

Long Short-Term Memory Neural Networks Based on Mean Generating Function for Runoff Prediction Modeling Research and Application

Jiansheng Wu¹, Yongsheng Xie¹, Long Jin²

¹School of Mathematical and Computer Sciences, Guangxi Science & Technology Normal University, Liuzhou Guangxi

²Guangxi Climate Center, Nanning Guangxi
Email: wjsh2002168@163.com, 551907834@qq.com

Received: May 6th, 2020; accepted: May 21st, 2020; published: May 28th, 2020

Abstract

Aiming at the problem that it is difficult to determine the modeling factors and the non-linear runoff prediction modeling in runoff prediction modeling, this paper uses a homogeneous function to generate a continuation factor matrix, and extracts a high information entropy matrix in the continuation matrix to generate a modeling factor. A long-term and short-term memory neural network is used to establish a Liujiang runoff prediction model. This model analyzes the characteristics of the original precipitation sequence and extracts different oscillation cycle characteristics of the runoff system to form modeling factors. The advantage of time series dynamic system control is to establish a Liujiang runoff model. This method makes full use of the homogeneous function to extract the sequence dependence and periodic relationship characteristics of the system sequence, and combines the long-term and short-term memory neural network memory characteristics of the non-linear time series to establish a Liujiang runoff simulation prediction model. The experimental results show that the method has good prediction accuracy and good stability, and provides a reliable and effective method for runoff prediction analysis.

Keywords

Mean Generating Function, Information Entropy, Recurrent Neural Network, Long Short-Term Memory Neural Networks, Runoff

基于均生函数的长短期记忆神经网络径流建模研究与应用

吴建生¹, 谢永盛¹, 金龙²

¹广西科技师范学院, 数学与计算机科学学院, 广西 柳州

²广西气象局, 广西 南宁

Email: wjsh2002168@163.com, 551907834@qq.com

收稿日期: 2020年5月6日; 录用日期: 2020年5月21日; 发布日期: 2020年5月28日

摘要

准确的径流建模预测是水文领域中一项具有挑战性的任务,也是减灾防灾的重要研究课题,精确可操作的径流预报模型对于水资源规划及综合利用,水利枢纽运行管理等重大决策问题提供基本决策依据,对国民经济健康发展具有十分重要的意义。针对径流预测建模中难以选择建模因子和难以确定非线性径流预测模型的问题,本文利用均生函数生成延拓因子矩阵,并提取延拓矩阵中信息熵高矩阵生成建模因子,进一步利用长短期记忆神经网络建立柳江径流预测模型,该模型通过分析原始降水序列特性,提取径流系统不同振荡周期特征形成建模因子,并结合长短期记忆神经网络对非线性时间序列记忆特征和非线性时序动力系统控制的优势,以此建立柳江径流模型。该方法充分利用均生函数提取系统序列前后依赖和周期关系特性,并结合长短期记忆神经网络对非线性时间序列记忆特性,此方法能有效地提高预报长度,并能获得较高的建模及预报精度。柳江径流实例进行验证,结果表明了方法预测的精度高,稳定性好,为径流预测分析提供了一种可靠、有效的方法。

关键词

均生函数, 信息熵, 循环神经网络, 长短期记忆神经网络, 径流

Copyright © 2020 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

精确的径流预报模型对于水资源规划及综合利用,水利枢纽运行管理等重大决策问题提供基本决策依据,对国民经济健康发展具有十分重要的意义。由于影响流域河川径流变化因素很多,如蒸发、降水、气候和地下河流等,它们之间的变化机制和相互影响的因素十分复杂,其过程表现出强烈的非线性特征[1] [2],要完整而准确地描述这样一种复杂非线性过程,传统的欧式几何和动力系统模型预测极为困难,因此流域内径流建模是一项具有挑战性的任务[3] [4] [5]。目前国内外对中长期径流预测研究主要的方法有成因分析法、统计分析法、灰色系统分析法、小波分析法等[6]。应用较为广泛统计建模方法主要有三种:自回归(Auto Regressive Model, AR)模型、移动平均(Moving Average Model, MA)模型和自回归移动平均混合(Auto Regressive Moving Average Model, ARMA)模型,这三种模型均可根据历史资料制作多步预测,但预测得到的结果都接近序列的平均值,对未来预测精度不高。而近年来随着计算机硬件和智能计算方法的发展,基于数据驱动的机器学习方法获得迅速发展,其本质上就是从观测样本数据中学习数据输入模式空间与输出空间的广义函数映射关系,建立数据输入和输出之间的知识和规律,并利用这些知识和规律对未来数据或无法观测的数据进行推断、决策和预测[7] [8] [9]。如经典有监督机器学习方法神经网络模型,在建立模型时不需要考虑的径流物理过程,将预测系统构造的神经网络学习矩阵看作是输入与输出之间的一种非线性映射,而实现这种非线性映射关系,并不需要事先知道所研究的

预报系统内部结构,而是通过对有限的样本作学习训练,以模拟预报系统输入与输出的内部结构关系[10] [11] [12]。由于神经网络方法具有很强的处理非线性问题的能力,比一般的方法具有更好的预测能力,它为径流建模分析和预测提供了新技术、新方法[13] [14] [15]。然而随着径流数据的收集越来越多和存储数据能力的不断增长以及计算机运算能力的飞速发展,对复杂数据建模的精度要求越来越高和预测时间区间尺度要求越来越精确,而神经网络方法需要有经验的工程技术人员通过反复训练和测试,网络才能在实际应用中获得一个稳定的解决方案,加之复杂数据的多源和异构,使得神经网络模型在使用中会出现过拟合或者欠学习的问题[16] [17],这极大限制神经网络在实际径流预测建模中的应用,从而使得研究新的机器学习方法越来越重要[18] [19]。

近年来在循环神经网络(Recurrent Neural Networks, RNN)基础上发展的长短期记忆神经网络(Long Short-Term Memory Neural Networks, LSTM),在处理时间序列数据时具有记忆过去和对过去信息采用非线性动力系统控制的优势[20] [21],通过误差反向传播和梯度下降算法求解网络参数达到目标误差最小,在时间序列建模问题上不但可以记忆相邻序列重要信息,而且能够记忆模型长期依赖的关键信息,忽略不必要的信息[22] [23] [24],已经成为深度机器学习重要模型之一。LSTM是在循环神经网络基础上提出的,循环神经网络最早的模型是1982年和1984年美国加州工学院物理学家 John Hopfield 提出的可用作联想存储器的互连网络 Hopfield 神经网络模型,他先后发表了两篇论文介绍了离散时间反馈神经网络网络(Discrete Hopfield Neural Network, DHNN)和连续时间反馈型神经网络网络(Continuous Hopfield Neural Network, CHNN) [25]。随后,在1986年Rumelhart等人, Werbos等人 and Elman等人改进 Hopfield 神经网络模型提出循环神经网络(Recurrent Neural Network, RNN),其主要是以时间序列数据建模和预测为目的,网络的结构与标准多层感知器的结构类似,在按照时间序列的推进方向进行递归且所有节点或者循环单元按链式法则连接的前向神经网络系统[26] [27] [28],不同之处在允许与时间延迟相关的隐藏单元之间进行连接,通过这些连接模型可以保留有关过去的信息,使它发现数据中彼此事件之间的时间相关性。

