基于 PCA-ESN 模型的潘家口水库水位预测研究

龚 莎[1],彭宏玉[2]

(1. 西南交通大学 唐山研究生院,唐山 河北 063000;
2. 唐山学院 计算机科学与技术系,唐山 河北 063000)

摘要:选择河北省潘家口水库为研究对象,采用 PCA 算法对数据进行预处理,选取新的主成分作为输入变量,再通过 ESN 模型对水库水位进行预测。实验结果表明,历史水位、降水量 2 个因素的变化对水位有较大的影响;ESN 预测模型能较好地预测水位变化趋势,误差小,精度高,应用在水位预测上具有可行性和有效性。

关键词:潘家口水库;回声状态网络模型;主成分分析法;水位预测
中图分类号:TP391.97;P332.3 **文献标志码**:A **文章编号**:1672-349X(2020)03-0037-05
DOI:10.16160/j.cnki.tsxyxb.2020.03.008

Research on Water Level Forecast Based on PCA-ESN Model

GONG Sha[1], PENG Hong-yu[2]

(1. Graduate School of Tangshan, Southwest Jiaotong University; Tangshan 063000, China;
2. Department of Computer Science and Technology, Tangshan University, Tanghan 063000, China)

Abstract: With Panjiakou Reservoir in Hebei Province as the research object, the PCA algorithm is used to preprocess the data. Then, the new principal component is selected as the input variable, and the water level of this reservoir is forecast through the PCA-ESN model. The experimental results show that the two factors of historical water level and the rainfall have a great impact on the water level. The ESN model can better forecast the trend of the water level changes with small errors and high accuracy, which is feasible and effective in the forecast of water level.

Key Words: Panjiakou Reservoir; Echo State Network(ESN) model; Principal Component Analysis(PCA); water level forecast

0 引言

滦河来水量在时间分布上很不均匀,一年之内的来水量主要集中在七、八、九三个月内,且来水量的年际变化悬殊[1]。潘家口水库是整个引滦工程的源头,对其水位进行预测,对于提高滦河水资源的利用率、加强水资源的分配管理以及防洪减灾都具有重要意义。

目前,很多学者致力于湖泊流域水位预测方法的研究,采用的技术主要有遥感技术、物联网技术、机器学习和神经网络。其中应用机器

基金项目:西南交通大学合作智慧水务项目(1200305);唐山市室内定位重点实验室建设项目(210050202)

作者简介:龚莎(1995—),女,四川江油人,硕士研究生,主要从事数据挖掘研究。

学习和神经网络进行预测在国内外已有不少研究成果。Makhtar 等[2]使用 Apriori 算法从数据集中生成最佳规则,用于查找频繁项集,通过建立洪水预报模型来发现水位与洪水面积之间的相关性,这项研究的结果证明了在水位预测中 Apriori 算法的可用性。Jangyodsuk 等[3]基于贝叶斯的方法,提出了一种新的因果发现算法,利用降水和水文数据来寻找未来洪水的影响特征,但此研究只考虑了降水这单一因素对洪水的影响。赵春雷等[4]采取历史资料回归和机器学习方法,对白洋淀水位随区域降水量变化的规律进行分析,并通过建立最低水位预测模型对已有的数据进行验证。但此研究只考虑了影响水位的雨季自然降水量和白洋淀基础水位这两个因素,造成一定的结果误差,预测精度有待提高。刘亚新等[5]提出了一种基于长短时记忆(Long Short-Term Memory,LSTM)的水位预测方法,用于葛洲坝水电站上下游水位的预测,此方法采用水位和出力等直接监测数据,避免了出入库流量等间接计算值带来的二次误差,但预测精度仍有待进一步提高。Ghorbani 等[6]将混合模型的预测能力与 FFA 集成在一起,作为具有多层感知器(MLP-FFA)的启发式优化工具,用于土耳其埃利迪尔湖水位的预测,实验表明 Firefly 算法作为优化器,可以使模型预测的准确性更高。Khan 等[7]为拉姆甘加河开发了一个人工神经网络模型,使用日常用水对模型化网络进行训练、验证和测试,由于河流流量和水位值难以测量,不能直接预测隐藏层神经元最佳数量,必须通过枚举技术获得最佳网络拓扑结构,计算成本较高,增加了运行时间。Xu 等[8]提出了一种基于 ARIMA-RNN 组合模型的水位预测方案,解决了单个预测模型不能同时考虑数据中线性和非线性成分的问题,实验结果证明,预测模型可以取得较好的效果,但精度仍有待提高。

