数MIC(D)的值,得到的最大信息系数MIC(D)的值在[0,1]区间内,最大信息系数MIC(D)的值越大,则溶解氧与其他环境变量的关联性越大,最大信息系数MIC(D)的值越小,则溶解氧与其他环境变量的关联性越小,选择与溶解氧关联性较大的其他环境变量作为预测模型的输入变量;

S3、长短时记忆网络模型建立:

S3-1、框架构建:所述长短时记忆网络模型包含1个输入层、1个输出层及多个隐藏层,每个隐藏层由多个记忆单元组成,所述记忆单元通过引入门控机制来控制历史信息的更新和利用,所述门控机制包括输入门i_t、遗忘门i_tf_t和输出门o_t,输入门i_t、遗忘门f_t和输出门o_t的取值均在[0,1]区间内表示以一定的比例让信息通过,对细胞状态定期重置避免细胞状态不短累加,细胞状态包括候选状态$\tilde{C}_t$、内部状态C_t和外部状态h_t,输入门i_t控制当前时刻的候选状态$\tilde{C}_t$有多少信息需要保存,遗忘门f_t控制上一个时刻的内部状态C_t需要遗忘多

少信息，输出门o_t控制当前时刻的内部状态C_t有多少信息需要输出给外部状态h_t，同时激活函数sigmoid(σ)和双曲正切函数层tanh，如下式所示：

$$\sigma(x)=\frac{1}{1+e^{-x}}$$

$$\tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$$

S3-2、初始化：将记忆单元的矩阵和向量进行初始化，用于保存模型参数和保存中间计算结果，保存输入层和输出层神经元个数、隐含层细胞个数、网络状态；

S3-3、前向传播计算：长短时记忆网络模型会决定从细胞状态中舍弃的信息，这个步骤由遗忘门完成，首先，针对当前时刻的输入信息x_t和上一时刻的隐藏层外部状态h_{t-1}的输出信息通过sigmoid(σ)函数层处理得到一个0到1之间的输出，作为上一时刻内部状态C_{t-1}的过滤值，即得遗忘门f_t的公式为：

$$f_t=\sigma(W_{xf}x_t+W_{hf}h_{t-1}+b_f)$$

式中，W为权重矩阵，W的下标代表具体两个单元之间的连接权重，b代表偏置项；

其次，长短时记忆网络模型判定存储到细胞状态中的信息，首先将当前时刻的输入信息x_t和上一时刻的隐藏层外部状态h_{t-1}的输出信息经过sigmoid函数层计算得到输入门i_t取值，如下式所示：

$$i_t=\sigma(W_{xi}x_t+W_{hi}h_{t-1}+b_i)$$

然后通过双曲正切函数层tanh产生一个候选状态$\tilde{C}_t$用于细胞状态的更新，如下式所示：

$$\tilde{C}_t=\tanh(W_{xC}x_t+W_{hC}h_{t-1}+b_c)$$

最后，长短时记忆网络模型决定细胞的输出信息，将当前时刻的输入信息x_t和上一时刻的隐藏层外部状态h_{t-1}的输出信息经过sigmoid(σ)函数层计算输出门o_t，如下式所示：

$$o_t=\sigma(W_{xo}x_t+W_{ho}h_{t-1}+b_o)$$

然后将当前细胞的内部状态C_t通过tanh函数压缩至[-1,1]的区间，最后将压缩后的细胞的内部状态C_t与输出门o_t相乘得到当前时刻的隐藏层外部状态h_t输出信息，如下式所示：

$$h_t=o_t\tanh(C_t)$$

记忆单元还会与长短时记忆网络模型中其他部分相连，当前时刻的隐藏层外部状态h_t的输出信息一方面作为隐藏层外部状态h_t的输入信息被传递到下个时刻，另一方面作为隐藏层外部状态h_t的输出信息被传递到下一层长短时记忆网络，当下一层长短时记忆网络为全连接层时，会对隐藏层结果做一个变换得到最终输出信息，从而得到时间序列的预测值$\hat{y}_t$，如下式所示：

$$\hat{y}_t=V_{out}h_t+b_y$$

式中，Vout为全连接层的权重矩阵，b代表偏置项；

S3-4、更新权重：求解长短时记忆网络的每一个权重的梯度，通过使用训练数据进行随机梯度下降找到最优解，由输出层往输入层的权重开始求梯度，依次更新各个权重，重置内部状态，设计误差函数，计算并检查梯度；

S3-5、均方根误差评估：通过长短时记忆网络模型对与溶解氧相关的其他环境变量的时序数据进行训练，将归一化及经过MIC筛选的其他环境变量的时序数据作为训练数据集对长短时记忆网络模型进行训练，为了缓解在多变量预测模型神经网络的训练过程中出现的过拟合问题，在隐藏层的训练机制中加入Dropout机制，训练完成后计算均方根误差来评估长短时记忆网络模型的预测结果，均方根误差如下式所示：

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}[\hat{y}(i)-y(i)]^2}\quad(i=1,\ 2,\ \ldots,\ n)$$

式中，$\hat{y}(i)$为溶解氧的预测值，y(i)为溶解氧的实测值；

S4、k折交叉验证：将步骤S2-4中得到的输入变量作为原始数据集分成k等份，每次选取k-1份作为训练集，剩下1份作为测试集，使用不同的超参数组合来训练k-1份和测试其余的1个部分，并计算测试集的RMSE值，重复上述步骤S3-2～S3-5中长短时记忆网络模型训练和测试的步骤，直到k份原始数据集中每个超参数组合都被测试完毕，并计算每个最终输出信息的RMSE平均值，RMSE平均值最小的参数组合为最优组合，如下式所示：

$$RMSE_{(k)}=\frac{1}{k}\sum_{i=1}^{k}RMSE_i$$

S5、计算及预测：使用感潮河网区水质自动站的实时数据经预处理后输入到建立好的长短时记忆网络模型中，通过长短时记忆网络模型输出的结果通过放缩得到溶解氧的预测值，采用滚动预报的方法，绘制出溶解氧的走势图。

