

基于信息熵与改进极限学习机的中长期径流预测

岳兆新^{1,2},艾萍^{1,3},熊传圣³,宋艳红¹,洪敏³,于家瑞³

(1. 河海大学计算机与信息学院,江苏 南京 211100; 2. 南京工业职业技术大学计算机与软件学院,江苏 南京 210023;
3. 河海大学水文水资源学院,江苏 南京 210098)

摘要:为提高流域中长期径流预测精度,提出一种基于信息熵与改进极限学习机的中长期径流预测方法。首先,基于不同水文站点的流域控制面积构造径流综合指数,在较宏观层面表征流域水情丰枯变化;其次,采用偏互信息法计算影响对象与径流综合指数之间的相关性,获得径流过程变化的关键因子集,形成预测模型输入;最后,结合K折交叉验证与改进粒子群算法优化极限学习机(ELM)参数,构建IPSO-ELM模型,用于中长期径流预测。以雅砻江流域为例,将所建模型与BP神经网络(BPNN)、支持向量机(SVM)、ELM和PSO-ELM等预测模型进行对比分析。结果表明:所提模型的 E_{mape} 、 E_{rmse} 、 E_{dc} 、 E_{qr} 和 E_{re} 等性能评价指标明显优于上述4种模型;5种预测模型在D1数据集上的预测效果整体上胜于D2。

关键词:中长期径流预测;径流综合指数;偏互信息法;粒子群;极限学习机;预测模型

中图分类号:TV124 **文献标志码:**A **文章编号:**1006-7647(2021)04-0007-08

Medium-long term runoff forecasting based on information entropy and improved extreme learning machine//YUE Zhaoxin^{1,2}, AI Ping^{1,3}, XIONG Chuansheng³, SONG Yanhong¹, HONG Min³, YU Jiarui³ (1. College of Computer and Information, Hohai University, Nanjing 211100, China; 2. School of Computer and Software, Nanjing Vocational University of Industry Technology, Nanjing 210023, China; 3. College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China)

Abstract: To improve the accuracy of the medium-long term runoff forecasting of the whole watershed, a combined method integrating information entropy and improved extreme learning machine (ELM) is proposed. Firstly, the comprehensive runoff index is constructed based on the controlling area of different hydrological stations to characterize the abundance and drought in the basin. Secondly, the partial mutual information (PMI) approach is applied to calculate the correlation between multiple factors and the comprehensive runoff index for inputs of the forecasting model. Finally, an improved ELM model by combining K-fold cross validation with improved particle swarm optimization (IPSO) is proposed to optimize parameters of ELM, together denoted as IPSO-ELM, for medium-long term forecasting. In a case study of the Yalong River basin, the proposed model was compared with classical forecasting models, i. e., backpropagation neural networks (BPNN), support vector machines (SVM), ELM and PSO-ELM models. The results shows that the performance evaluation indexes of the proposed model perform much better than the above four data-driven models in terms of E_{mape} , E_{rmse} , E_{dc} , E_{qr} , and E_{re} . The five forecasting models demonstrate better results for the D1 dataset compared to that of the D2 dataset.

Key words: medium-long term runoff forecasting; comprehensive runoff index; partial mutual information; particle swarm; extreme learning machine; forecasting model

水文中长期预报是指基于水文现象演化的客观规律,根据前期和历史水文、气象等信息,运用成因分析和数学建模等方法,对未来较长时间内水文情势做出定性或定量预报^[1-2]。及时、准确的中长期水文预报可为水资源高效利用、水利工程建设与运行,以及防汛抗旱指挥决策等提供重要的基础数据和科

学的决策依据。当前,中长期水文预报仍然处于探索、发展阶段,预报精度还不能满足各生产部门的实际需求。中长期径流预测是中长期水文预报中的一个重要研究方向,也是水信息学科研究与应用的重要难题之一。

目前,常用的中长期径流预测方法主要有成因

分析方法、统计学方法、基于智能算法的预测方法和基于数值天气预报的预测方法四大类。其中,成因分析方法^[3]和统计学方法^[4-5]是水文学科的典型方法,具有一定的适用范畴,但也存在诸多需要研究的问题。比如影响径流序列长期变化的物理成因复杂,难以完全掌握其客观规律。而统计学方法多以线性方法为主,难以适应径流变化影响要素的复杂非线性特性,具有一定的局限性。基于智能算法和数值天气预报的综合预测方法是近些年发展起来的新方法,是伴随计算机信息技术的发展和新数学建模方法的涌现而发展起来的新技术。前者具有较好的非线性映射、泛化和容错能力,被广泛应用于径流预测领域^[6-7];后者则在水文预报中耦合一定预见期内的数值天气预报产品,在探索增长径流预报预见期方面,具有一定的研究意义^[8]。基于智能算法的中长期径流预测模型主要根据输入输出变量之间的函数关系构建基于人工神经网络、支持向量机(support vector machine, SVM)、小波分析等预测模型或者综合采用多个模型对未来中长期径流进行预测分析,并取得了诸多成果^[9-13]。尽管上述基于智能算法的预测模型应用广泛,但模型结构相对复杂,参数在训练过程中需要初始化以及不断优化调整,效率相对较低,且 BP 神经网络(backpropagation neural networks, BPNN)采用基于梯度下降的方法,容易陷入局部最小值问题,算法需要多次迭代,因而整体效率不高。极限学习机^[14](extreme learning machine, ELM)是一种单隐层前馈神经网络,具有参数设置简单、计算速度快、误差小、泛化能力强等优点,被广泛应用于故障诊断、图像处理等领域。