针对应用传统神经网络进行水位预测存在的问题,本文提出了一种基于 PCA-ESN 模型的水位预测方法。首先,在数据预处理部分利用主成分分析法(Principal Component Analysis,PCA)有效提取多元时间序列数据的特征并对原始数据进行重组,降低了水位数据的信息冗余;其次,使用回声状态网络(Echo State Network,ESN)建立水位预测模型,克服了传统递归神经网络梯度消失和梯度爆炸问题,网络具有较好的信息处理能力和泛化能力。具体过程为,根据河北省潘家口水库的水位日数据,先通过 PCA 算法选取影响水位的相关变量,然后设计 ESN 模型预测水位。

1 数据预处理

对数据预处理是实验结果准确的前提。作为多元统计中常用的数据分析方法之一,主成分分析法能够在降低原始数据变量维数的同时有效提取各个变量的特征,产生新成分,新成分能够克服因原始变量信息重叠而对数据分析结果造成的不良影响[9]。

设原始数据集包括 n 个数据样本,每个样本具有 p 个变量,对此数据集的主成分分析计算流程如下。

(a) 对原始数据集进行标准化处理,组成标准化数据矩阵 $\boldsymbol{Z}$。

(b) 引入 Pearson 相关系数(式(1))计算各个变量数据间的相关性,组成相关系数矩阵 $\boldsymbol{R}$。

$$r_{ij}=\frac{\sum_{k=1}^{n}(X_{ki}-\overline{X}_i)(X_{kj}-\overline{X}_j)}{\sqrt{\sum_{k=1}^{n}(X_{ki}-\overline{X}_i)^2\sum_{k=1}^{n}(X_{kj}-\overline{X}_j)^2}},\quad(1)$$

其中,$\sum_{k=1}^{n}(X_{ki}-\overline{X}_i)^2$ 表示数据 i 的离均差平方和,$\sum_{k=1}^{n}(X_{kj}-\overline{X}_j)^2$ 表示 j 的离均差平方和,$\sum_{k=1}^{n}(X_{ki}-\overline{X}_i)(X_{kj}-\overline{X}_j)$ 表示 i 与 j 间的离均差积和。

(c) 求解相关系数矩阵 $\boldsymbol{R}$ 的特征方程,对求出的特征值按从大到小的顺序进行排序 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$,并求出每个特征值相对应的单位特征向量 l_i,l_i 组成如式(2)的主成分得分矩阵 $\boldsymbol{L}$。

$$\boldsymbol{L}=\begin{bmatrix}l_{11}&\cdots&l_{1p}\\ \vdots&\ddots&\vdots\\ l_{n1}&\cdots&l_{np}\end{bmatrix}=(L_1,L_2,\cdots,L_p)^{\mathrm{T}},\quad(2)$$

其中，$l_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,p)$。

(d) 根据式(3) 计算累计贡献率 k_i，保留累计贡献率在 85% 以上的前 m 个成分作为新主成分。

$$k_i=\sum_{i=1}^{m}\lambda_i/\sum_{i=1}^{p}\lambda_i。\tag{3}$$

(e) 根据第(c) 步、第(d) 步计算新主成分 Y 的各个主成分，得到原始数据集经过 PCA 处理后的重组数据集，第 m 个新主成分的数学模型 Y_m 如式(4) 所示。

$$Y_m=l_{1m}\times X_1+l_{2m}\times X_2+\cdots+l_{mn}\times X_p。\tag{4}$$

原始数据为河北省潘家口水库 2009 年 9 月 3 日至 2019 年 12 月 29 日的 3 770 组水位日数据，数据来源于实验室搭建的野外人工湖水位智慧预警系统的数据采集部分。该系统的数据采集部分比较完善，经过了长时间的运行，已经采集到定量的水位及相关信息。部分历史水位及相关信息数据如表 1 所示。