2.根据权利要求1所述的一种感潮河网区溶解氧预测方法，其特征在于，步骤S1中水质时间序列数据的其他环境变量包括pH、水温、电导率、浊度、水位、流量、氨氮、总磷、高锰酸盐指数、化学需氧量、总氮以及DO_{25h}，所述DO_{25h}为修正后的溶解氧时间序列数据，DO_{25h}修正方法为：一个潮汐周期的时长为24h50min，增加滞后时间至25h，此时得到的修正后的溶解氧时间序列数据即为DO_{25h}。

3.根据权利要求1所述的一种感潮河网区溶解氧预测方法，其特征在于，所述步骤S1中预处理的方法为：对采集到的水质时间序列数据进行缺失值插补以及归一化处理；

S1-1、缺失值插补：当水质时间序列数据出现缺失时使用相邻两个时刻数据的平均值插补；

S1-2、归一化处理：归一化处理的公式为：

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

式中，x'为归一化处理后的水质时间序列数据，x为归一化处理前的水质时间序列数据，x_{min}为水质时间序列数据中的最小值，x_{max}为水质时间序列数据中的最大值。

4.根据权利要求1所述的一种感潮河网区溶解氧预测方法，其特征在于，所述步骤S2-4中最大信息系数MIC(D)的值大于0.8时认为其他环境变量与溶解氧关联性较大。

5.根据权利要求1所述的一种感潮河网区溶解氧预测方法，其特征在于，所述步骤S3-2中模型参数包括权重矩阵W和偏置项b，中间计算结果包括外部状态h_t的输出信息、输入门f_t、遗忘门i_t、输出门o_t。

6.根据权利要求1所述的一种感潮河网区溶解氧预测方法,其特征在于,所述步骤S3-5中Dropout机制为:在其他环境变量的时序数据进行训练的过程中随机丢失神经单元及其连接。

7.根据权利要求1所述的一种感潮河网区溶解氧预测方法,其特征在于,所述步骤S3-1中搭建长短时记忆网络模型是基于TensorFlow深度学习框架搭建。

8.根据权利要求1所述的一种感潮河网区溶解氧预测方法,其特征在于,所述步骤S5中滚动预报的方法具体为:依据已有溶解氧的预测值的采样间隔,设置合理的预测时间步长,假设预测的时间为n日,长短时记忆网络模型会根据测试集中的t日溶解氧数据及S2所述方法筛选出的重要参数,对t+n日的溶解氧数据进行计算并输出得到溶解氧真实值,然后在t+2n日,运用t+n日的溶解氧真实值以及S2所述方法筛选出其他环境变量,采用滚动预报的方法及时更新序列信息。

一种感潮河网区溶解氧预测方法

技术领域

[0001] 本发明涉及水质预测技术领域,具体是涉及一种感潮河网区溶解氧预测方法。

背景技术

[0002] 溶解氧是水环境的一个关键度量指标,通常被用于评估水生生态系统的健康状况,水体缺氧会极大地影响水生生物的代谢、遗传和繁殖。感潮河网区受到径流与潮汐双重影响,动力条件复杂,温度、盐度、水体层化等因素均会影响水体复氧,导致感潮河网区时常出现低氧现象(溶解氧浓度≤3mg/L)。预测感潮河网区溶解氧浓度变化有利于对水环境突发的低氧事件进行预警预报和风险优化控制,提升对感潮河网区的水质风险防控和决策支持能力。

[0003] 溶解氧预测模型主要分为过程驱动模型和数据驱动模型。过程驱动模型基于物理定律,能够捕捉到水体动力学和营养组分循环的非线性相互作用以及水体中的化学和生物过程,充分模拟水污染过程的机理,但建模过程对环境数据的需求量和依赖性很大,求解过程复杂,需要大量的计算成本,当数据缺失或环境发生变化就难以模拟水污染过程。数据驱动模型有别于过程驱动模型,不依赖于物理机制,能捕捉到目标变量与解释变量之间复杂的非线性关系并通过动态自适应地修正模型元素(如结构、算法和参数),可用于非线性和高度随机的预测,已广泛地应用于水文水环境领域的相关研究。经典的数据驱动模型时间序列预测模型要求数据具有一定的平稳性和线性相关性,不能处理非线性问题;支持向量机(SVM)、Boosting算法、最大熵方法(MaxEnt)等都属于浅层机器学习的范畴,体系结构中通常最多包含一到两层非线性特征转换,在解决许多简单或约束良好的问题时表现出有效性,但它们有限的建模和表达能力在处理更复杂的现实问题时会造成困难。

[0004] 长短时记忆网络(Long Short-Term Memory Network,LSTM),是深度学习机器模型的其中一种,通过在循环神经网络的基础上引入了输入门、遗忘门和输出门来实现信息的自动化保留与舍弃,能够在预测过程中实现过去、现在和将来信息之间的有效关联,并解决了传统循环网络中的梯度消失的问题,相较于传统的浅层学习网络具有更好的预测性能。在实际预测中,过量的输入变量会增加模型计算的复杂性和降低模型性能,此时识别并筛选出驱动溶解氧变化的重要因素作为预测模型的输入变量对于预测溶解氧具有重要意义,最大互信息系数(Maximal Information Coefficient,MIC)可以有效地捕获变量之间的线性与非线性关系,被广泛用于各种研究领域内的输入变量的筛选。感潮河网区的溶解氧变化具有很强的日周期性,每天相同时刻溶解氧变化有着相似的变化趋势,但是,目前还没有很好地将长短时记忆网络结合溶解氧变化的日周期性对溶解氧做出更加精确预测的方法。

发明内容

[0005] 针对上述存在的问题,本发明提供了一种感潮河网区溶解氧预测方法。

[0006] 本发明的技术方案是:

[0007] 一种感潮河网区溶解氧预测方法，包括以下步骤：

[0008] S1、数据采集：在需要进行溶解氧预测的感潮河网区建立水质自动站，通过水质自动站采集水质时间序列数据，并对采集到的水质时间序列数据进行预处理，水质时间序列数据包括溶解氧和其他环境变量；

[0009] S2、数据筛选：计算步骤S1得到的水质时间序列数据中溶解氧和其他环境变量的最大互信息系数，筛选出与溶解氧相关性较大的其他环境变量，作为长短时记忆网络的输入变量；