循环神经网络具有高维的隐藏状态和非线性动力学,使它们能够记忆和处理过去的信息,并且参数共享并且图灵完备,因此在对序列的非线性特征进行学习时具有一定优势[29],先后被成功应用在自然语言处理、语音识别、语言建模、机器翻译等领域有应用,也被用于各类时间序列预测建模[30]。但是,循环神经网络实际应用也存在很多缺陷,其中最大缺陷就是训练网络参数时,误差梯度在经过多个时间步的反向传播后容易导致参数与隐藏状态的动力学之间的关系非常不稳定,包括梯度消失和梯度爆炸[31]。在1997年由 Hochreiter & Schmid Huber 在循环神经网络的隐层引入了存储单元状态(Cell State)和门结构调节网络中信息流的门控机制取代传统的感知器架构扩展了 RNN 架构[32],初步形成 LSTM 的架构,梯度消失问题得以解决,并在近期被 Alex Graves 进行了改良和推广[33]。目前 LSTM 已变得越来越流行,机器学习范式并被成功应用到经济、交通量、语言识别和手写体识别等多个领域[34] [35] [36]。

在使用机器学习方法建立径流模型时,有两个重要因素需要考虑:一是选择数据的特征,二是选择恰当的机器学习模型。在径流序列建模研究时如何选取与径流相关性的建模因子,并且排除多重变量相关因素之间的影响是径流建模首要考虑的因素。为此本文依据径流时间序列的系统状态具有不同时期前后依赖的特征,利用魏凤英、曹鸿兴提出的均生函数模型思想构建一组周期函数[37] [38],通过分析原序列与这组周期函数间的统计关系,提取系统不同振荡周期特征形成建模因子,并结合长短期记忆神经网络对非线性时间序列记忆特征和非线性时序动力系统控制的优势,建立柳江径流机器学习模型。

2. 基于均生函数的长短期记忆神经网络模型

2.1. 均生函数建模的原理和方法

均生函数(Mean Generating Function, MGF)建模主要是基于动态时间序列系统状态前后具有依赖的记

忆思想,把原始序列以及它的一阶差分、二阶差分按照不同时间间隔计算均值,构建周期函数及其延拓序列。通过分析原序列与这组周期函数间及其延拓序列的统计关系,建立回归预测方程,其目的就是将动态的时间序列分析转化为静态数据分析,将单一的要素通过不同的因子生成方式,转化为包含不同预报信息的多个预报因子序列[39] [40]。原理为:

设某一时间序列为

$$y(t) = \{y(1), y(2), \dots, y(n)\} \tag{1}$$

式中 n 为样本长度,则序列的平均值为 $\bar{x}_1(1) = \frac{1}{n} \sum_{i=1}^n y(i)$, 则均值生成函数为:

$$\bar{x}_l(i) = \frac{1}{n_l} \sum_{j=0}^{n_l} y(i + j * l), \quad i = 1, 2, \dots, l, \quad 1 \leq l \leq m \tag{2}$$

式中 $n_l = INT\left(\frac{n}{l}\right), m = INT\left(\frac{n}{2}\right)$, INT 表示取整,对均生函数 $\bar{x}_l(i)$ 按照式(3)做周期延拓计算由此构造均生函数延拓序列矩阵

$$x_l(t) = \bar{y}_l \left[t - l \cdot INT\left(\frac{t-1}{l}\right) \right] \tag{3}$$

$$X = \begin{bmatrix} \bar{x}_1(1) & \bar{x}_2(1) & \bar{x}_3(1) & \dots & \bar{x}_M(1) \\ \bar{x}_1(2) & \bar{x}_2(2) & \bar{x}_3(2) & \dots & \bar{x}_M(2) \\ \bar{x}_1(3) & \bar{x}_2(3) & \bar{x}_3(3) & \dots & \bar{x}_M(3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1(i_2) & \bar{x}_2(i_2) & \bar{x}_3(i_2) & \dots & \bar{x}_M(i_2) \end{bmatrix} \tag{4}$$

同时为对原始序列 $y(t)$ 再分别计算一阶差分和二阶差分后再次计算均生函数并延拓获得一阶差分和二阶差分矩阵,对 3 个延拓的均生函数序列矩阵的信息熵计算采用由此可以获得:

$$2I = 2 \left[\sum_{i=1}^G \sum_{j=1}^G n_{ij} \ln n_{ij} + N \ln N - \left(\sum_{i=1}^G n_i \ln n_i + \sum_{j=1}^G n_j \ln n_j \right) \right] \tag{5}$$

其中 I 为对分类预测的信息熵评分, n_{ij} 为 i 类事件与 j 类估计事件的列联表中的个数,挑选出预报信息熵 (Information Entropy, IE) 相对高的延拓均生函数序列作为建模因子。

2.2. 长短期记忆神经网络建模方法

假设给定一个时间序列 $x(t) = \{x(1), x(2), \dots, x(n)\}$, 传统时间序列模型是建立自回归模型进行预测,具体表达式如下:

$$x(t) = \omega_0 + \sum_i^p \omega_i x(t-i) + \varepsilon_t \tag{6}$$

其中 p 为延迟阶数, ω_0 为回归系数, $\varepsilon_t \sim N(0, \sigma^2)$ 为白噪声, LSTM 针对时间序列问题,后续每一时刻有外部输入时,则模型演变成有外部输入的非线性自回归模型时间序列模型:

$$y(t) = f \{x(t-1), x(t-2), \dots, x(t-p), y(t-1), y(t-2), \dots, y(t-p)\} \tag{7}$$

其中 $f(\cdot)$ 为利用神经网络模拟估计逼近函数。原始 RNN 的隐藏层只有一个 h 状态,它对于短期的输入非常敏感,很难实现自身反馈和记忆状态。LSTM 通过增加一个存储单元状态 c 来实现,让网络保存长期的状态。在 t 时刻, LSTM 单元模块有三个输入: (1) 当前时刻网络的输入值 x_t ; (2) 上一时刻 LSTM

的网络隐层输出值 h_{t-1} ; (3) 上一时刻网络单元状态的输出 c_{t-1} ; LSTM 的网络输出有两个: 当前时刻 LSTM 网络隐层输出值 h_t , 和当前时刻网络单元状态的输出 c_t , 具体 LSTM 网络模块结构如图 1 所示。

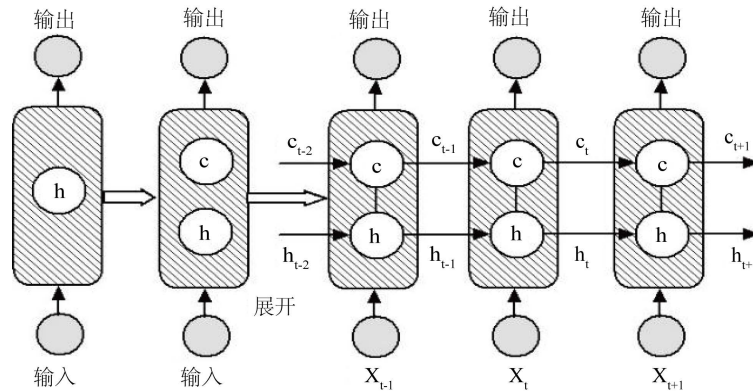


Figure 1. LSTM neural network model structure and fold diagram
图 1. LSTM 神经网络模型结构和展开图