表 1　部分历史水位及相关信息数据

日期	压力 /Mpa	水温 /℃	湿度 (%)	降水量 /mm	水位 /m
2019/7/1	0.02	20	26	0	178.5
2019/7/2	0.02	20	31	4.9	178.6
2019/7/3	0.01	20	28	0	178.6
2019/7/4	0.01	20	20	0	178.6
2019/7/5	0.01	20	30	6.4	178.7
2019/7/6	0.02	19	78	77.7	179.1
2019/7/7	0.01	19	69	7.2	179.2
2019/7/8	0.01	19	47	0	179.2
2019/7/9	0.02	19	68	16.2	179.3
2019/7/10	0.02	18	53	0.9	179.3
2019/7/11	0.02	18	54	0.9	179.4
2019/7/12	0	18	42	0.1	179.4
2019/7/13	0	18	44	1.4	179.5
2019/7/14	0	18	52	22.3	179.6
2019/7/15	0.01	18	45	0	179.7
2019/7/16	0.01	18	49	0	179.8
2019/7/17	0.02	19	68	2.7	179.8
2019/7/18	0.02	19	51	2.6	179.9
2019/7/19	0.02	19	48	0	179.9
2019/7/20	0.02	19	62	0	180

在进行 ESN 模型构建之前，要把 ESN 结构中输入层因子的具体数值进行预处理，输入层因子的选择和主成分分析相关性很强，无论是因子初选还有因子精选都是其范围内的，因子选择过程本质上来说就是主成分分析过程。

将历史水位、降水量、水温、湿度、压力作为主成分分析过程的备选因子。数据集的变量分别为历史水位(X_1)、降水量(X_2)、湿度(X_3)、水温(X_4)、压力(X_5)。对 5 个初始因子进行主成分分析，结果如表 2 所示。

表 2　PCA 分析结果表

成分	数据		
	特征值	贡献率(%)	累计贡献率(%)
PA_1	5.634	68.932	68.932
PA_2	1.966	24.054	92.986
PA_3	0.477	5.841	98.827
PA_4	0.053	0.646	99.473
PA_5	0.013	0.163	100

经主成分分析，得到各个主成分的特征值、贡献率、累积贡献率。由表 2 可知，历史水位和降水量 2 个主成分的累计贡献率已经达到 92.986%，说明这 2 个主成分取代先前的 5 个因子可以让信息丢失的程度降低，主成分的分析效果更稳定。

数据集经过 PCA 处理得到 2 个新主成分 PA_1，PA_2，由主成分得分矩阵计算得到新主成分的模型如式(5) 所示：

$$\begin{cases}PA_1=0.021X_1+0.538X_2+\\0.130X_3+0.013X_4+0.832X_5\\PA_2=0.014X_1+0.833X_2-\\0.065X_3-0.011X_4-0.549X_5\end{cases}。\tag{5}$$

2　模型设计

回声状态网络 ESN 由 2001 年 Jaeger 提出[10]，是典型的储备池计算网络，具有复杂的动力学特征。目前 ESN 已经在智能控制、语音识别、非线性时间序列预测等领域取得了广泛的应用。ESN 主要由输入层、储备层和输出层组成，其特点是储备层由一个包含大量神经元的动态储备池(DR) 构成，储备池内的神经元采用随机、稀疏的方式连接，其蕴含了网络的运行状态，具有短期记忆功能。储备池是回声状态网络结构的核心部分，对于网络最终的性能起着至关重要的作用。ESN 结构如图 1 所示。