[0010] S2-1、互信息定义：互信息是衡量其他环境变量与溶解氧之间相关程度的指标，给定变量A＝{x_i，i＝1，2，...，n}和B＝{y_i，i＝1，2，...，n}，其中，n为样本数量，A和B的互信息I(A;B)定义公式为：

[0011] $$I(A;B)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

[0012] 式中，p(x,y)为A和B的联合概率密度，p(x)为A的边缘概率密度，p(y)为B的边缘概率密度；

[0013] S2-2、值域划分：假设D＝{(a_i,b)，i＝1，2，...，n}为一个有限集合，同时，将变量A和变量B的值域分别划分为x段和y段，得到x×y的网格G，再在得到的每一种网格划分的内部计算互信息MI(A,B)，得到互信息MI(A,B)的最大值G，则定义最大值G条件下有限集合D的最大归一化值公式为：

[0014] MI*(D,x,y)＝maxMI(D|G)

[0015] 式中，D|G为有限集合D在使用网格G进行划分，MI*(D,x,y)为最大归一化值；

[0016] S2-3、求取最大值：对每一种网格划分下得到的最大归一化值组成的特征矩阵求取最大值，即得最大信息系数的公式为：

[0017] $$MIC(D)=\max_{x\times y<B(n)}\left\{\frac{MI*(D,x,y)}{\log\min\{x,y\}}\right\}$$

[0018] 式中，MIC(D)为最大信息系数；

[0019] S2-4、关联性分析：将溶解氧作为变量A，其他环境变量作为变量B，计算溶解氧与其他环境变量的最大信息系数MIC(D)的值，得到的最大信息系数MIC(D)的值在[0,1]区间内，最大信息系数MIC(D)的值越大，则溶解氧与其他环境变量的关联性越大，最大信息系数MIC(D)的值越小，则溶解氧与其他环境变量的关联性越小，选择与溶解氧关联性较大的其他环境变量作为预测模型的输入变量；

[0020] S3、长短时记忆网络模型建立：

[0021] S3-1、框架构建：所述长短时记忆网络模型包含1个输入层、1个输出层及多个隐藏层，每个隐藏层由多个记忆单元组成，所述记忆单元通过引入门控机制来控制历史信息的更新和利用，所述门控机制包括输入门i_t、遗忘门i_tf_t和输出门o_t，输入门i_t、遗忘门f_t和输出门o_t的取值均在[0,1]区间内表示以一定的比例让信息通过，对细胞状态定期重置避免细胞状态不短累加，细胞状态包括候选状态$\tilde{C}_t$、内部状态C_t和外部状态h_t，输入门i_t控制当前时刻的候选状态$\tilde{C}_t$有多少信息需要保存，遗忘门f_t控制上一个时刻的内部状态C_t需要遗

忘多少信息，输出门o_t控制当前时刻的内部状态C_t有多少信息需要输出给外部状态h_t，同时激活函数sigmoid(σ)和双曲正切函数层tanh，如下式所示：

[0022] $$\sigma(x)=\frac{1}{1+e^{-x}}$$

[0023] $$\tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$$

[0024] S3-2、初始化：将记忆单元的矩阵和向量进行初始化，用于保存模型参数和保存中间计算结果，保存输入层和输出层神经元个数、隐含层细胞个数、网络状态；

[0025] S3-3、前向传播计算：长短时记忆网络模型会决定从细胞状态中舍弃的信息，这个步骤由遗忘门完成，首先，针对当前时刻的输入信息x_t和上一时刻的隐藏层外部状态h_{t-1}的输出信息通过sigmoid(σ)函数层处理得到一个0到1之间的输出，作为上一时刻内部状态C_{t-1}的过滤值，即得遗忘门f_t的公式为：

[0026] $$f_t=\sigma(W_{xf}x_t+W_{hf}h_{t-1}+b_f)$$

[0027] 式中，W为权重矩阵，W的下标代表具体两个单元之间的连接权重，b代表偏置项；

[0028] 其次，长短时记忆网络模型判定存储到细胞状态中的信息，首先将当前时刻的输入信息x_t和上一时刻的隐藏层外部状态h_{t-1}的输出信息经过sigmoid函数层计算得到输入门i_t取值，如下式所示：

[0029] $$i_t=\sigma(W_{xi}x_t+W_{hi}h_{t-1}+b_i)$$

[0030] 然后通过双曲正切函数层tanh产生一个候选状态$\tilde{C}_t$用于细胞状态的更新，如下式所示：

[0031] $$\tilde{C}_t=\tanh(W_{xC}x_t+W_{hC}h_{t-1}+b_c)$$

[0032] 最后，长短时记忆网络模型决定细胞的输出信息，将当前时刻的输入信息x_t和上一时刻的隐藏层外部状态h_{t-1}的输出信息经过sigmoid(σ)函数层计算输出门o_t，如下式所示：

[0033] $$o_t=\sigma(W_{xo}x_t+W_{ho}h_{t-1}+b_o)$$

[0034] 然后将当前细胞的内部状态C_t通过tanh函数压缩至[-1,1]的区间，最后将压缩后的细胞的内部状态C_t与输出门o_t相乘得到当前时刻的隐藏层外部状态h_t输出信息，如下式所示：

[0035] $$h_t=o_t\tanh(C_t)$$

[0036] 记忆单元还会与长短时记忆网络模型中其他部分相连，当前时刻的隐藏层外部状态h_t的输出信息一方面作为隐藏层外部状态h_t的输入信息被传递到下个时刻，另一方面作为隐藏层外部状态h_t的输出信息被传递到下一层长短时记忆网络，当下一层长短时记忆网络为全连接层时，会对隐藏层结果做一个变换得到最终输出信息，从而得到时间序列的预测值$\hat{y}_t$，如下式所示：