LSTM 隐层记忆单元具有的遗忘门、输入门和输出门使得网络具有记忆单元长短时记忆机制, 该机制一致地确定要保留的信息, 要保留的时间以及何时从存储单元读取信息。在存储单元状态中配备三个门结构门, 即: 遗忘门、输入门和输出门。通过门控单元可以对存储单元添加和删除信息, 有选择地决定信息是否通过, 采用 sigmoid 神经网络层和一个成对乘法操作组成, 该层的输出是一个介于 0 到 1 的数, 表示允许信息通过的多少, 0 表示完全不允许通过, 1 表示允许完全通过[41] [42] [43]。LSTM 的建模过程隐层单元的结构图如图 2 所示。

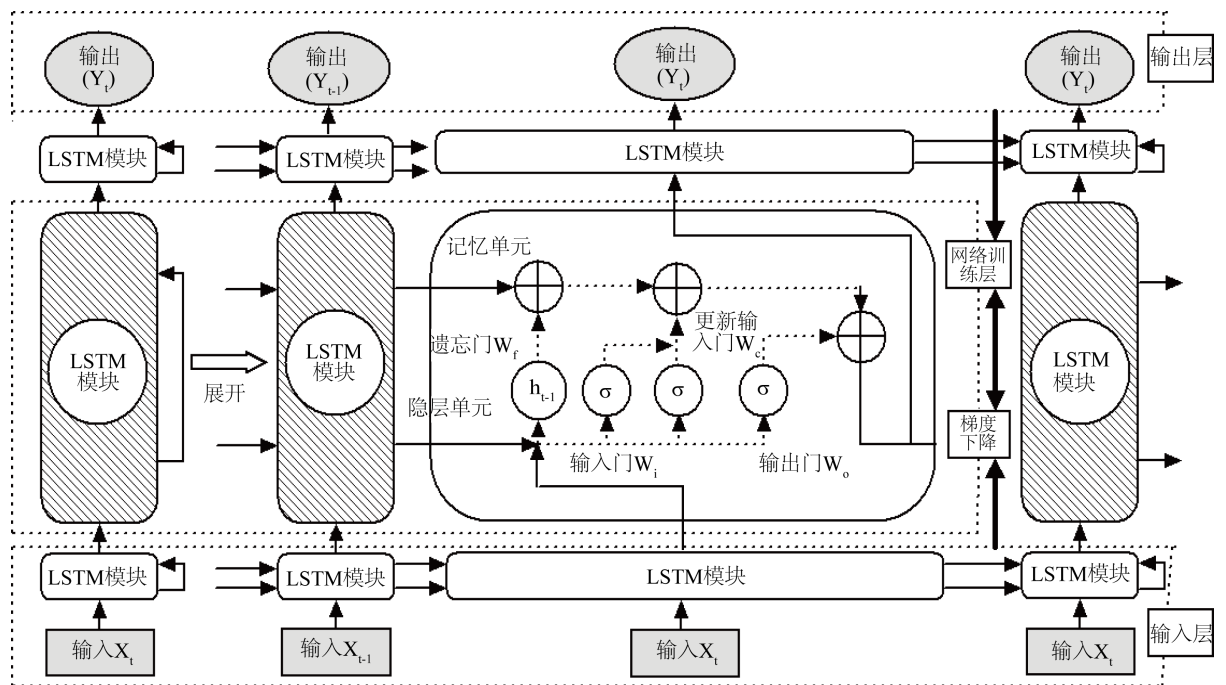


Figure 2. Structure diagram of the hidden layer unit of LSTM
图 2. LSTM 的隐层单元的结构图

遗忘门决定上一时刻的单元状态有多少“记忆”可以保留到当前时刻，其计算模型如下：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{6}$$

而输入门根据上一次的输出 h_{t-1} 和本次输入 x_t 来计算当前输入的单元状态，避免当前无关紧要的内容进入记忆，决定当前时刻网络输入 x_t 有多少保存在单元状态 c_t 中，其计算模型如下：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{8}$$

用于描述当前输入状态 c_t' 的计算模型如下：

$$Z_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{9}$$

这是一个 \tanh 激活函数生成新的候选值 z_t' ，它作为当前层产生的候选值会添加到单元状态中，当前时刻输入的单元状态 c_t 计算模型如下：

$$c_t = f_t \cdot c_{t-1} + i_t \cdot z_t' \tag{10}$$

则模型的输出 o_t, h_t 的计算，首先是先通过 σ 层来得到一个初始输出，然后使用 \tanh 将 c_t 缩放到-1到1间，再与 σ 得到的输出逐对相乘，从而得到模型的输出，具体计算如下：

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{11}$$

$$h_t = o_t * \tanh(c_t) \tag{12}$$

以上公式中 W_f, W_i, W_c, W_o 是权重矩阵， b_f, b_i, b_c, b_o 是偏置项， $[h_{t-1}, x_t]$ 表示把两个向量连接成一个更长的向量， σ 是门的激活函数，表达式如下：

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{13}$$

LSTM 主要是求解以上 8 个参数，损失函数是交叉熵损失函数，表示为：

$$L = \frac{1}{N} \sum_{t=1}^N \left(- \sum_{j=1}^V (y_{tj} * \ln \hat{y}_{tj}) \right) \tag{14}$$

采用梯度下降的误差反向传播算法(Back-Propagation Through Time, BPTT)求解模型的参数，首先确定参数的初始化值，然后向前计算每个神经元的输出值；反向计算每个神经元的误差项值，LSTM 误差项的反向传播包括两个层面：空间层面上将误差项向网络的上一层传播，时间层面上沿时间反向传播，即从当前 t 时刻开始，计算每个时刻的误差。每个权重(即参数)具体推导和计算过程见文献[44] [45]。

2.3. 柳江径流建模过程

本文利用柳江原始径流和一阶差分序列的均生函数生成径流序列延拓矩阵，并利用信息熵筛选建模因子作为神经网络输入训练集，建立非线性自回归外生输入神经网络(Nonlinear Auto Regressive with External Input by Neural Network, NARX-NN)和长短期记忆神经网络(Information Entropy and Long Short-Term Memory Neural Network, IE-LSTMNN)的径流模型，具体计算过程如图 3 和图 4 所示。

3. 应用实例及其结果分析

3.1. 建模数据

本文建模的数据选取广西柳州市老桥口柳江径流 2001 年 1 月 1 日~2010 年 12 月 31 日期间 10 年每天 12 时的径流(m^3/s)数据进行应用实例分析，全部的数据 3652 个，图 5 为 3652 个柳江径流数据实况。