此外,影响径流过程变化的关键因子筛选也是中长期径流预测需要研究的重要内容。因子筛选方法主要包括先验知识法、相关系数法、主成分分析法和信息熵法^[15-17]。先验知识法主要依赖于人工经验,主观性较强,具有一定的局限性。相关系数法和主成分分析法,整体上属于线性方法,难以适应中长期径流过程影响因子的复杂非线性特性,具有一定的适用范围。信息熵法,尤其是互信息方法忽略了变量分布,适用于备选因子间的线性和非线性相关关系。偏互信息方法是在互信息方法基础上的改进,可以有效避免对已入选因子的影响,减少冗余变量,降低计算复杂度。

鉴于此,本文提出一种基于信息熵与改进极限学习机的中长期径流预测方法。首先,基于不同水文站点的流域控制面积构造径流综合指数,在较宏观层面表征流域水情丰枯变化;其次,采用偏互信息方法计算影响对象与径流综合指数之间的相关性,

获得径流过程变化的关键因子集,形成预测模型的输入;最后,结合 K 折交叉验证与改进粒子群算法优化 ELM 参数,构建 IPSO-ELM(improved particle swarm optimization, IPSO-ELM)模型,用于中长期径流预测。

1 基于信息熵与改进极限学习机的中长期径流预测方法

1.1 径流综合指数构造

粒计算作为人工智能研究领域中的一种新理念方法,其目的是在问题的求解过程中,用粒度合适的“粒”作为处理对象,从而在保证求得满意解的前提下,提高解决问题的效率^[18]。当前中长期径流预测,主要基于流域内某些典型站点(单一粒度)的径流变化来预测中长期径流变化情势,并没有结合研究区域内多个水文时空对象进行多粒度综合分析,导致复杂环境下流域中长期径流预测的宏观研判能力不足。因此,本文依据粒计算理论,通过研究分析流域中不同水文站点(细粒度)月平均径流的一致性,构造描述流域月均径流丰枯情况的流域径流综合指数(粗粒度),在较宏观的层面研究整个流域的径流变化情势,以提高流域中长期径流预测的准确性和可靠性。以此为基础,本文通过研究流域内多个不同水文站点月平均径流的一致性,构造具有多粒度特性的流域径流综合指数(comprehensive runoff index, COM),以表征流域径流的丰枯情况。为了避免测站之间的累积影响,更客观地使用各测站描述流域径流丰枯情况,本文在确定所选水文站点的权重时,采用削减流域下游水文站对整体指数的贡献度原则,基于站点流域控制面积构建流域径流综合指数。

假设流域站点数量为 n_s ,第 i 个水文站的控制面积百分比为 S_i ,第 i 个水文站,第 j 个月的月平均径流为 c_{ij} ,则第 i 个水文站的权重 w_i 和第 j 个月的径流综合指数 c_j 分别为

$$w_i = \frac{1/S_i}{\sum_{j=1}^{n_s} 1/S_j} \quad (1)$$

$$c_j = \sum_{i=1}^{n_s} w_i c_{ij} \quad (2)$$

1.2 基于信息熵的因子筛选

互信息(mutual information, MI)以信息熵理论为基础,既能够度量输入变量与预测对象间的线性和非线性关系,也能够度量一个变量中含有的关于另一个变量的信息量^[19]。但采用 MI 对输入变量进行筛选时,由于输入变量之间的耦合关系会对 MI

的计算结果产生影响,从而导致误选或漏选。因此, May 等^[20-21]提出了偏互信息法 (partial mutual information, PMI),通过计算条件期望消除了变量之间的联系,从而保证了变量选择的可靠性和准确性。偏互信息定义为

$$P_{MI} = \iint f_{X',Y'}(x',y') \ln \frac{f_{X',Y'}(x',y')}{f_{X'}(x')f_{Y'}(y')} dx' dy' \quad (3)$$

$$x' = x - E[x | z] \quad (4)$$

$$y' = y - E[y | z] \quad (5)$$

式中: E 为期望值; x 为备选输入因子; y 为预测对象。

给定 N 个离散样本,偏互信息可采用如下离散形式定义:

$$P_{MI} = \frac{1}{N} \sum_{i=1}^N \ln \frac{f_{X',Y'}(x'_i, y'_i)}{f_{X'}(x'_i) f_{Y'}(y'_i)} \quad (6)$$

1.3 IPSO-ELM 预测模型构建

1.3.1 极限学习机

假设给定任意 N 个不同样本 $(\mathbf{X}_i, \mathbf{t}_i)$ 。其中, $\mathbf{X}_i = (x_{i1}, x_{i2}, \dots, x_{in})^T \in \mathbf{R}_n$, $\mathbf{t}_i = (t_{i1}, t_{i2}, \dots, t_{im})^T \in \mathbf{R}_m$, 目标函数定义如下:

$$\sum_{i=1}^C \beta_j g(\mathbf{W}_i \cdot \mathbf{X}_j + b_i) = o_j \quad (j = 1, 2, \dots, N) \quad (7)$$

式中: $g(x)$ 为激活函数; $\mathbf{W}_i$ 为输入层与隐含层之间的权重矩阵, $\mathbf{W}_i = (w_{i1}, w_{i2}, \dots, w_{in})^T$; β_i 为隐含层与输出层之间的输出权重矩阵, $\beta_i = (\beta_{i1}, \beta_{i2}, \dots, \beta_{im})^T$; b_i 为第 i 个隐含层神经元的偏置; o_j 为第 j 个样本的网络输出值; C 为隐含层神经元个数。

预测值与真实值误差最小,可表示为

$$\sum_{j=1}^C \|o_j - t_j\| = 0 \quad (8)$$

也就是存在 $\beta_i, b_i, \mathbf{W}_i$ 使得:

$$\sum_{i=1}^C \beta_i g(\mathbf{W}_i \cdot \mathbf{X}_j + b_i) = t_j \quad (j = 1, 2, \dots, N) \quad (9)$$