[0037] $$\hat{y}_t=V_{out}h_t+b_y$$

[0038] 式中，Vout为全连接层的权重矩阵，b代表偏置项；

[0039] S3-4、更新权重：求解长短时记忆网络的每一个权重的梯度，通过使用训练数据进

行随机梯度下降找到最优解，由输出层往输入层的权重开始求梯度，依次更新各个权重，重置内部状态，设计误差函数，计算并检查梯度；

[0040] S3-5、均方根误差评估：通过长短时记忆网络模型对与溶解氧相关的其他环境变量的时序数据进行训练，将归一化及经过MIC筛选的其他环境变量的时序数据作为训练数据集对长短时记忆网络模型进行训练，为了缓解在多变量预测模型神经网络的训练过程中出现的过拟合问题，在隐藏层的训练机制中加入Dropout机制，训练完成后计算均方根误差来评估长短时记忆网络模型的预测结果，均方根误差如下式所示：

[0041] $$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}[\hat{y}(i)-y(i)]^2}\quad(i=1,\ 2,\ \ldots,\ n)$$

[0042] 式中，$\hat{y}(i)$为溶解氧的预测值，y(i)为溶解氧的实测值；

[0043] S4、k折交叉验证：将步骤S2-4中得到的输入变量作为原始数据集分成k等份，每次选取k-1份作为训练集，剩下1份作为测试集，使用不同的超参数组合来训练k-1份和测试其余的1个部分，并计算测试集的RMSE值，重复上述步骤S3-2～S3-5中长短时记忆网络模型训练和测试的步骤，直到k份原始数据集中每个超参数组合都被测试完毕，并计算每个最终输出信息的RMSE平均值，RMSE平均值最小的参数组合为最优组合，如下式所示：

[0044] $$RMSE_{(k)}=\frac{1}{k}\sum_{i=1}^{k}RMSE_i$$

[0045] S5、计算及预测：使用感潮河网区水质自动站的实时数据经预处理后输入到建立好的长短时记忆网络模型中，通过长短时记忆网络模型输出的结果通过放缩得到溶解氧的预测值，采用滚动预报的方法，绘制出溶解氧的走势图。

[0046] 进一步地，步骤S1中水质时间序列数据的其他环境变量包括pH、水温、电导率、浊度、水位、流量、氨氮、总磷、高锰酸盐指数、化学需氧量、总氮以及DO_{25h}，所述DO_{25h}为修正后的溶解氧时间序列数据，DO_{25h}修正方法为：一个潮汐周期的时长为24h50min，增加滞后时间至25h，此时得到的修正后的溶解氧时间序列数据即为DO_{25h}。

[0047] 进一步地，所述步骤S1中预处理的方法为：对采集到的水质时间序列数据进行缺失值插补以及归一化处理；

[0048] S1-1、缺失值插补：当水质时间序列数据出现缺失时使用相邻两个时刻数据的平均值插补；

[0049] S1-2、归一化处理：归一化处理的公式为：

[0050] $$x'=\frac{x-x_{min}}{x_{max}-x_{min}}$$

[0051] 式中，x'为归一化处理后的水质时间序列数据，x为归一化处理前的水质时间序列数据，x_{min}为水质时间序列数据中的最小值，x_{max}为水质时间序列数据中的最大值。

[0052] 进一步地，所述步骤S2-4中最大信息系数MIC(D)的值大于0.8时认为其他环境变量与溶解氧关联性较大。通常溶解氧与DO_{25}的MIC(D)值较大，约为0.7-0.8左右，关联性的计算由python模块sklearn.metrics.cluster中的normalized_mutual_info_score完成。

[0053] 进一步地，所述步骤S3-2中模型参数包括权重矩阵W和偏置项b，中间计算结果包

括外部状态h_t的输出信息、输入门f_t、遗忘门i_t、输出门o_t。

[0054] 进一步地，所述步骤S3-5中Dropout机制为：在其他环境变量的时序数据进行训练的过程中随机丢失神经单元及其连接。

[0055] 进一步地，所述步骤S3-1中搭建长短时记忆网络模型是基于TensorFlow深度学习框架搭建。

[0056] 进一步地，所述步骤S5中滚动预报的方法具体为：依据已有溶解氧的预测值的采样间隔，设置合理的预测时间步长，假设预测的时间为n日，长短时记忆网络模型会根据测试集中的t日溶解氧数据及S2所述方法筛选出的重要参数，对t+n日的溶解氧数据进行计算并输出得到溶解氧真实值，然后在t+2n日，运用t+n日的溶解氧真实值以及S2所述方法筛选出其他环境变量，采用滚动预报的方法及时更新序列信息，避免误差积累。

[0057] 本发明的有益效果是：

[0058] 本发明的感潮河网区溶解氧预测方法提供了针对感潮河网区溶解氧预测的解决方案，充分考虑了感潮河网区受潮汐影响、溶解氧呈现周期性变化的特点，选取时间滞后的溶解氧数据作为输入变量，并通过最大互信息系数(Maximal Information Coefficient，MIC)方法识别出影响溶解氧变化的关键因素作为输入变量，使用深度机器学习模型建立长短时记忆网络(Long Short-Term Memory Network，LSTM)有效解决了传统循环网络中的梯度消失的问题，并使用K折交叉验证网格搜索的方法选择模型最优超参数组合，提升了感潮河网区溶解氧预测的精度。

附图说明

[0059] 图1是本发明的感潮河网区溶解氧预测方法流程图；

[0060] 图2是本发明的感潮河网区溶解氧预测方法实验例中步骤S3的示意图；

[0061] 图3是本发明的感潮河网区溶解氧预测方法实验例1中长短时记忆网络模型的测试及训练结果示意图；

[0062] 图4是本发明的感潮河网区溶解氧预测方法实验例2中长短时记忆网络模型的测试及训练结果示意图；

[0063] 图5是本发明的感潮河网区溶解氧预测方法实验例3中长短时记忆网络模型的测试及训练结果示意图。

具体实施方式

[0064] 实施例1

[0065] 一种感潮河网区溶解氧预测方法，如图1所示，包括以下步骤：

[0066] S1、数据采集：在需要进行溶解氧预测的感潮河网区建立水质自动站，通过水质自动站采集水质时间序列数据，并对采集到的水质时间序列数据进行预处理，水质时间序列数据包括溶解氧和其他环境变量，水质时间序列数据的其他环境变量包括pH、水温、电导率、浊度、水位、流量、氨氮、总磷、高锰酸盐指数、化学需氧量、总氮以及DO_{25h}，所述DO_{25h}为修正后的溶解氧时间序列数据，DO_{25h}修正方法为：一个潮汐周期的时长为24h50min，增加滞后时间至25h，此时得到的修正后的溶解氧时间序列数据即为DO_{25h}；