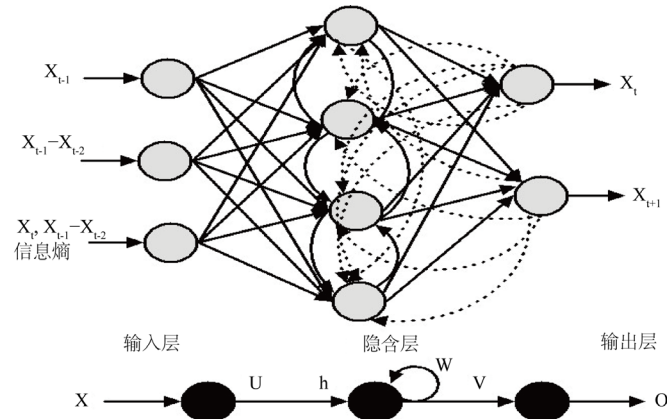


Figure 3. Flow chart of NARX-NN runoff modeling
图 3. NARX-NN 径流建模流程图

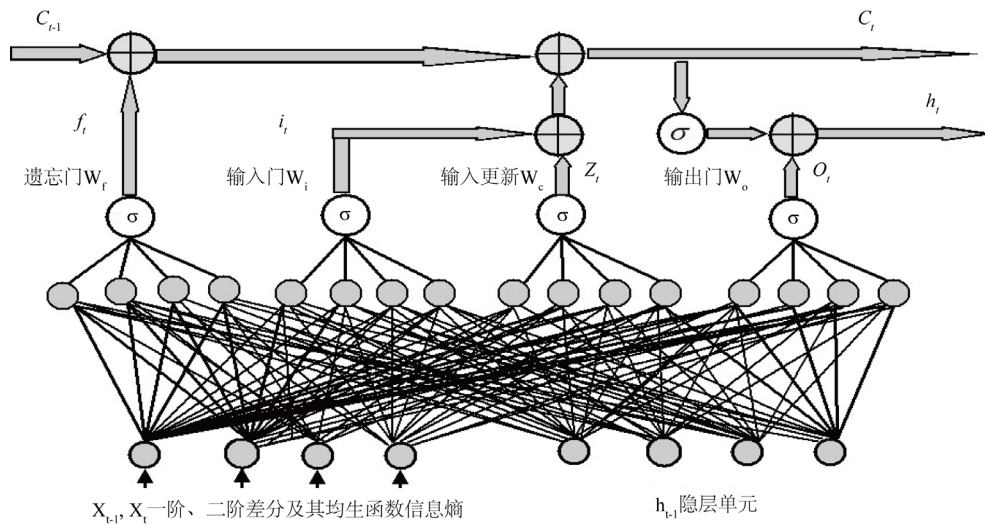


Figure 4. Flow chart of IE-LSTM runoff modeling
图 4. IE-LSTM 径流建模流程图

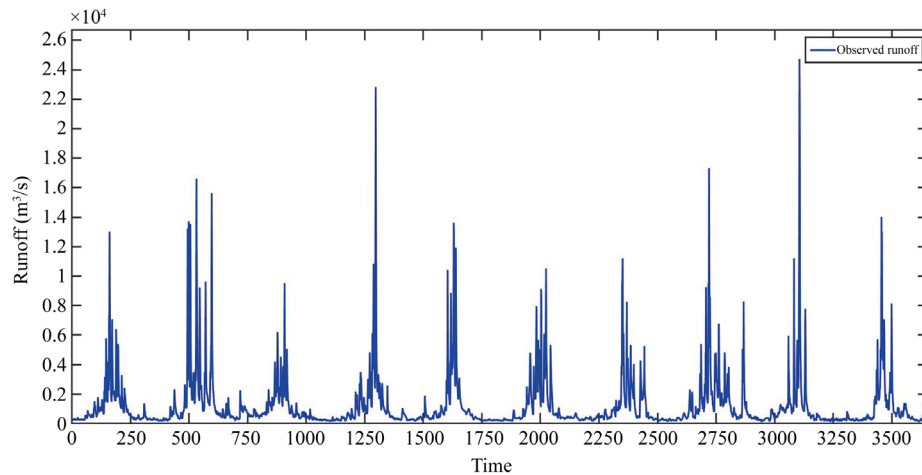


Figure 5. Liujiang River runoff of 3652 actual observed from January 1, 2001 to December 31, 2010
图 5. 从 2001 年 1 月 1 日~2010 年 12 月 31 日 3652 个柳江径流实际观测值

在机器学习方法中, 利用有监督学习方法建立模型时, 通常会设置一定比例的验证数据集(Validation Set), 使用训练数据集去训练模型, 使用验证集去评估模型的泛化能力, 其目的防止对训练数据集(Training set)过度训练而导致最小化过度拟合。为此把数据分为训练数据集训练模型参数、验证数据集调整优化模型参数和测试数据集检验模型实际预测效果, 其中 2001 年 1 月 1 日到 2007 年 12 月 31 日共计 2556 个数据作为训练数据集建立径流拟合模型, 2008 年 1 月 1 日至 2009 年 12 月 31 日共计 731 个数据作为验证数据集优化校验模型, 2010 年 1 月 1 日到 2010 年 12 月 31 日共计 365 个数据为测试数据集检验预测模型效果。柳江每年 5 月份进入汛期, 10 月份汛期结束, 为了较为准确预测灾害性洪水, 本文把径流大于 5000 (m²/s)作为洪水峰值来预测, 表 1 是训练数据、验证数据和测试数据样本和洪水峰值样本总数情况, 图 6 是不同数据集中径流峰值分布示意图。

Table 1. Training, validation and test data samples and statistics table of the total number of flood peaks
表 1. 训练、验证和测试数据样本和洪水峰值总数统计表

数据集	数据起止时间	样本数	径流峰值个数
训练数据集	2001 年 1 月 1 日~2007 年 12 月 31 日	2556	99
验证数据集	2008 年 1 月 1 日~2009 年 12 月 31 日	731	34
测试数据集	2010 年 1 月 1 日~2010 年 12 月 31 日	365	15

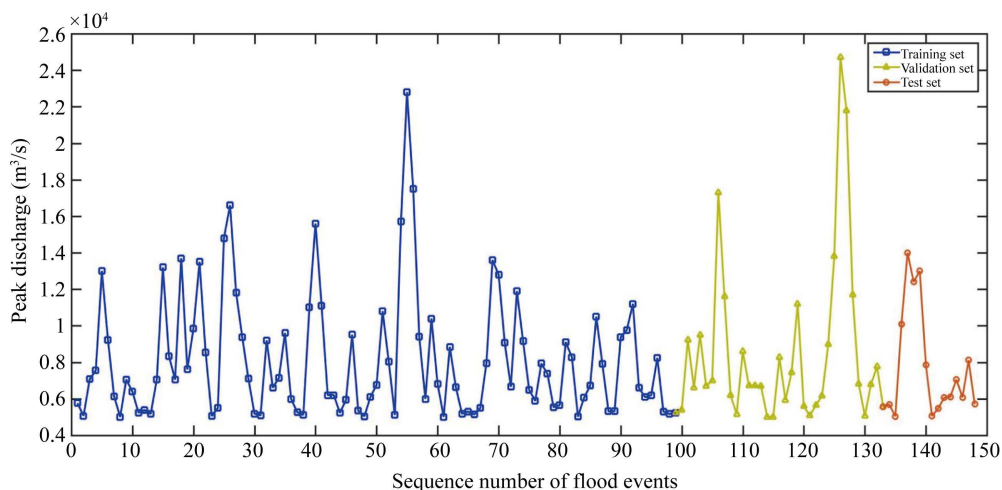


Figure 6. Distribution of peak runoff in training, verification and test data sets
图 6. 训练、验证和测试数据集中径流峰值分布情况

3.2. 模型性能评价

本文引入以下 4 种统计指标:

- (1) 均方根误差(Root Mean Square Error, RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \tag{15}$$

- (2) 平均绝对百分比误差(Mean Absolute Percentage Error, MAPE):

$$MAPE = \frac{100}{n} \sum_{t=1}^n \frac{|y_t - \hat{y}_t|}{|y_t|} \tag{16}$$

(3) Pearson 相关系数(Pearson correlation coefficient, PCC)

$$PCC = \frac{\sum_{t=1}^n (y_t - \bar{y}_t) \sum_{t=1}^n (\hat{y}_t - \bar{\hat{y}}_t)}{\sqrt{\sum_{t=1}^n (y_t - \bar{y}_t)^2 \sum_{t=1}^n (\hat{y}_t - \bar{\hat{y}}_t)^2}} \quad (17)$$

(4) 纳什效率系数(Nash-Sutcliffe Efficiency coefficient, NSE):

$$NSE = 1 - \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (y_t - \bar{y})^2} \quad (18)$$