用矩阵表示为

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{T} \quad (10)$$

其中 $\mathbf{H}(\mathbf{W}_1, \dots, \mathbf{W}_C, b_1, \dots, b_C, \mathbf{X}_1, \dots, \mathbf{X}_C) =$

$$\begin{bmatrix} g(\mathbf{W}_1 \cdot \mathbf{X}_1 + b_1) & g(\mathbf{W}_2 \cdot \mathbf{X}_1 + b_2) & \dots & g(\mathbf{W}_C \cdot \mathbf{X}_1 + b_C) \\ g(\mathbf{W}_1 \cdot \mathbf{X}_2 + b_1) & g(\mathbf{W}_2 \cdot \mathbf{X}_2 + b_2) & \dots & g(\mathbf{W}_C \cdot \mathbf{X}_2 + b_C) \\ \vdots & \vdots & \ddots & \vdots \\ g(\mathbf{W}_1 \cdot \mathbf{X}_N + b_1) & g(\mathbf{W}_2 \cdot \mathbf{X}_N + b_2) & \dots & g(\mathbf{W}_C \cdot \mathbf{X}_N + b_C) \end{bmatrix}_{N \times C}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_C^T \end{bmatrix}_{C \times m} \quad \mathbf{T} = \begin{bmatrix} T_1^T \\ \vdots \\ T_N^T \end{bmatrix}_{N \times m}$$

式中: $\mathbf{H}$ 为隐层节点的输出; $\boldsymbol{\beta}$ 为输出权重; $\mathbf{T}$ 为期望输出。

望输出。

一般 $\mathbf{W}_i$ 和 b_i 随机给定,权重估计 $\hat{\boldsymbol{\beta}} = \mathbf{H}^+ \mathbf{T}$, $\mathbf{H}^+$ 为矩阵 $\mathbf{H}$ 的广义逆。

根据上述原理,ELM 模型的算法步骤为:①随机赋值输入权重 $\mathbf{W}_i$ 和隐含层偏置 $b_i, i=1, 2, \dots, C$; ②计算隐含层输出矩阵 $\mathbf{H}$; ③计算输出层权重 $\hat{\boldsymbol{\beta}}$ 。

1.3.2 结合 K 折交叉验证与改进粒子群的 ELM 参数优化方法

粒子群 (particle swarm optimization, PSO) 算法具有算法简单、收敛速度快、可调参数少、寻优能力强等优势,但是还存在一些不足^[22]: 随机产生初始位置,导致部分粒子位置距离最优解较远,影响了计算效率;参数设定较大时容易错过最优解,导致算法不收敛,或者其他粒子可能错过最优解,进而影响收敛速度和精度;可能出现“早熟”现象,导致局部极值点的出现。为避免出现上述情形,本文首先对 PSO 算法进行参数改进和变异操作,再以此为基础,提出结合 K 折交叉验证 (K-fold cross validation, K-CV) 与 IPSO 的 ELM 参数优化方法。具体步骤主要包括参数初始化、IPSO 适应度函数选取、个体极值与群体极值的迭代更新和 ELM 最优参数生成。具体如下:

步骤 1: 初始化。给定训练样本 $[x_i, y_i]$ ($x_i \in \mathbf{R}_{n_{ip}}, n_{ip}$ 为输入神经元个数, $i=1, 2, \dots, N_{ip}$, N_{ip} 为训练样本个数), 确定激励函数,并设置隐含层节点数 C 。初始化 $N_{p,ip}$ 个维数为 D 的参数向量 $\mathbf{t}_{r,g}$ ($r=1, 2, \dots, N_{p,ip}$), 其中任意一维的取值范围为 $[-1, 1]$, g 表示迭代次数, $D=C(n_{ip}+1)$ 。

粒子种群个体 $\mathbf{t}$ 由极限学习机的输入权值向量 $\mathbf{a}=(a_1, a_2, \dots, a_c)$ 和隐含层偏置矩阵 $\mathbf{d}$ 组成, $\mathbf{t}=(a_{11}, a_{12}, \dots, a_{1n_{ip}}, a_{21}, a_{22}, \dots, a_{2n_{ip}}, \dots, a_{c1}, a_{c2}, \dots, a_{cn_{ip}}, d_1, d_2, \dots, d_c)$ 。对于每个种群个体 $\mathbf{t}_{r,g}$, 计算隐含层输出矩阵 $\mathbf{H}$, 并计算输出权重 $\boldsymbol{\beta}$ 。

步骤 2: 适应度函数选取。计算 10-CV 的均方根误差作为 IPSO 的适应度,寻找平均均方根误差最小的个体。

步骤 3: 迭代更新。更新位置 x_i 和速度 v_j , 同时引入变异算子,在粒子更新之前有一定的概率初始化粒子速度和位置,计算适应度值,更新粒子的个体极值和群体极值。

步骤 4: 获得 ELM 最优参数。判断是否达到终止条件 (达到适应度值预设精度或满足最小误差值或最大迭代次数), 则停止迭代,获得 ELM 最优参数组合; 否则,回到步骤 3。

1.4 评价指标

本文选用水文预报领域常用的评价指标,包括平

均绝对百分比误差 E_{mape} 、均方根误差 E_{rmse} 、确定性系数 E_{dc} 、相对误差 E_{re} 和合格率 E_{qr} 5 种,以此综合评价预测模型的性能。各项评价指标的计算公式如下:

$$E_{\text{mape}} = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \tag{11}$$

$$E_{\text{rmse}} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \tag{12}$$

$$E_{\text{dc}} = 1 - \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (y_t - \bar{y}_t)^2} \tag{13}$$

$$E_{\text{re}} = \left| \frac{y_t - \hat{y}_t}{y_t} \right| \times 100\% \tag{14}$$

$$E_{\text{qr}} = \frac{m}{n} \times 100\% \tag{15}$$

式中: y_t 为第 t 个时刻的流域径流综合指数观测值; $\hat{y}_t$ 为第 t 个时刻的流域径流综合指数预测值; $\bar{y}_t$ 为流域径流综合指数观测均值; m 为预报合格次数。