[0067] 预处理的方法为：对采集到的水质时间序列数据进行缺失值插补以及归一化处

理；

[0068] S1-1、缺失值插补：当水质时间序列数据出现缺失时使用相邻两个时刻数据的平均值插补；

[0069] 对数据的异常值(在数据二次表格中用L或者多个000标识)和缺省值进行识别，标记为nan。当水质时间序列数据出现缺失时使用相邻两个时刻数据的平均值插补；

[0070] 并对采样频率的统一化，在水质自动站的数据记录中，会出现非整点或者整天记录的情况，对此类情况进行甄别，依照各个站点的实际情况，统一到整天或者整小时；

[0071] 缺失值插补：依照统一化的各站点数据采样频率，在对应时间点无有效数据的情况下，利用最近的有效数据进行填补，如果缺失数据大于12个时间步长，则运用线性插值进行插补。

[0072] S1-2、归一化处理：归一化处理的公式为：

[0073] $$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

[0074] 式中，x’为归一化处理后的水质时间序列数据，x为归一化处理前的水质时间序列数据，x_{min}为水质时间序列数据中的最小值，x_{max}为水质时间序列数据中的最大值；

[0075] S2、数据筛选：计算步骤S1得到的水质时间序列数据中溶解氧和其他环境变量的最大互信息系数，筛选出与溶解氧相关性较大的其他环境变量，作为长短时记忆网络的输入变量；

[0076] S2-1、互信息定义：互信息是衡量其他环境变量与溶解氧之间相关程度的指标，给定变量A＝{x_i，i＝1，2，...，n}和B＝{y_i，i＝1，2，...，n}，其中，n为样本数量，A和B的互信息I(A;B)定义公式为：

[0077] $$I(A;B) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)$$

[0078] 式中，p(x,y)为A和B的联合概率密度，p(x)为A的边缘概率密度，p(y)为B的边缘概率密度；

[0079] S2-2、值域划分：假设D＝{(a_i，b)，i＝1，2，...，n}为一个有限集合，同时，将变量A和变量B的值域分别划分为x段和y段，得到x×y的网格G，再在得到的每一种网格划分的内部计算互信息MI(A,B)，得到互信息MI(A,B)的最大值G，则定义最大值G条件下有限集合D的最大归一化值公式为：

[0080] MI*(D,x,y)＝maxMI(D|G)

[0081] 式中，D|G为有限集合D在使用网格G进行划分，MI*(D,x,y)为最大归一化值；

[0082] S2-3、求取最大值：对每一种网格划分下得到的最大归一化值组成的特征矩阵求取最大值，即得最大信息系数的公式为：

[0083] $$MIC(D) = \max_{x \times y < B(n)} \left\{\frac{MI*(D,x,y)}{\log \min\{x,y\}}\right\}$$

[0084] 式中，MIC(D)为最大信息系数；

[0085] S2-4、关联性分析：将溶解氧作为变量A，其他环境变量作为变量B，计算溶解氧与其他环境变量的最大信息系数MIC(D)的值，得到的最大信息系数MIC(D)的值在[0,1]区间

内，最大信息系数MIC(D)的值越大，则溶解氧与其他环境变量的关联性越大，最大信息系数MIC(D)的值越小，则溶解氧与其他环境变量的关联性越小，选择与溶解氧关联性较大的其他环境变量作为预测模型的输入变量，最大信息系数MIC(D)的值大于0.8时认为其他环境变量与溶解氧关联性较大；

[0086] S3、长短时记忆网络模型建立：

[0087] S3-1、框架构建：基于TensorFlow深度学习框架搭建长短时记忆网络模型，所述长短时记忆网络模型包含1个输入层、1个输出层及3个隐藏层，每个隐藏层由20个记忆单元组成，所述记忆单元通过引入门控机制来控制历史信息的更新和利用，所述门控机制包括输入门i_t、遗忘门i_tf_t和输出门o_t，输入门i_t、遗忘门f_t和输出门o_t的取值均在[0,1]区间内表示以一定的比例让信息通过，对细胞状态定期重置避免细胞状态不短累加，细胞状态包括候选状态$\tilde{C}_t$、内部状态C_t和外部状态h_t，输入门i_t控制当前时刻的候选状态$\tilde{C}_t$有多少信息需要保存，遗忘门f_t控制上一个时刻的内部状态C_t需要遗忘多少信息，输出门o_t控制当前时刻的内部状态C_t有多少信息需要输出给外部状态h_t，同时激活函数sigmoid(σ)和双曲正切函数层tanh，如下式所示：

[0088] $$\sigma(x)=\frac{1}{1+e^{-x}}$$

[0089] $$\tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$$

[0090] S3-2、初始化：将记忆单元的矩阵和向量进行初始化，用于保存模型参数和保存中间计算结果，模型参数包括权重矩阵W和偏置项b，中间计算结果包括外部状态h_t的输出信息、输入门f_t、遗忘门i_t、输出门o_t，保存输入层和输出层神经元个数、隐含层细胞个数、网络状态；

[0091] S3-3、前向传播计算：长短时记忆网络模型会决定从细胞状态中舍弃的信息，这个步骤由遗忘门完成，首先，针对当前时刻的输入信息x_t和上一时刻的隐藏层外部状态h_{t-1}的输出信息通过sigmoid(σ)函数层处理得到一个0到1之间的输出，作为上一时刻内部状态C_{t-1}的过滤值，即得遗忘门f_t的公式为：

[0092] $f_t=\sigma(W_{xf}x_t+W_{hf}h_{t-1}+b_f)$

[0093] 式中，W为权重矩阵，W的下标代表具体两个单元之间的连接权重，b代表偏置项；

[0094] 其次，长短时记忆网络模型判定存储到细胞状态中的信息，首先将当前时刻的输入信息x_t和上一时刻的隐藏层外部状态h_{t-1}的输出信息经过sigmoid函数层计算得到输入门i_t取值，如下式所示：