式(15)~(18)中、 y_t, \hat{y}_t 分别代表径流实际值和模型的输出值, $\bar{y}_t, \bar{\hat{y}}_t$ 分别表示径流实际值平均值和模型的输出值平均值, n 为样本个数, 在纳什效率系数的基础上, 提出洪水峰值模型预测性能评价指标:

(5) 峰值纳什效率系数

$$NSE_{peak} = 1 - \frac{\sum_{i=1}^p (y_{pi} - \hat{y}_{pi})^2}{\sum_{i=1}^p (y_{pi} - \bar{y}_{pi})^2} \quad (19)$$

式(19)中、 y_{pi}, \hat{y}_{pi} 分别表示洪水峰值实际值和预测洪水峰值输出值, \bar{y}_{pi} 表示洪水峰值的平均值, p 为洪水峰值期间样本个数。性能评价指标(1)和(2)是衡量实际径流值和模型计算值的偏离程度, 其值越小说明模型的预测性能越好; 性能指标(3)主要是看模型能否对径流的趋势做出正确的判断, 其值越大就说明模型能够较为准确把握径流未来趋势; 性能指标(4)和(5)是验证径流模型结果的好坏, NSE 取值为负无穷至 1, 越接近 1 表示模型的效果越好, 模型可信度高; NSE 接近 0 表示模型结果接近观测值的平均水平, 即模型总体结果可信, 如果 NSE 远远小于 0 则模型是不可信的。

3.3. 结果分析

表 2 是两种模型对柳江径流训练数据集的拟合结果性能统计表, 图 7 是两种模型对训练数据集中径流峰值拟合效果, 从表 2 的结果可以看出 IE-LSTM 模型对训练数据集中的 RMSE 和 MAPE 性能指标均小于 NARX-NN 模型, 说明 IE-LSTM 模型拟合值和真实值之间的偏离程度最小, 拟合精度高于 NARX-NN 模型; 同时可以看出 IE-LSTM 模型对训练数据集的 Pearson 相关系数为 0.9978, 而 NARX-NN 模型为 0.8807, 二者比较接近, 说明这两种模型均可以对径流趋势做出较为准确的拟合, IE-LSTM 模型拟合精度高于 NARX-NN 模型。

Table 2. Performance indicators of NARX-NN and IE-LSTM models for training samples

表 2. NARX-NN 和 IE-LSTM 模型对训练样本计算结果性能指标统计表

训练数据集(2001 年 1 月 1 日~2007 年 12 月 31 日)					
模型	RMSE	MAPE	PCC	NSE	NSE _{peak}
NARX-NN	847.1152	47.5727%	0.8807	0.7746	0.1784
IE-LSTM	126.2828	8.5085%	0.9978	0.9950	0.9850

从纳什效率系数指标考察这两种模型, NARX-NN 模型置信度只有 17.84%, 置信度很低; 而 IE-LSTM 模型对径流峰值置信度很高, 可以达到 98.50%。说明 IE-LSTM 模型相对 NARX-NN 模型具有很好的学习能力、对径流峰值跟踪能力。本文涉及两种神经网络方法均是通过训练数据集来调整网络的权重, 通过验证数据集校验网络性能, 在训练过程中训练数据集性能误差减小, 而验证数据集的性能误差保持不

变或升高，则停止网络训练过程，以此找出效果最佳的模型所对应的参数，用来对测试数据集(Test set)预测。

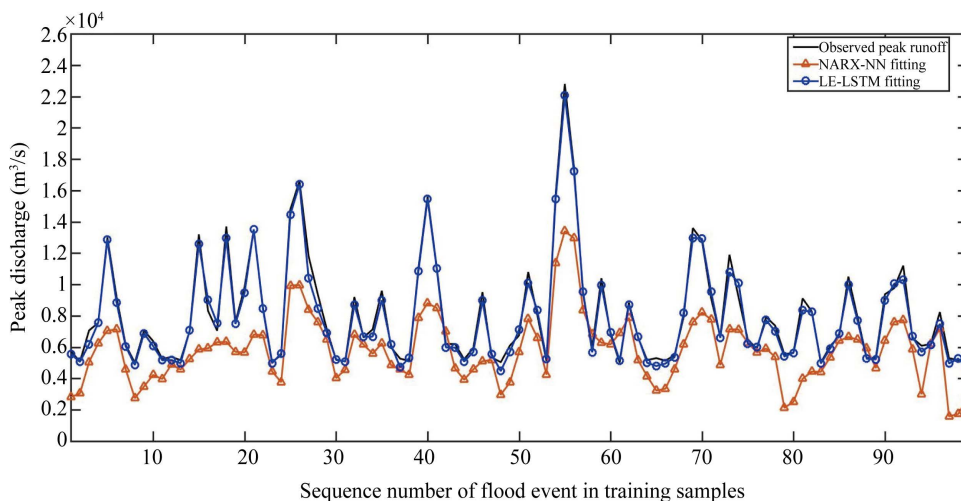


Figure 7. The fitting result of two different models on the peak discharge in the training data sets
图 7. 两种不同模型对训练数据集中径流峰值拟合效果

表 3 是两种模型对柳江径流验证数据集的预测结果性能统计表，图 8 是两种模型对验证数据集中径流峰值预测效果，从表 3 的结果可以看出，IE-LSTM 模型的性能指标 RMSE 和 MAPE 均小于 NARX-NN 模型，性能指标 PCC 和 NSE 均大于 NARX-NN 模型，在对验证数据集中径流峰值预测中，IE-LSTM 模型置信度高达 0.9917，而 NARX-NN 模型置信度只有 0.2130，这些结果进一步说明 IE-LSTM 模型对验证数据集的预测误差精度小、趋势预测和实际观测值高度一致，对径流峰值预测置信度很高；对比表 2 和表 3，结果显示 IE-LSTM 模型对训练数据集的 RMSE 为 126.2828，对验证数据集的 RMSE 为 129.1887，二者非常接近，说明 IE-LSTM 模型没有过拟合，可以用训练好的模型对未来进行预测。

Table 3. Forecasting performance indicators of two models for validation samples
表 3. NARX-NN 和 IE-LSTM 模型对验证样本预测结果性能指标统计表

验证数据(2008 年 1 月 1 日~2009 年 12 月 31 日)					
模型	RMSE	MAPE	PCC	NSE	NSE _{peak}
NARX-NN	858.0234	44.2265%	0.7916	0.4505	0.2130
IE-LSTM	129.1887	8.7650%	0.9972	0.9935	0.9917

有监督机器学习的主要任务就是利用建立的模型并实现对未来的精准预测，训练数据集和验证数据集训练得到的模型能否适用于待测试数据集的新数据，即模型具有泛化能力。因此，当在评估模型的性能时需要重点考察训练好的机器学习模型在新数据上准确率如何，这样才有把握用它对未来做出高质量的预测结果。表 4 是两种模型对柳江径流测试数据集的预测结果性能统计表，图 9 是 NARX-NN 模型对 365 个测试数据集径流预测效果，图 10 是 IE-LSTM 模型对 365 个测试数据集径流预测效果。图 11 是两种模型期望输出和目标输出之间相关性以及回归效果。