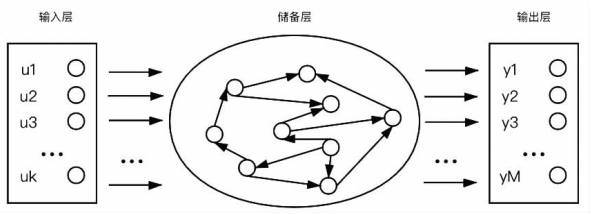

图 1 ESN 结构图

设输入矩阵、状态矩阵、输出矩阵分别为：

$$\boldsymbol{u}(t) = (u_1(t), u_2(t), \cdots, u_K(t))^{\mathrm{T}}, \quad (6)$$

$$\boldsymbol{x}(t) = (x_1(t), x_2(t), \cdots, x_N(t))^{\mathrm{T}}, \quad (7)$$

$$\boldsymbol{y}(t) = (y_1(t), y_2(t), \cdots, y_M(t))^{\mathrm{T}}, \quad (8)$$

其中，K 为输入维数，N 为储备池内部神经元个数，M 为输出维数，$t = 1, 2, \cdots, T$。

状态矩阵的更新如式(9) 所示：

$$x(t+1) = f(\boldsymbol{W}_{in}u(t+1) + \boldsymbol{W}x(t)), \quad (9)$$

$$y(t) = f_{out}(\boldsymbol{W}_{out}x_{out}(t)), \quad (10)$$

其中，$\boldsymbol{W}_{in}$ 是 $N \times K$ 输入层到储备层的权重矩阵，$\boldsymbol{W}$ 是储备层权重矩阵，$\boldsymbol{W}_{out}$ 是 $M \times K$ 储备层到输出层的权重矩阵，此矩阵的更新如式(11) 所示：

$$\boldsymbol{W}_{out} = ((\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S} + \beta\boldsymbol{I})^{-1}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{D})^{\mathrm{T}}, \quad (11)$$

其中，β 表示非负正则化系数，$\boldsymbol{S}$ 表示全部状态矩阵，$\boldsymbol{I}$ 表示单位矩阵，$\boldsymbol{D}$ 为全部输出矩阵。

3 算法设计

在设计算法前首先需要进行关键参数选择，具体参数包括储备池规模、稀疏度、谱半径、输入缩放因子。

(ⅰ) 储备池规模 N 是指 ESN 储备池内神经元的个数。储备池是随机生成的，其规模必须足够大，以捕捉潜在的数据特征。一般来说，如果采取适当的正则化措施，则储备池规模越大，获得的网络性能就越好。但过大也会导致“过拟合”。文中选取 N 为 500。

(ⅱ) 稀疏度 s 是指储备池中存在相互连接的神经元个数与神经元总个数的百分比，反映了储备池神经元间连接的稀疏程度。ESN 储备池内神经元是稀疏连接的，即连接输入层与储备层的输入权重矩阵 $\boldsymbol{W}_{in}$ 中大部分元素值为 0。文中设定稀疏度为 5%。

(ⅲ) 谱半径 $\rho(\boldsymbol{W})$ 是指储备层权重矩阵 $\boldsymbol{W}$ 的特征值绝对值中的最大值。当谱半径介于 [0,1] 之间时，回声状态网络具有回声状态特性。但由于激活函数引入非线性因素，最佳 $\rho(\boldsymbol{W})$ 值有时可能会比 1 大得多，意味着 $\rho(\boldsymbol{W}) < 1$ 并不是网络具有回声状态特性的必要条件。因此，实际任务中的谱半径应更大，需要更大的输入存储空间。

(ⅳ) 输入缩放因子是指在输入信号传送到网络储备池前，对输入权重矩阵 $\boldsymbol{W}_{in}$ 进行尺度变换的一个缩放因子。输入缩放因子的大小与网络处理问题的非线性程度有关，非线性程度越强，输入缩放因子值越大。

ESN 模型训练：

Step1：进行初始化操作，先确定储备池的规模，即神经元的个数。

Step2：随机生成输入权重矩阵 $\boldsymbol{W}_{in}$ 和储备层权重矩阵 $\boldsymbol{W}$。调整输入缩放因子，使 $\rho(\boldsymbol{W})$ 谱半径小于 1。

Step3：样本数据依次加载到输入、输出，更新储备池内部状态。

Step4：前 n 个数据因受到初始瞬变的影响，所以删除 $x(1)$ 至 $x(n)$，即前 n 个数据不用于学习 $\boldsymbol{W}_{out}$，并收集第 n 个数据以后的状态变量。