[0095] $i_t=\sigma(W_{xi}x_t+W_{hi}h_{t-1}+b_i)$

[0096] 然后通过双曲正切函数层tanh产生一个候选状态$\tilde{C}_t$用于细胞状态的更新，如下式所示：

[0097] $$\tilde{C}_t=\tanh(W_{xC}x_t+W_{hC}h_{t-1}+b_c)$$

[0098] 最后，长短时记忆网络模型决定细胞的输出信息，将当前时刻的输入信息x_t和上一时刻的隐藏层外部状态h_{t-1}的输出信息经过sigmoid(σ)函数层计算输出门o_t，如下式所

示：

[0099] $o_t=\sigma(W_{xo}x_t+W_{ho}h_{t-1}+b_o)$

[0100] 然后将当前细胞的内部状态C_t通过tanh函数压缩至[-1,1]的区间，最后将压缩后的细胞的内部状态C_t与输出门o_t相乘得到当前时刻的隐藏层外部状态h_t输出信息，如下式所示：

[0101] $h_t=o_t\tanh(C_t)$

[0102] 记忆单元还会与长短时记忆网络模型中其他部分相连，当前时刻的隐藏层外部状态h_t的输出信息一方面作为隐藏层外部状态h_t的输入信息被传递到下个时刻，另一方面作为隐藏层外部状态h_t的输出信息被传递到下一层长短时记忆网络，当下一层长短时记忆网络为全连接层时，会对隐藏层结果做一个变换得到最终输出信息，从而得到时间序列的预测值$\hat{y}_t$，如下式所示：

[0103] $$\hat{y}_t=V_{out}h_t+b_y$$

[0104] 式中，Vout为全连接层的权重矩阵，b代表偏置项；

[0105] S3-4、更新权重：求解长短时记忆网络的每一个权重的梯度，通过使用训练数据进行随机梯度下降找到最优解，由输出层往输入层的权重开始求梯度，依次更新各个权重，重置内部状态，设计误差函数，计算并检查梯度；

[0106] S3-5、均方根误差评估：通过长短时记忆网络模型对与溶解氧相关的其他环境变量的时序数据进行训练，将归一化及经过MIC筛选的其他环境变量的时序数据作为训练数据集对长短时记忆网络模型进行训练，为了缓解在多变量预测模型神经网络的训练过程中出现的过拟合问题，在隐藏层的训练机制中加入Dropout机制，Dropout机制为：在其他环境变量的时序数据进行训练的过程中随机丢失神经单元及其连接，训练完成后计算均方根误差来评估长短时记忆网络模型的预测结果，均方根误差如下式所示：

[0107] $$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}[\hat{y}(i)-y(i)]^2}\quad(i=1,\ 2,\ \ldots,\ n)$$

[0108] 式中，$\hat{y}(i)$为溶解氧的预测值，y(i)为溶解氧的实测值；

[0109] S4、k折交叉验证：将步骤S2-4中得到的输入变量作为原始数据集分成k等份，k取5，每次选取k-1份作为训练集，剩下1份作为测试集，使用不同的超参数组合来训练k-1份和测试其余的1个部分，并计算测试集的RMSE值，重复上述步骤S3-2～S3-5中长短时记忆网络模型训练和测试的步骤，直到k份原始数据集中每个超参数组合都被测试完毕，并计算每个最终输出信息的RMSE平均值，RMSE平均值最小的参数组合为最优组合，如下式所示：

[0110] $$RMSE_{(k)}=\frac{1}{k}\sum_{i=1}^{k}RMSE_i$$

[0111] S5、计算及预测：使用感潮河网区水质自动站的实时数据经预处理后输入到建立好的长短时记忆网络模型中，通过长短时记忆网络模型输出的结果通过放缩得到溶解氧的预测值，采用滚动预报的方法，绘制出溶解氧的走势图，所述步骤S5中滚动预报的方法具体为：依据已有溶解氧的预测值的采样间隔，设置合理的预测时间步长，假设预测的时间为n

日，长短时记忆网络模型会根据测试集中的t日溶解氧数据及S2所述方法筛选出的重要参数，对t+n日的溶解氧数据进行计算并输出得到溶解氧真实值，然后在t+2n日，运用t+n日的溶解氧真实值以及S2所述方法筛选出其他环境变量，采用滚动预报的方法及时更新序列信息，避免误差积累。

[0112] 实施例2

[0113] 本实施例与实施例1基本相同，其不同之处在于：步骤S3-1框架构建中隐藏层的个数不同。

[0114] S3-1、框架构建：基于TensorFlow深度学习框架搭建长短时记忆网络模型，所述长短时记忆网络模型包含1个输入层、1个输出层及3个隐藏层。

[0115] 实施例3

[0116] 本实施例与实施例1基本相同，其不同之处在于：步骤S2-4中最大信息系数MIC(D)的值不同。最大信息系数MIC(D)为0.5，用于预测的变量包含氨氮和总磷。

[0117] 实验例1

[0118] 为了验证本发明的实际应用效果，选择了某水质自动在线站点实际运行的实测水质在线观测数据进行验证。以实施例1中的感潮河网区溶解氧预测方法进行预测，所选择的站点为大龙涌站点，时间跨度为2019年1月1日到2021年3月29日。高锰酸盐指数、氨氮、总磷、总氮的采样频率为4小时，其余变量的时间采样频率为1小时，如表1所示。

[0119] 在步骤S1数据采集中处理后的时间序列样本共有8832个，在步骤S2中分别计算温度、pH、DO25、电导率、浊度、高锰酸盐指数、氨氮、总磷、总氮与溶解氧的MIC(D)值，按MIC(D)值，以0.85为阈值，选取DO25、电导率、水温、氨氮、总氮浓度作为长短时记忆网络模型的预测变量；在步骤S3中，基于主流的TensorFlow深度学习框架搭建长短时记忆网络模型，针对预测模型中涉及的超参数，如图2所示，在步骤S4中采用k折交叉验证网格搜索法进行寻优，以得到最优的超参数组合，选取样本中67％的数据作为训练集，对长短时记忆网络模型进行训练，剩余33％的样本作为测试集，训练及测试结果见图3所示，各个相关变量的计算结果如表1所示，模型参数设置及结果评价列表如表2所示。训练完成后计算均方根误差评估模型性能，其中训练集RMSE为0.29，测试集RMSE为0.22。