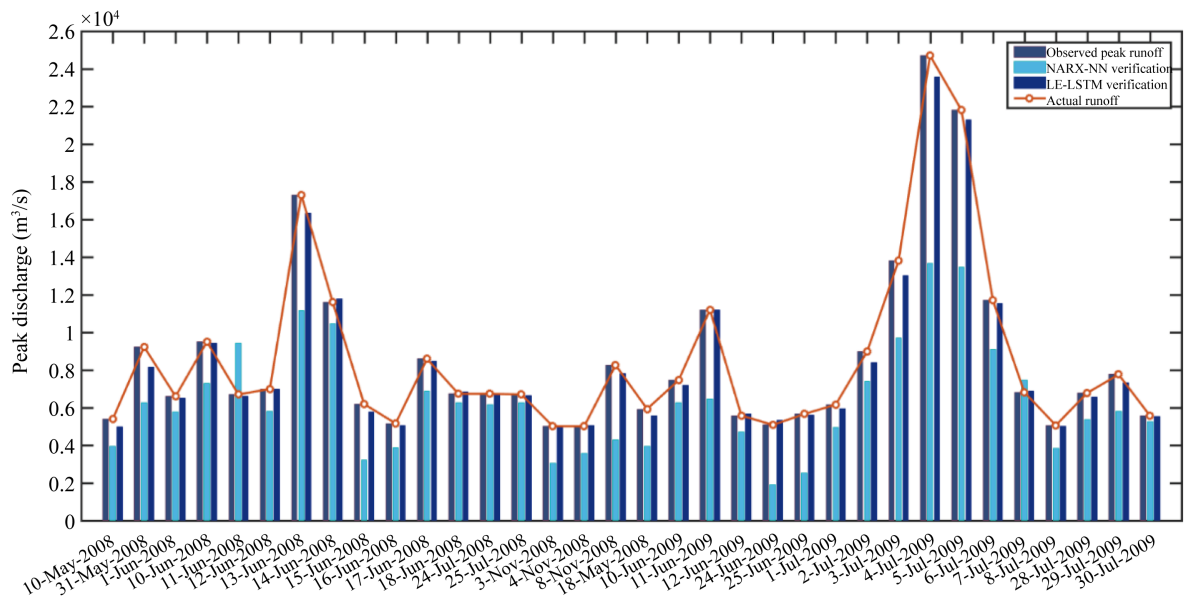


Figure 8. The forecasting result of two different models on the peak discharge in the training data sets

图 8. 两种不同模型对验证数据集中径流峰值预测效果

Table 4. Forecasting performance indicators of two models for test samples

表 4. NARX-NN 和 IE-LSTM 模型对测试样本预测结果性能指标统计表

测试数据(2010年1月1日~2010年12月31日)					
模型	RMSE	MAPE	PCC	NSE	NSE _{peak}
NARX-NN	843.2273	82.8031%	0.6730	0.6660	0.2919
IE-LSTM	93.0750	11.1170%	0.9941	0.9971	0.9972

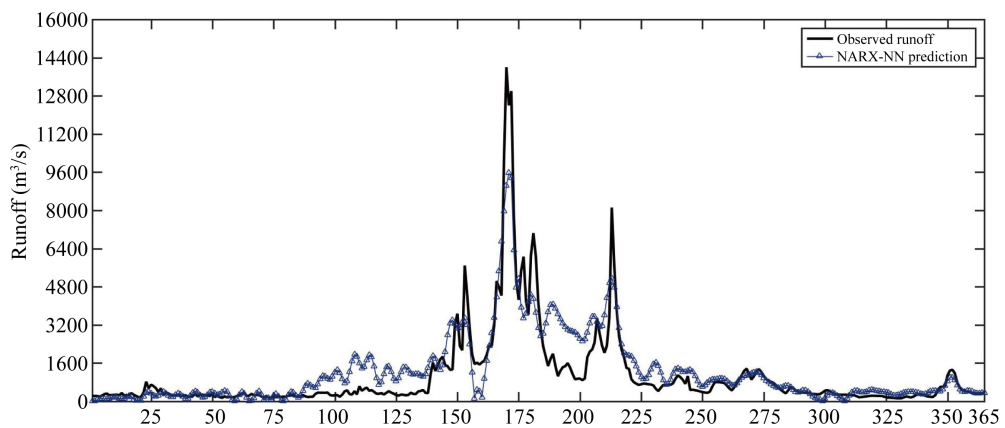


Figure 9. Effect of NARX-NN model on runoff prediction of 365 test sets

图 9. NARX-NN 模型对 365 个测试数据集径流预测效果

从表 4 结果可以看出 NARX-NN 模型对测试数据集预测的 RMSE 为 843.2273, IE-LSTM 模型对测试数据集的预测 RMSE 为 93.0750, 显然 IE-LSTM 模型的 RMSE 比 NARX-NN 模型小的多, 同样 IE-LSTM 模型的 MAPE 也比 IE-LSTM 模型小很多, 这说明 IE-LSTM 模型的预测值和实际值偏差小, 其预测的精

度高；对于 PCC 性能指标，NARX-NN 模型是 0.6730，IE-LSTM 模型是 0.9941，这说明 IE-LSTM 模型的预测值和实际观测值高度相关。

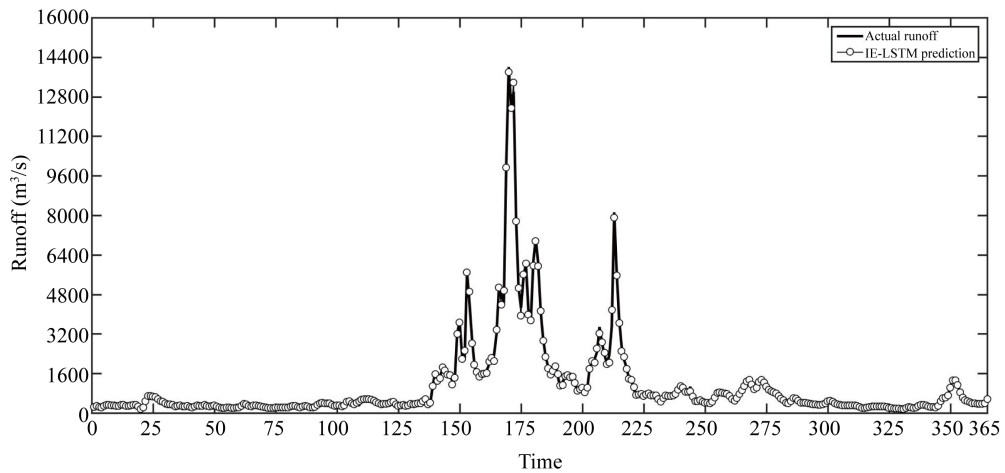


Figure 10. Effect of IE-LSTM model on runoff prediction of 365 test sets
图 10. IE-LSTM 模型对 365 个测试数据集径流预测效果

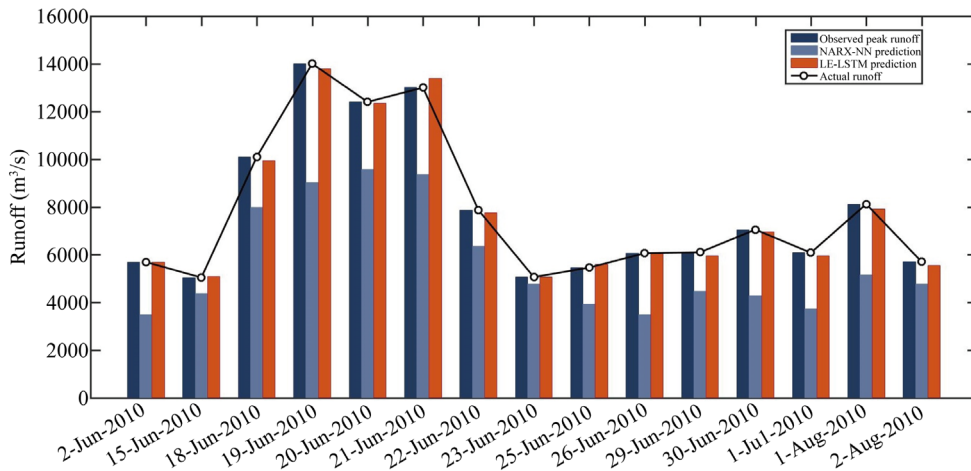


Figure 11. The forecasting result of two different models on the peak discharge in the test data sets
图 11. 两种不同模型对测试数据集中径流峰值预测效果