根据 GBT 22482—2008《水文情报预报规范》,中长期径流预报的相对误差小于 20% 为合格。

2 实例验证

2.1 研究区域与数据资料

雅砻江流域位于青藏高原东部,地理位置界于 $96^{\circ}52' \sim 102^{\circ}48' \text{E}$ 、 $26^{\circ}32' \sim 33^{\circ}58' \text{N}$ 之间,北以巴颜喀拉山与黄河分水,东以大雪山与大渡河分界,西以雀儿山、沙鲁里山与金沙江上段相邻,南接滇东北高原的金沙江谷地。流域东、北、西三面大部分为海拔 4 000 m 以上的高山包围,其主峰均在 5 500 m 以上,南面分水岭高程较低,约 2 000 m 左右^[23]。雅砻江流域及其站点分布如图 1 所示。用于雅砻江流域中长期径流预测的相关资料包括:①1951 年 1 月—2011 年 12 月,130 项大气环流因子数据(本文选取与雅砻江流域相关的 21 项遥相关气候因子);②1960 年 1 月—2021 年 9 月的气压、温度、湿度、降水、风速、日照等气

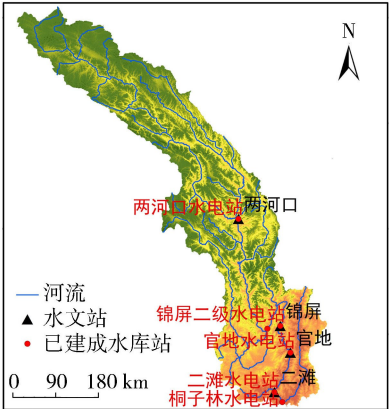


图 1 雅砻江流域及其站点分布

象资料;③1960 年 1 月—2016 年 12 月的两河口、锦屏、官地、二滩水文站径流资料;④1998 年 4 月—2008 年 7 月的归一化植被指数(normalized difference vegetation index, NDVI)数据(旬尺度)。

考虑水文时间序列的一致性,以及比较不同水文要素对流域径流的影响,本文将数据集分为 D1 和 D2 两组。其中,D1 数据集时间跨度为 1998 年 4 月—2008 年 7 月,包括前期流域径流综合指数、面雨量指数、遥相关气候因子共计 124 组样本;D2 数据集时间跨度为 1998 年 4 月—2008 年 7 月,包括前期流域径流综合指数、面雨量指数、遥相关气候因子,以及植被指数数据共计 124 组样本。

2.2 径流综合指数构造结果

本文首先对雅砻江流域 4 个水文站(两河口、锦屏、官地、二滩)进行月平均径流一致性分析,结果表明上述站点之间保持着很高的相关性(图 2),因此基于上述水文站点的月平均径流(细粒度),构造流域径流综合指数(粗粒度);然后根据式(1)(2),计算不同水文站点权重,结果如表 1 所示;最后构建径流综合指数,并与 4 个水文站的月平均径流进行 Pearson 相关性分析,结果表明构建的径流综合指数与 4 个水文站的月平均径流高度相关(表 2)。

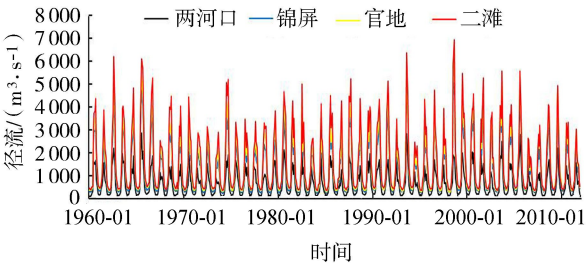


图 2 流域内 4 个水文站月平均径流一致性对比

表 1 基于 4 个水文站点流域控制面积构建流域径流综合指数权重

测站	测站控制面积/%	归一化权重
两河口	51	0.366 510
锦屏	81	0.230 766
官地	89	0.210 023
二滩	97	0.192 701

表 2 径流综合指数与 4 个水文站月平均径流量 Pearson 相关性分析

站点	两河口	锦屏	官地	二滩	径流综合指数
两河口	1.000	0.966	0.963	0.947	0.975
锦屏	0.966	1.000	0.998	0.993	0.998
官地	0.963	0.998	1.000	0.994	0.998
二滩	0.947	0.993	0.994	1.000	0.993
径流综合指数	0.975	0.998	0.998	0.993	1.000

2.3 基于偏互信息法的因子筛选结果

D1 数据集的候选因子包括雅砻江流域径流综合指数因子 $f_{com}(f_{com}(t-1),f_{com}(t-2),\cdots,f_{com}(t-12))$ 、面雨量指数因子 $f_{rain}(f_{rain}(t-1),f_{rain}(t-2),\cdots,f_{rain}(t-12))$ 、21 个遥相关气候因子 $f_{atml}(f_{atml}(t-1),f_{atml}(t-2),\cdots,f_{atml}(t-12)),f_{atm2}(f_{atm2}(t-1),f_{atm2}(t-2),\cdots,f_{atm2}(t-12)),\cdots,f_{atm21}(f_{atm21}(t-1),f_{atm21}(t-2),\cdots,f_{atm21}(t-12))$ 等 23 个对象前期 12 个月的观测值作为备选特征,总数为 276(23×12)个;D2 的候选因子则包括雅砻江流域径流综合指数因子、面雨量指数因子、21 个遥感相关气候因子,以及植被指数因子 $f_{ndvi}(f_{ndvi}(t-1),f_{ndvi}(t-2),\cdots,f_{ndvi}(t-12))$ 等 24 个对象前期 12 个月的观测值作为备选特征,总数为 288(24×12)个。具体如表 3 所示。