[0120] 实验例2

[0121] 本实验例与实验例1基本相同，其不同之处在于：选择的观测站点不同，选择墩头基的数据对模型进行训练和预测测试，各个相关变量的计算结果如表1所示，模型参数设置及结果评价列表如表2所示，训练及测试结果如图4所示。

[0122] 实验例3

[0123] 本实验例与实验例2基本相同，其不同之处在于：选择的网格层数不同，各个相关变量的计算结果如表1所示，模型参数设置及结果评价列表如表2所示，训练及测试结果如图5所示。

[0124] 实验例4

[0125] 本实验例与实验例2基本相同，其不同之处在于：基于实施例3中的最大信息系数MIC(D)为0.5，用于预测的变量包含氨氮和总磷，各个相关变量的计算结果如表1所示，模型参数设置及结果评价列表如表2所示。

[0126] 实验例5

[0127] 本实验例与实验例1基本相同，其不同之处在于：步长有所改变，运用更多的输入和输出时间步长，各个相关变量的计算结果如表1所示，模型参数设置及结果评价列表如表2所示。

[0128] 表1大龙涌站点和墩头基站点中各个相关变量的最大信息系数MIC(D)计算结果列表

[0129]

变量名	MIC系数	
	大龙涌	墩头基
电导率	0.9	0.8
温度	0.85	0.47
PH	0.84	0.4
浊度	0.77	0.7
高锰酸盐指数	0.72	0.69
氨氮	0.91	0.55
总磷	0.8	0.53
总氮	0.87	0.65
溶解氧25	0.92	0.8

[0130] 表2实验案例1-5中模型参数设置及结果评价列表

[0131]

实施案例ID	站位	MIC阈值	优化器	输入时间步长（lag_time）	输出时间步长	网络层数	训练RMSE	测试RMSE
1	大龙涌	0.85	adam	3	1	4	0.29	0.22
2	墩头基	0.65	adam	3	1	4	0.40	0.36
3	墩头基	0.65	adam	3	1	5	0.46	0.54
4	墩头基	0.5	adam	3	1	4	0.46	0.54
5	大龙涌	0.85	adam	5	3	4	0.33	0.39

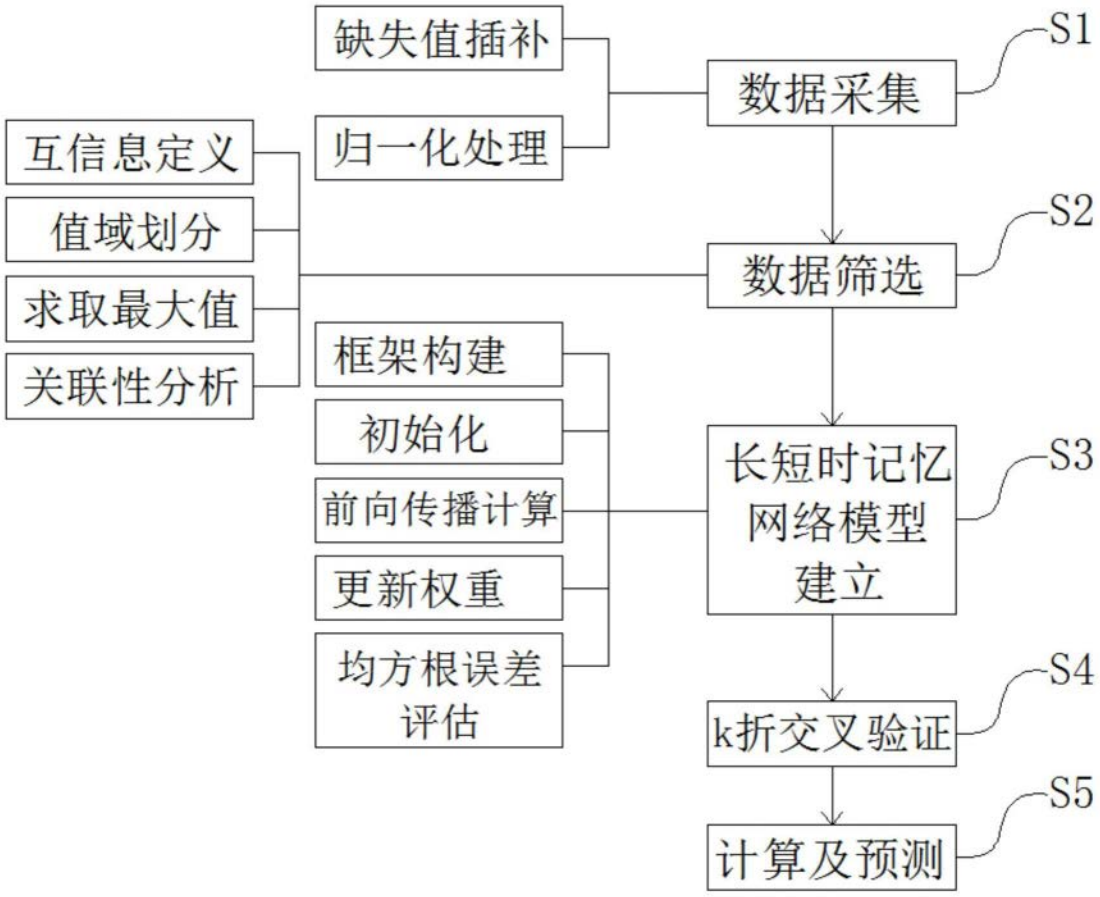

缺失值插补
归一化处理
数据采集
S1
互信息定义
值域划分
求取最大值
关联性分析
数据筛选
S2
框架构建
初始化
前向传播计算
更新权重
均方根误差评估
长短时记忆网络模型建立
S3
k折交叉验证
S4
计算及预测
S5