大量数据中离群点(Outline)检测、识别和预测是机器学习挖掘数据中潜在特征重要组成部分，它的任务就是寻找观测值和参照值之间有意义的较大偏差，对这些异常记录的检测、识别和解释有很重要的意义。在一年中绝大多数时间径流数据序列处于稳定状态，每年汛期径流峰值会有很大波动，即整个数据中离群点。如何利用机器学习识别和预测径流峰值是预测灾害性洪水达到时刻和对径流峰值趋势跟踪重要指标，也是检验模型能否预测灾害性洪水以及在实际应用的重要指标。图 11 是两种不同模型对测试数据 2010 年 6 月~8 月径流峰值预测效果，从表 4 的数据性能指标和图 11 预测效果的结果可以看出，在 2010 年全年 365 个数据中，流量大于 5000 (m³/s)有 15 个，其中 2010 年 6 月 18 日径流峰值流量为：10100 (m³/s)，6 月 19 日径流峰值流量为：14000 (m³/s)，6 月 20 日径流峰值流量为：12400 (m³/s)，6 月 21 日径流峰值流量为 13000 (m³/s)，IE-LSTM 模型可以学习径流数据中离群点，并能准确跟踪离群点，及就是能够预测径流峰值的趋势，而且精度最高。这些结果均说明 IE-LSTM 模型可以用于预测灾害性洪水，并且能够在

实际径流预测建模中应用，而且结果稳定。

4. 结语

流域内径流是水文系统中研究人员关注的重大课题，它关系到水利工程建设和水资源的开发利用，由于径流系统包含着系统空间和时间变化的非线性特性，传统的分析方法很难对它们的演化规律和变化特征做出清晰的了解和准确的预测。针对径流预测建模过程中难以筛选合适的建模因子问题和难以建立有效非线性径流时间序列模型问题，本文利用均生函数延拓矩阵分析原始时间序列周期函数间的统计关系，提取原始序列不同振荡周期特征形成径流建模因子，并结合长短期记忆神经网络对非线性时间序列记忆特征和非线性时序动力系统控制的优势，建立柳江径流机器学习模型。通过实例计算得出如下结论：

(1) 在利用 LSTM 建立大规模、非同源异构数据机器学习模型时，如何选择有效数据特征是影响模型性能因素之一，本文利用一阶差分、二阶差分平稳后处理时间序列，在利用均生函数和信息熵复杂数据提取径流时间序列中系统空间和时间变化的非线性特性是一种有效时间序列特征提取技术。

(2) 针对径流非线性时间序列建模过程中，采用“理论 + 实验 + 计算”模式分析数据建立 IE-LSTM 径流模型，实例计算结果 IE-LSTM 适宜建立径流模型，对径流汛期峰值趋势跟踪预测效果最好，预测结果精度高，稳定性好。

(3) 本文 LSTM 建立径流模型，LSTM 实质是一种深度神经网络模型，建立模型拟合和预测效果好，但是还有一些问题需要探讨，如：深度神经网络模型是一种黑箱结构，模型接受输入 - 输出完成复杂非线性映射，但是无法解释模型表现好的具体原因；机器学习模式是建立在数据满足独立同步分布条件下获得的效果，但是无法判断整个训练数据满足独立同分布的条件；预测数据和验证数据集的增添无法判断其中反馈关系等，这将是下一步需要深入探讨的问题。

基金项目

国家自然科学基金资助项目(No. 41575051, No. 41565005)，广西科技厅科技基地和人才专项(No. AD16450003)，广西科技厅创新项目(No. 2018AB14003)，广西高校中青年教师基础能力提升项目(No. 2018KY0699, No. 2018KY0700, No. 2018KY0701)，广西来宾市科学研究与技术开发项目(来科能 193305)。

参考文献

- [1] Wang, W., Chau, K., Xu, D., et al. (2015) Improving Forecasting Accuracy of Annual Runoff Time Series Using ARIMA Based on EEMD Decomposition. *Water Resources Management*, **29**, 2655-2675. <https://xs.scribbr.com/https://doi.org/10.1007/s11269-015-0962-6>
- [2] Farajzadeh, J., Fard, A.F. and Lotfi, S. (2014) Modeling of Monthly Rainfall and Runoff of Urmia Lake Basin Using “Feed-forward Neural Network” and “Time Series Analysis” Model. *Water Resources and Industry*, **7**, 38-48. <https://doi.org/10.1016/j.wri.2014.10.003>
- [3] 黄强, 赵雪花, 等. 河川径流时间序列分析预测理论与方法[M]. 郑州: 黄河水利出版社, 2008: 2-12.
- [4] 金龙著. 神经网络气象预报建模理论与方法[M]. 北京: 气象出版社, 2004: 14-15.
- [5] Wu, J.S. and Xie, Y.S. (2019) Hybrid Support Vector Regression with Parallel Co-Evolution Algorithm Based on GA and PSO for Forecasting Monthly Rainfall. *Journal of Software Engineering and Applications*, **12**, 524-539. <https://doi.org/10.4236/jsea.2019.1212032>
- [6] 李彦彬, 黄强, 徐建新, 等. 河川径流混沌特征及预测理论与实践[M]. 北京: 中国水利水电出版社, 2011: 10-12.
- [7] 高新波, 张军平. 机器学习及其应用 2015[M]. 北京: 清华大学出版社, 2015: 8-9.
- [8] Wu, J.S., Jin, L. and Liu, M. (2010) A Hybrid Support Vector Regression Approach for Rainfall Forecasting Using Particle Swarm Optimization and Projection Pursuit Technology. *International Journal of Computational Intelligence and Applications*, **9**, 87-104. <https://doi.org/10.1142/S1469026810002793>