表 3 雅砻江流域中长期径流过程变化影响候选因子

候选因子	描述
径流综合指数(f_{com})	表征雅砻江流域径流的丰枯情况
面雨量指数(f_{rain})	选用雅砻江流域内 22 个气象站点所代表的面积赋权构建面雨量指数,各站所代表的面积由泰森多边形计算
f_{atml}	北非副高面积指数
f_{atm2}	大西洋欧洲区极涡面积指数
f_{atm3}	北半球极涡面积指数
f_{atm4}	亚洲纬向环流指数
f_{atm5}	东亚槽强度(CQ)
f_{atm6}	West Pacific Pattern(WP)
f_{atm7}	East Atlantic/West Russia Pattern(EA/WR)
f_{atm8}	Polar/ Eurasia Pattern(POL)
f_{atm9}	赤道中东太平洋 200hPa 纬向风指数
f_{atm10}	850hPa 中太平洋信风 CPAC850 指数
f_{atm11}	登陆台风数
f_{atm12}	NINO 3 区海表温度距平指数
f_{atm13}	NINO 4 区海表温度距平指数
f_{atm14}	NINO B 区海表温度距平指数
f_{atm15}	IOWPA 印度洋暖池面积指数
f_{atm16}	IOWPS 印度洋暖池强度指数
f_{atm17}	WPWPS 西太平洋暖池强度指数
f_{atm18}	AMON. US 大西洋年代际振荡指数
f_{atm19}	西风漂流区海温指数
f_{atm20}	ENSO Modoki
f_{atm21}	TIOD 热带印度洋海温偶极子指数
植被指数(f_{ndvi})	基于 SPOT 影像的流域归一化植被指数

为研究不同水文要素对中长期径流预测效果的影响,并考虑自回归项对整个因子筛选的影响太过显著等特点,基于赤池信息准则方法并不完全适用于雅砻江流域径流变化过程的因子筛选。因此,本文将在 PMI 方法基础上结合人工挑选方式进行因子选择:采用 PMI 方法对所有备选因子按照相关性大小进行排序,并在此基础上分别选择各自相关性大小排名前 20 的备选因子;以此为基础,再通过人工挑选方式分别在 D1 与 D2 数据集上选取上述几类水文对象中相关性大小各自排序前列的因子(最

多排名前五),最终组合形成新的综合筛选结果。其中,D1 数据集筛选后的因子为 11 个,D2 为 13 个,具体如表 4 所示。

表 4 不同数据集的因子筛选结果

流域径流综合影响对象	D1 筛选结果	D2 筛选结果
径流综合指数	$f_{com}(t-12),f_{com}(t-1),f_{com}(t-11)$	$f_{com}(t-12),f_{com}(t-1),f_{com}(t-11),f_{com}(t-2)$
面雨量指数	$f_{rain}(t-1),f_{rain}(t-2),f_{rain}(t-12)$	$f_{rain}(t-1),f_{rain}(t-12)$
气候因子	$f_{atml6}(t-5),f_{atml5}(t-6),f_{atml}(t-1),f_{atml}(t-7),f_{atml0}(t-4)$	$f_{atml6}(t-5),f_{atml5}(t-6),f_{atml}(t-1),f_{atml}(t-7),f_{atml0}(t-4)$
植被指数		$f_{ndvi}(t-12),f_{ndvi}(t-1)$

2.4 基于 IPSO-ELM 模型的流域中长期径流预测

2.4.1 数据集划分

本文将 D1 与 D2 数据集划分为两部分,一部分数据用于预测模型的 10 折交叉验证(10-CV),称为交叉验证期;另一部分数据用于模型测试。其中,用于 10-CV 的数据为 1998 年 4 月—2006 年 7 月共计 100 组样本(随机选取 90 组用于训练,余下 10 组用于验证模型),测试数据为 2006 年 8 月—2008 年 7 月共 24 组样本。交叉验证期与测试期数据划分如表 5 所示。

表 5 交叉验证期及测试期数据划分

数据集	交叉验证期	测试期	流域径流综合影响对象
D1	1998-04—2006-07	2006-08—2008-07	径流综合指数、面雨量指数、气候因子
D2	1998-04—2006-07	2006-08—2008-07	径流综合指数、面雨量指数、气候因子、植被指数

2.4.2 参数设置

粒子群算法初始化为:种群规模为 40,最大迭代次数为 400,粒子位置区间为 $[-2,2]$,粒子速度区间 $[-0.5,0.5]$,其他参数设置为 c_{10} 取值 2.2, c_{11} 取值 1.2, c_{20} 取值 0.3, c_{21} 取值 2.2, p_0 取值 0.01, p_1 取值 0.28, w_0 取值 1, w_1 取值 0.1,学习速率为 0.1,训练目标为 0.001。适应度函数选择 ELM 的 10 折交叉验证的平均均方根误差,ELM 的激活函数选择“sigmoid”。

为了保证对比试验的可靠性,本文对 BPNN 和 SVM 算法进行了相应优化,其初始参数设置分别为:BPNN 基于 10-CV 方法,采用与 ELM 相同的结构,训练函数选择“tansig”,学习函数选择“logsig”,最大训练次数为 600,学习速率为 0.1,训练采用 LM 算法,动量因子为 0.9,期望误差为 0.001;支持向量机回归模型选择径向基核函数,采用 10-CV 方法,通过多次试验寻找最佳参数:径向基核函数中的 $\sigma=0.5$,惩罚参数 $C_a=1,\varepsilon=0.001$ 。