Step5：计算输出权重矩阵 $\boldsymbol{W}_{out}$。

Step6：用新输入和训练好的 $\boldsymbol{W}_{out}$ 计算相应输出进行测试。

4 仿真实验

将基于 PCA-ESN 的预测模型用于潘家口水库水位的预测。

首先对输入数据进行归一化处理：

$$x' = (x_y - x_{\min})/(x_{\max} - x_{\min}), \quad (12)$$

其中，x_y 为数据序列的原始值，x' 为归一化后的值，$x_{\min}$ 为序列的最小值，$x_{\max}$ 为序列的最大值。

将 3 770 组数据归一化处理后，利用 PCA 算法对数据进行分析处理，选取新主成分。取前 2 000 组数据用作训练样本，后面 1 770 组数据用作测试样本。ESN 网络的输入节点为 2，储备池神经元个数为 500，稀疏度为 5%，输出节点数为 1。

基于处理的数据集，对 PCA-ESN 模型进

行水位预测的实验模拟。将模型运行 30 次以求得较为稳定的模拟结果，采用 30 次的平均值绘制 ESN 测试效果图图 2。由图 2 可知，ESN 的测试输出接近真实水位数据，效果较好。

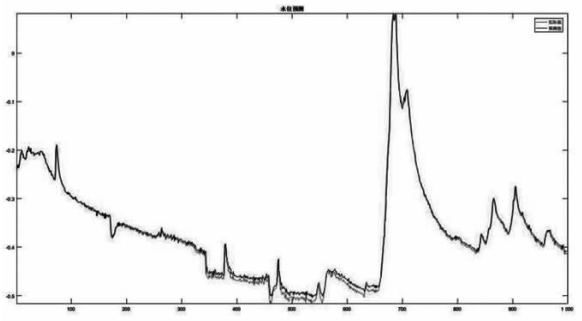

图 2　ESN 测试效果图

为了进一步说明 ESN 的优越性，本实验选取均方误差（*MSE*）作为模型预测性能的评价指标，计算公式如下：

$$MSE = \frac{\sum_{t=1}^{T}(y(t) - y_d(t))^2}{T}。\quad (13)$$

ESN 误差测试如图 3 所示。由于 ESN 模型的输入权重矩阵与储备层权重矩阵在每次训练时均需随机生成，因此 ESN 模型的误差分布有一定波动，但基本比较平稳，误差控制在 7.278E－05 左右，预测效果较好。说明 ESN 的神经网络数据处理能力强，训练效果好。

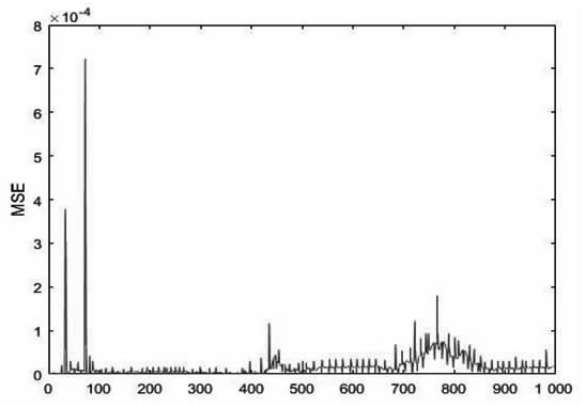

图 3　ESN 误差测试图

5　结论

采用 PCA 算法对数据集进行预处理，选取新的主成分作为输入变量，建立 ESN 预测模型，对潘家口水库水位进行预测。仿真实验结果表明，采用 PCA 算法对数据进行预处理，提取到历史水位和降水量的累计贡献率达到 92.986%，将这 2 个主成分作为输入变量，减少了重复率，降低了复杂度。实验展示了 ESN 预测非线性时间序列数据的优势，收到了较好的预测效果。因此，基于 PCA-ESN 模型预测水位的误差小，准确性高，具有一定的实用价值。

参考文献：

[1] 姚德贵，郭旭峰. 滦河潘家口、大黑汀水库水污染诊断与治理对策[J]. 中国防汛抗旱，2020，30(4)：48－51.