- [9] Makungo, R., Odiyo, J.O. and Mwak, B. (2010) Rainfall-Runoff Modelling Approach for Ungauged Catchments: A Case Study of Nzhelele River Sub-Quaternary Catchment. *Physics and Chemistry of the Earth*, **35**, 45-62. <https://doi.org/10.1016/j.pce.2010.08.001>
- [10] Wu, J.S. (2011) An Effective Hybrid Semi-Parametric Regression Strategy for Rainfall Forecasting Combining Linear and Nonlinear Regression. *International Journal of Applied Evolutionary Computation*, **2**, 50-65. <https://doi.org/10.4018/jaec.2011100104>
- [11] Refona, J., Lakshmi, M., Dhamodaran, S., et al. (2019) Machine Learning Techniques for Rainfall Prediction Using Neural Network. *Journal of Computational and Theoretical Nanoscience*, **16**, 3319-3323. <https://doi.org/10.1166/jctn.2019.8185>
- [12] Mosavi, A., Ozturk, P. and Chau, K. (2018) Flood Prediction Using Machine Learning Models: Literature Review. *Water*, **10**, 1536. <https://doi.org/10.3390/w10111536>
- [13] Mishra, N., Soni, H.K., Sharma, S., et al. (2018) Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data. *International Journal of Intelligent Systems and Applications*, **11**, 16. <https://doi.org/10.5815/ijisa.2018.01.03>
- [14] Liu, Q., Zhou, Y., Liu, X. and Nigel, L. (2019) A Survey on Rainfall Forecasting Using Artificial Neural Network. *International Journal of Embedded Systems*, **11**, 240-249. <https://doi.org/10.1504/IJES.2019.098300>
- [15] Yaseen, Z.M., Ghareb, M.I., Ebtehaj, I., et al. (2018) Rainfall Pattern Forecasting Using Novel Hybrid Intelligent Model Based ANFIS-FFA. *Water Resources Management*, **32**, 105-122. <https://doi.org/10.1007/s11269-017-1797-0>
- [16] Graves, A., Wayne, G., Reynolds, M., et al. (2016) Hybrid Computing Using a Neural Network with Dynamic External Memory. *Nature*, **538**, 471-476. <https://doi.org/10.1038/nature20101>
- [17] Van, G.M. and Bohte, S. (2017) Artificial Neural Networks as Models of Neural Information Processing. *Frontiers in Computational Neuroscience*, **11**, 114. <https://doi.org/10.3389/fncom.2017.00114>
- [18] Niu, W., Feng, Z., Cheng, C., et al. (2018) Forecasting Daily Runoff by Extreme Learning Machine Based on Quantum-Behaved Particle Swarm Optimization. *Journal of Hydrologic Engineering*, **23**, Article ID: 04018002. [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001625](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001625)
- [19] Feng, Z., Niu, W., Tang, Z., et al. (2020) Monthly Runoff Time Series Prediction by Variational Mode Decomposition and Support Vector Machine Based on Quantum-Behaved Particle Swarm Optimization. *Journal of Hydrology*, **583**, Article ID: 124627. <https://doi.org/10.1016/j.jhydrol.2020.124627>
- [20] 刘炳春, 齐鑫, 王庆山. 北京城市代谢预测研究-基于长短期记忆神经网络模型[J]. 地理科学进展, 2019, 38(6): 851-860.
- [21] Ayvaz, E., Kaplan, K. and Kuncan, M. (2020) An Integrated LSTM Neural Networks Approach to Sustainable Balanced Scorecard-Based Early Warning System. *IEEE Access*, **8**, 37958-37966. <https://doi.org/10.1109/access.2020.2973514>
- [22] 毛雪娇, 谈健, 姚颖蓓, 等. 基于长短期记忆神经网络的饱和负荷预测方法及应用[J]. 水电能源科学, 2019(6): 192-195.
- [23] Sagheer, A. and Kotb, M. (2019) Time Series Forecasting of Petroleum Production Using Deep LSTM Recurrent Networks. *Neurocomputing*, **323**, 203-213. <https://doi.org/10.1016/j.neucom.2018.09.082>
- [24] Karevan, Z. and Suykens, J.A.K. (2020) Transudative LSTM for Time-Series Prediction: An Application to Weather Forecasting. *Neural Networks*, **125**, 1-9. <https://doi.org/10.1016/j.neunet.2019.12.030>
- [25] Mandic, D.P. and Chambers, J. (2001) Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability. John Wiley & Sons, Inc., New York.
- [26] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Representations by Back-Propagating Errors. *Nature*, **323**, 533-536. <https://doi.org/10.7551/mitpress/1888.003.0013>
- [27] Werbos, P.J. (1988) Generalization of Back-Propagation with Application to a Recurrent Gas Market Model. *Neural Networks*, **1**, 339-356. [https://doi.org/10.1016/0893-6080\(88\)90007-X](https://doi.org/10.1016/0893-6080(88)90007-X)
- [28] Elman, J. (1990) Finding Structure in Time. *Cognitive Science*, **14**, 179-211. https://doi.org/10.1207/s15516709cog1402_1
- [29] 邱锡鹏. 神经网络与深度学习[M]//第六章 循环神经网络.
- [30] [https://baike.baidu.com/item/%E5%BE%AA%E7%8E%AF%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/23199490?fr=aladdin#reference-\[4\]-23754076-wrap](https://baike.baidu.com/item/%E5%BE%AA%E7%8E%AF%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/23199490?fr=aladdin#reference-[4]-23754076-wrap)
- [31] Bengio, Y., Simard, P. and Frasconi, P. (1994) Learning Long-Term Dependencies with Gradient Descent Is Difficult. *IEEE Transactions on Neural Networks*, **5**, 157-166. <https://doi.org/10.1109/72.279181>
- [32] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780.

-
- <https://doi.org/10.1162/neco.1997.9.8.1735>
- [33] Zia, T. and Zahid, U. (2019) Long Short-Term Memory Recurrent Neural Network Architectures for Urdu Acoustic Modeling. *International Journal of Speech Technology*, **22**, 21-30. <https://doi.org/10.1007/s10772-018-09573-7>
- [34] Kong, W., Dong, Z.Y., Jia, Y., *et al.* (2017) Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network. *IEEE Transactions on Smart Grid*, **10**, 841-851. <https://doi.org/10.1109/TSG.2017.2753802>
- [35] Kim, H.Y. and Won, C.H. (2018) Forecasting The Volatility of Stock Price Index: A Hybrid Model Integrating LSTM with Multiple GARCH-Type Models. *Expert Systems with Applications*, **103**, 25-37. <https://doi.org/10.1016/j.eswa.2018.03.002>
- [36] Oh, S.L., Ng, E.Y.K., San, T.R., *et al.* (2018) Automated Diagnosis of Arrhythmia Using Combination of CNN and LSTM Techniques with Variable Length Heart Beats. *Computers in Biology and Medicine*, **102**, 278-287. <https://doi.org/10.1016/j.compbiomed.2018.06.002>
- [37] 魏凤英, 曹鸿兴. 长期预测的数学模型及其应用[M]. 北京: 气象出版社, 1990. 9-12.
- [38] 魏生生, 曹鸿兴. 均生函数模型预报检验[J]. 气象, 1998, 24(5): 42-46.
- [39] 李文辉, 陈妍, 梁嘉雯, 等. 时间序列均生函数模型在降水量预测中的应用[J]. 广东气象, 2019, 41(4): 11-14.
- [40] 刘婷婷, 李月瑶. 时间序列均生函数模型在降水量预测中的应用[J]. 吉林水利, 2013, 371(4): 32-34.
- [41] <http://blog.csdn.net/pipisorry/article/details/78361778>
- [42] <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- [43] Abdel-Nasser, M. and Mahmoud, K. (2019) Accurate Photovoltaic Power Forecasting Models Using Deep LSTM-RNN. *Neural Computing & Applications*, **31**, 2727-2740. <https://doi.org/10.1007/s00521-017-3225-z>
- [44] Zhao, J.C., Deng, F., Cai, Y.Y., *et al.* (2019) Long Short-Term Memory-Fully Connected (LSTM-FC) Neural Network for PM2.5 Concentration Prediction. *Chemosphere*, **220**, 486-492. <https://doi.org/10.1016/j.chemosphere.2018.12.128>
- [45] Gers, F.A. and Schmidhuber, E. (2001) LSTM Recurrent Networks Learn Simple Context-free and Context-sensitive Languages. *IEEE Transactions on Neural Networks*, **12**, 1333-1340. <https://doi.org/10.1109/72.963769>