2.4.3 训练结果及分析

在 10-CV 阶段,5 种预测模型在 D1 与 D2 数据集上均具有较好的训练效果,其中 IPSO-ELM 模型在 E_{mape} 、 E_{rmse} 和 E_{dc} 3 个指标方面整体优于其他 4 种预测模型,显示了本文所提算法在该阶段具有较好的拟合效果和泛化能力。另外,5 种预测模型在 D1 数据集上的交叉验证效果整体优于 D2,说明当样本较少时,考虑参与的径流影响要素越多,效果反而较差。不同预测模型在 D1 与 D2 数据集上的 10-CV 性能对比如表 6 所示。

表 6 不同模型在 D1 与 D2 数据集上 10-CV 的性能比较

模型	D1 数据集			D2 数据集		
	$E_{\text{mape}}/\%$	$E_{\text{rmse}}/(\text{m}^3 \cdot \text{s}^{-1})$	E_{dc}	$E_{\text{mape}}/\%$	$E_{\text{rmse}}/(\text{m}^3 \cdot \text{s}^{-1})$	E_{dc}
BPNN	20.45	288.34	0.893 2	24.14	342.46	0.812 8
SVM	22.36	250.46	0.860 2	30.17	289.32	0.846 4
ELM	28.35	233.68	0.881 6	28.56	310.98	0.835 6
PSO-ELM	20.18	201.89	0.891 5	23.79	245.45	0.867 8
IPSO-ELM	18.32	183.34	0.943 2	21.08	203.14	0.910 2

2.4.4 预测结果及分析

考虑到模型评测结果的可靠性,在 D1 与 D2 数

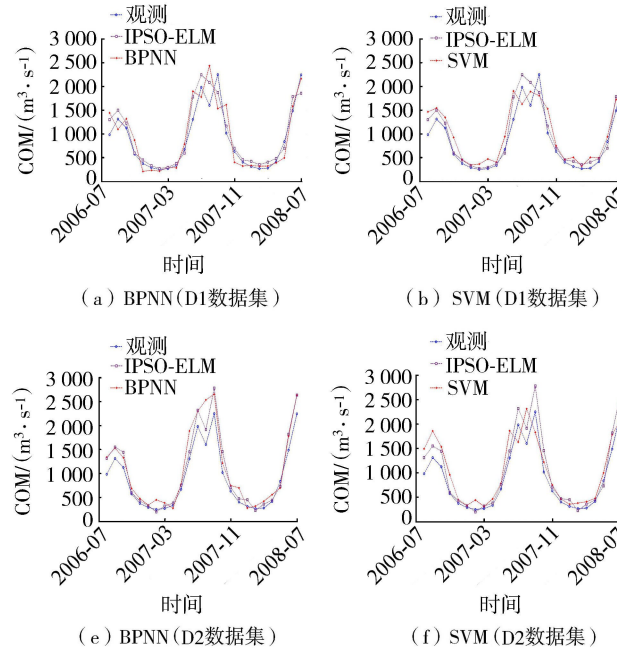


图 3 IPSO-ELM 模型与其他 4 种模型在 D1 与 D2 数据集上的预测结果对比

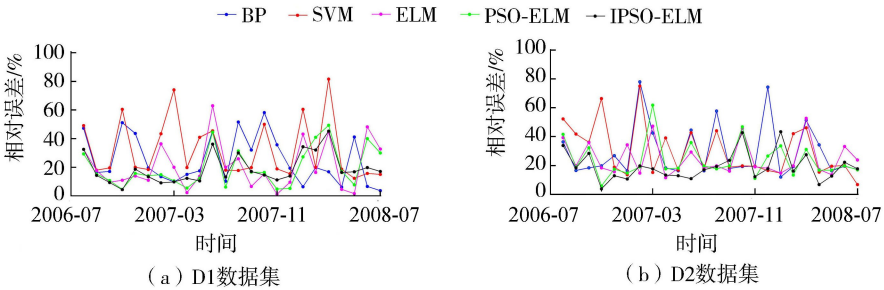


图 4 不同模型在 D1 与 D2 数据集上的相对误差对比曲线

据集上,测试集设计为 2006 年 8 月至 2008 年 7 月共 24 组样本数据,对比模型选用 BPNN、SVM、ELM 和 PSO-ELM。IPSO-ELM 模型分别与其他 4 种模型在 D1 与 D2 数据集上的预测结果对比如图 3 所示。相对误差对比曲线如图 4 所示,综合性能对比如表 7 所示。

2.4.4.1 不同模型预测结果比较

在 D1 数据集上,5 种预测模型均具有较好的预测效果,但在不同的评测指标上展现出具有差异性的结果。其中:在 E_{mape} 指标上,IPSO-ELM 模型最优,相比较,SVM 在该项指标上得分较低; E_{rmse} 指标与研究区域的年来水量紧密相关,年来水量大的流域,其 E_{rmse} 指标也较大,其中 IPSO-ELM 模型的 E_{rmse} 最小,相比较,BPNN 模型较大;在 E_{dc} 指标上,5 种预测模型的确定系数都较高,其中 IPSO-ELM 模型最优,说明 5 种预测模型在基于信息熵的因子筛选基础上对流域中长期径流预测都具有较好的拟合效果;在 E_{qr} 指标上,5 种预测模型的合格率都较高,其中 IPSO-ELM 模型最优 (IPSO-ELM 为 75%,BPNN 为 62.5%,SVM 为 58.3%,ELM 为 66.7%,PSO-ELM 为 66.7%),说明 5 种预测模型均可以应用于