[2] MAKHTAR M，HARUN N A，AZIZ A A，et al. An association rule mining approach in predicting flood areas[J]. Advances in Intelligent Systems and Computing，2016(549)：437－446.

[3] JANGYODSUK P，SEO D，ELMASRI R，et al. Flood prediction and mining influential spatial features on future flood with causal discovery[C]. 2015 IEEE International Conference on Data Mining Workshop(ICDMW)，Atlantic City，NJ，2015：1462－1469.

[4] 赵春雷，钱拴，黄强，等. 基于降水量的白洋淀最低水位预测研究[J]. 中国生态农业学报，2019，27(8)：1238－1244.

[5] 刘亚新，樊启祥，尚毅梓，等. 基于 LSTM 神经网络的水电站短期水位预测方法[J]. 水利水电科技进展，2019，39(2)：56－60.

[6] GHORBANI M A，DEO R C，KARIMI V，et al. Implementation of a hybrid MLP-FFA model for water level prediction of Lake Egirdir，Turkey[J]. Stoch Environ Res Risk Assess，2018(32)：1683－1697.

[7] KHAN M Y A，HASAN F，PANWAR S，et al. Neural network model for discharge and water-level prediction for Ramganga River catchment of Ganga Basin，India[J]. Hydrological Sciences Journal，2016，61(11)：2084－2095.　（下转第 67 页）

车方向。机动车通行高峰时段的仿真实验结果显示，与动态连续车道管理方法和静态车道管理方法相比，复合动态车道管理方法能够改善道路交叉路口的整体性能，能够更加均衡地利用干道双向的道路空间，这种性能优势随着交叉路口机动车交通负荷的提高而愈发显著，而且结合现有道路上的直行待行区，可以增加车辆的通行效率，缓解交通拥堵。

参考文献：

[1] 张卓然. BP 神经网络和自适应模糊推理系统在多传感器信息融合系统中的研究[D]. 武汉：武汉工业学院，2012.

[2] 韩鹏. 基于短时交通流预测交叉口优化配时研究[D]. 合肥：中国科学技术大学，2017.

[3] 唐凡. 基于神经网络的城市交通流预测研究[J]. 四川水泥，2019(4)：335.

[4] 胡枫. 基于马尔科夫模型的短时交通流预测研究[D]. 南京：南京邮电大学，2013.

[5] 杨慧慧. 城市交通流短时预测模型研究[D]. 焦作：河南理工大学，2015.

（责任编校：夏玉玲）

（上接第 26 页）

[4] 薛雅丽. 基于 STM32W108 的粮食储备系统粮情监测[J]. 唐山学院学报，2016，29(3)：35－37.

[5] 任丽棉. 基于 STM32 的电力谐波测试仪的设计[J]. 唐山学院学报，2014，27(6)：67－68.

[6] 范大勇. 基于 STM32 的低成本高精度电能测量装置设计[J]. 传感器与微系统，2019(10)：85－89.

[7] 戴捷，胡晓吉. 基于光传输的 USB 键盘鼠标一体化设计[J]. 计算机工程与设计，2012，33(7)：2620－2627.

[8] 刘春林. 基于 USB 协议键盘信息快捷输入技术研究[D]. 哈尔滨：哈尔滨工业大学，2017.

[9] 滕鹏，姜昌华，王春慧，等. 基于 HID 类的 USB 手势输入系统设计与实现[J]. 数字技术与应用，2017(2)：180－181.

（责任编校：李秀荣）

（上接第 41 页）

[8] XU G，CHENG Y，LIU F et al. A water level prediction model based on ARIMA-RNN[C]. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)，Newark，CA，USA，2019：221－226.

[9] 韩敏，王亚楠. 基于储备池主成分分析的多元时间序列预测研究[J]. 控制与决策，2009，24(10)：1526－1530.

[10] JAEGER H. The “echo state” approach to analysing and training recurrent neural networks with an Erratum note[R]. Bonn：German National Research Center for Information Technology，2001.

（责任编校：李秀荣）