图1

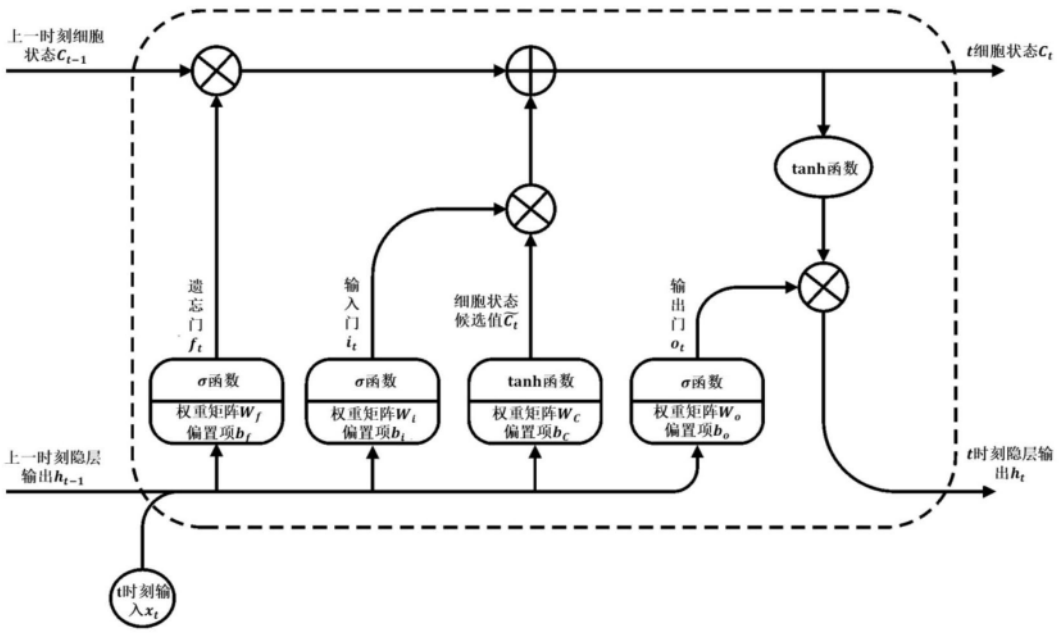

上一时刻细胞状态C_{t-1}
t细胞状态C_t
tanh函数
遗忘门f_t
输入门i_t
细胞状态候选值$\tilde{C}_t$
输出门o_t
σ函数
权重矩阵W_f
偏置项b_f
σ函数
权重矩阵W_i
偏置项b_i
tanh函数
权重矩阵W_C
偏置项b_C
σ函数
权重矩阵W_o
偏置项b_o
上一时刻隐层输出h_{t-1}
t时刻隐层输出h_t
t时刻输入x_t

图2

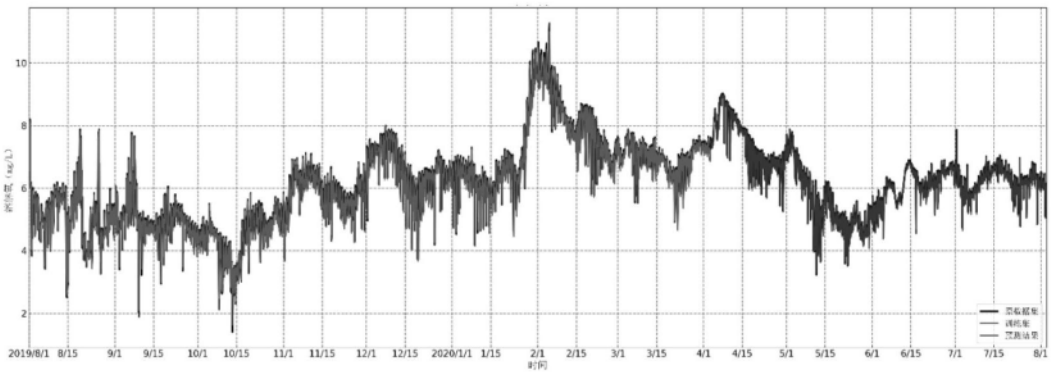

10
8
6
4
2
2019/8/1 8/15 9/1 9/15 10/1 10/15 11/1 11/15 12/1 12/15 2020/1/1 1/15 2/1 2/15 3/1 3/15 4/1 4/15 5/1 5/15 6/1 6/15 7/1 7/15 8/1

图3

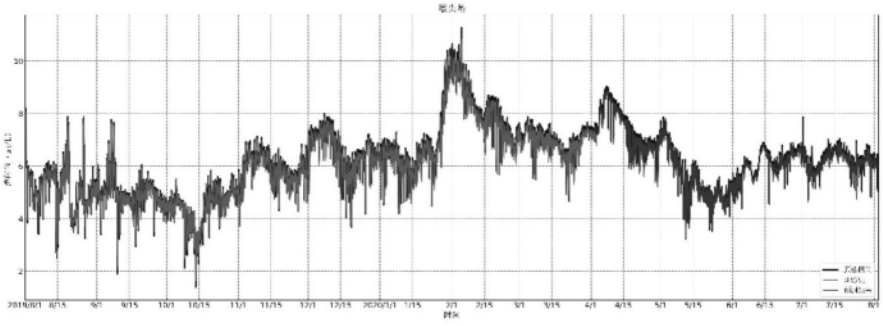


图4

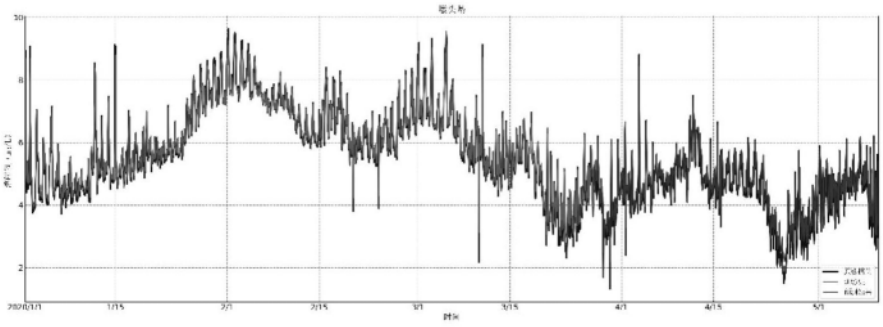


图5