表 7 不同模型在 D1 与 D2 数据集上的综合性能对比

模型	D1 数据集				D2 数据集			
	$E_{\text{mape}}/\%$	$E_{\text{rmse}}/(\text{m}^3 \cdot \text{s}^{-1})$	E_{dc}	$E_{\text{qr}}/\%$	$E_{\text{mape}}/\%$	$E_{\text{rmse}}/(\text{m}^3 \cdot \text{s}^{-1})$	E_{dc}	$E_{\text{qr}}/\%$
BPNN	25.04	327.55	0.733 0	62.5	29.50	303.17	0.771 2	62.5
SVM	32.15	281.83	0.802 3	58.3	29.87	308.55	0.763 0	58.3
ELM	20.75	311.32	0.758 8	66.7	24.82	270.19	0.818 3	62.5
PSO-ELM	19.49	271.49	0.815 6	66.7	23.87	256.94	0.835 7	66.7
IPSO-ELM	18.76	218.86	0.880 8	75.0	19.16	232.27	0.865 7	70.8

水文作业,其中基于 IPSO-ELM 模型的预报方案属于乙等,可用于向水文部门正式提供预报成果。

在 D2 数据集上,5 种预测模型均具有较好的预测效果,但在不同的评测指标上展现出具有差异性的结果。其中:在 E_{mape} 指标上,IPSO-ELM 模型最优,相比较,BPNN 和 SVM 在该项指标上得分较低;在 E_{rmse} 指标上,IPSO-ELM 模型最优,相比较,BPNN 和 SVM 的均方根误差较大;在 E_{dc} 指标上,IPSO-ELM 模型较好,相比较,BP 和 SVM 的确定性系数较低;在 E_{qr} 指标上,5 种模型的合格率都较高,其中 IPSO-ELM 模型最优 (IPSO-ELM 合格率为 70.8%,BPNN 为 62.5%,SVM 为 58.3%,ELM 为 62.5%,PSO-ELM 为 66.7%)。

综上所述,5 种预测模型在 D1 与 D2 数据集上均具有较好的预测效果。其中,IPSO-ELM 模型的预测效果最佳。主要原因在于:BPNN 和 SVM 模型结构相对复杂,参数在训练过程中需要初始化与不断优化调整,整体效率不高,而 ELM 具有参数设置简单、计算速度快、误差小、泛化能力强等优点,因而整体预测效果优于上述两种常用模型;针对传统 ELM 模型输入权值和隐含层阈值随机给定,可能导致部分隐含层节点失效问题,本文结合 K 折交叉验证与 IPSO 算法,加快了 ELM 模型参数寻优速度,因而提高了预测效果。

2.4.4.2 不同数据集上的预测效果比较

在增加径流影响要素的情况下,5 种预测模型在 D1 数据集上的预测效果整体上胜于 D2(增加了 NDVI 植被指数),主要原因在于:流域径流影响要素增多导致模型的输入变量增加,从而增加了模型的计算复杂度,导致在 D2 数据集上的整体运算效率不如 D1;与中长期径流预测有关的时间序列较短(训练样本较少),模型难以充分学习,导致 D2 整体预测效果差于 D1。因此,在流域径流影响要素增多时,模型的输入维度也相应增加,而时间序列又相对较短(训练样本较少)时,预测模型难以充分训练学习,整体预测效果可能会相对较差。

3 结 论

a. 基于各站控制面积的流域径流综合指数能

够较好地反映流域水情的丰枯变化。在此基础上采用基于信息熵的因子筛选方法,获得了影响流域中长期径流过程变化的关键因子集合,形成 IPSO-ELM 模型的特征输入。

b. 以 K 折交叉验证法求得的均方根误差作为粒子的适应度,以适应度值为基础对粒子进行寻优,通过迭代更新找到最优的个体粒子,获得了 ELM 模型的最优参数,由此构建了 IPSO-ELM 模型,并用于流域中长期径流预测,提高了预测精度。

c. 实例计算与对比分析结果表明,所提方法具有较好的实用性,且所建模型性能优于 BPNN、SVM、ELM 和 PSO-ELM 等预测模型,可为流域中长期径流变化趋势预测分析提供一定参考。

参考文献:

[1] 谭乔凤,陈然,朱阳,等. 基于多因子最近邻抽样回归模型的径流相似性预报[J]. 河海大学学报(自然科学版),2020,48(6) :521-527. (TAN Qiaofeng, CHEN Ran, ZHU Yang, et al. Runoff similarity forecast based on multi-factor nearest neighbor bootstrapping regressive model [J]. Journal of Hohai University (Natural Sciences),2020,48(6) :521-527. (in Chinese))

[2] 芮孝芳. 数值天气预报的成功经验对洪水预报的启示[J]. 水利水电科技进展,2019,39(1) :1-6. (RUI Xiaofang. Inspirations of successful experience of numerical weather prediction to flood forecasting [J]. Advances in Science and Technology of Water Resources, 2019,39(1) :1-6. (in Chinese))

[3] 张树磊,杨大文,杨汉波,等. 1960—2010 年中国主要流域径流量减小原因探讨分析[J]. 水科学进展,2015,26(5) :605-613. (ZHANG Shulei, YANG Dawen, YANG Hanbo, et al. Analysis of the dominant causes for runoff reduction in five major basins over China during 1960—2010 [J]. Advances in Water Science,2015,26(5) :605-613. (in Chinese))

[4] ZHANG Q, WANG B D, HE B, et al. Singular spectrum analysis and ARIMA hybrid model for annual runoff forecasting[J]. Water Resources Management, 2011, 25(11) : 2683-2703.

[5] VALIPOUR M. Long-term runoff study using SARIMA and ARIMA models in the United States[J]. Meteorological Applications, 2015, 22(3) : 592-598.

- [6] ESCALANTE-SANDOVAL C , AMORES-ROVELO L. Regional monthly runoff forecast in southern Canada using ANN, K-means, and L-moments techniques [J]. Canadian Water Resources Journal, 2017, 42 (3): 205-222.
- [7] SHOAI B M, SHAMSELDIN A Y, MELVILLE B W, et al. A comparison between wavelet based static and dynamic neural network approaches for runoff prediction [J]. Journal of Hydrology, 2016, 535: 211-225.
- [8] 金君良, 舒章康, 陈敏, 等. 基于数值天气预报产品的气象水文耦合径流预报 [J]. 水科学进展, 2019, 30 (3): 316-325. (JIN Junliang, SHU Zhangkang, CHEN Min, et al. Meteorological-hydrological coupled runoff forecasting based on numerical weather prediction products [J]. Advances in Water Science, 2019, 30 (3): 316-325. (in Chinese))
- [9] LIU Y, SANG Y F, LI X, et al. Long-term streamflow forecasting based on relevance vector machine model [J]. Water, 2017, 9 (1): 9.
- [10] TAN Q F, LEI X H, WANG X, et al. An adaptive middle and long-term runoff forecast model using EEMD-ANN hybrid approach [J]. Journal of Hydrology, 2018, 567: 767-780.
- [11] LIANG Z, LI Y, HU Y, et al. A data-driven SVR model for long-term runoff prediction and uncertainty analysis based on the Bayesian framework [J]. Theoretical and Applied Climatology, 2018, 133 (1/2): 137-149.
- [12] 崔东文. 多组群教学优化算法-神经网络-支持向量机组合模型在径流预测中的应用 [J]. 水利水电科技进展, 2019, 39 (4): 41-48. (CUI Dongwen. Combined model of multi-group teaching optimization algorithm-neural network-support vector machine in runoff prediction application [J]. Advances in Science and Technology of Water Resources, 2019, 39 (4): 41-48. (in Chinese))
- [13] 王亦斌, 孙涛, 梁雪春, 等. 基于 EMD-LSTM 模型的河流水量水位预测 [J]. 水利水电科技进展, 2020, 40 (6): 40-47. (WANG Yibin, SUN Tao, LIANG Xuechun, et al. Prediction of river water flow and water level based on EMD-LSTM model [J]. Advances in Science and Technology of Water Resources, 2020, 40 (6): 40-47. (in Chinese))
- [14] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: theory and applications [J]. Neurocomputing, 2006, 70: 489-501.
- [15] 荣艳淑, 胡玉恒, 冯瑞瑞, 等. 广义相加模型在乌江夏季径流预报中的应用 [J]. 河海大学学报 (自然科学版), 2021, 49 (2): 121-126. (RONG Yanshu, HU Yuheng, FENG Ruirui, et al. Application of generalized additive model in summer runoff forecasting of Wujiang Basin [J]. Journal of Hohai University (Natural Sciences), 2021, 49 (2): 121-126. (in Chinese))
- [16] 赵钢铁, 杨大文. 神经网络径流预报模型中基于互信息的预报因子选择方法 [J]. 水力发电学报, 2011, 30 (1): 24-30. (ZHAO Tongtiengang, YANG Dawen. Mutual information-based input variable selection method for runoff-forecasting neural network model [J]. Journal of Hydroelectric Engineering, 2011, 30 (1): 24-30. (in Chinese))
- [17] 麦紫君, 曾小凡, 周建中, 等. 基于偏互信息法遴选因子的长江中长期径流预报 [J]. 人民长江, 2018, 49 (3): 52-56. (MAI Zijun, ZENG Xiaofan, ZHOU Jianzhong, et al. Medium and long-term runoff forecasting based on input variables selected by partial mutual information [J]. Yangtze River, 2018, 49 (3): 52-56. (in Chinese))
- [18] 徐计, 王国胤, 于洪. 基于粒计算的大数据处理 [J]. 计算机学报, 2015, 38 (8): 1497-1517. (XU Ji, WANG Guoyin, YU Hong. Review of big data processing based on granular computing [J]. Chinese Journal of Computers, 2015, 38 (8): 1497-1517. (in Chinese))
- [19] 张金萍, 肖宏林, 张鑫. 基于经验模态分解方法和信息熵的水沙关系研究 [J]. 水资源保护, 2019, 35 (4): 30-34. (ZHANG Jinping, XIAO Honglin, ZHANG Xin. Study on relationship between water and sediment based on empirical mode decomposition and information entropy [J]. Water Resources Protection, 2019, 35 (4): 30-34. (in Chinese))
- [20] MAY R J, DANDY G C, MAIER H R, et al. Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems [J]. Environmental Modelling & Software, 2008, 23 (10/11): 1289-1299.
- [21] HE J, VALEO C, CHU A, et al. Prediction of event-based stormwater runoff quantity and quality by ANNs developed using PMI-based input selection [J]. Journal of Hydrology, 2011, 400 (1/2): 10-23.
- [22] WANG G G, HOSSEIN G A, YANG X S, et al. A novel improved accelerated particle swarm optimization algorithm for global numerical optimization [J]. Engineering Computations, 2014, 31 (7): 1198-1220.
- [23] 袁定波, 艾萍, 洪敏, 等. 基于地理空间要素的雅砻江流域面雨量估算 [J]. 水科学进展, 2018, 29 (6): 779-787. (YUAN Dingbo, AI Ping, HONG Min, et al. Estimation of areal rainfall in Yalong River basin based on geospatial factors [J]. Advances in Water Science, 2018, 29 (6): 779-787. (in Chinese))

(收稿日期: 2020-04-25 编辑: 郑孝宇)