

DOI:10.3876/j.issn.1000-1980.2020.06.002

基于遗传算法的 SVM-AR 改进模型与应用

王红瑞¹, 魏豪杉^{2,3}, 胡立堂¹, 赵自阳¹, 娄和震¹

(1. 北京师范大学水科学研究院, 北京 100875; 2. 中国科学院地理科学与资源研究所, 北京 100101;
3. 中国科学院大学资源与环境学院, 北京 100190)

摘要: 为提升河流流量的预测精度, 将支持向量机与 AR 进行耦合, 并构造三核混合核函数的流量预测支持向量机模型。以渭河流域的月径流量为例, 首先, 通过时间序列分析, 将渭河流域的径流序列划分为趋势序列、季节序列和随机波动序列, 然后利用 AR 模型构造适用于支持向量机算法的数据集, 并将数据集按 4 : 1 划分为训练集和检验集; 其次, 利用线性组合构造由多项式核函数、径向基核函数与 Sigmoid 核函数构成的三核混合核函数, 在训练集上, 采用遗传算法确定相关参数, 随后在检验集上进行预测。结果表明: 遗传算法确定参数会带来较大的不确定性, 导致结果差异较大, 从而着重讨论遗传算法带来的参数不确定性; 通过函数构造与统计分析, 给出三核混合核函数参数选择的一般性方法与流程, 并进行验证, 该参数选取方法能够降低遗传算法的不确定性, 得到精度较高的流量预测结果, 预测流量与实际流量的均方误差从 150 左右降低到 130 左右。

关键词: 支持向量机; 遗传算法; 混合核函数; 时间序列分析; 自回归模型; 径流预报; 渭河流域

中图分类号: P338 **文献标志码:** A **文章编号:** 1000-1980(2020)06-0488-10

Improved model and application of SVM-AR based on genetic algorithm

WANG Hongrui¹, WEI Haoshan^{2,3}, HU Litang¹, ZHAO Ziyang¹, LOU Hezhen¹
(1. College of Water Sciences, Beijing Normal University, Beijing 100875, China;
2. Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China;
3. College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In order to improve the prediction accuracy of river flow, the support vector machine (SVM) and the AR model are coupled to construct the SVM model for the river flow prediction with three-core hybrid kernel function. Taking the monthly runoff in the Weihe River Basin as an example, firstly through the time series analysis, the runoff data of Weihe River Basin was divided into trend data, seasonal data and random fluctuation data. A data set suitable for support vector machine algorithms was constructed with the AR model, and then was divided into the training set and the test set by 4 : 1. Secondly, with the linear combination, this study constructed a three-core hybrid kernel function composed of polynomial kernel function, radial basis kernel function and Sigmoid kernel function. On the training set, the genetic algorithm was used to determine the relevant parameters, and predictions were conducted on the test set. It was found that when the genetic algorithm is used to determine the parameters, it will bring greater uncertainty with greater differences in results, thus it need more discussions on the parameter uncertainty brought by genetic algorithm. Through the function construction and statistical analysis, the general method and process of parameter selection of the three-core hybrid kernel function are given and verified. With this method, the uncertainty of the genetic algorithm can be reduced on the test set, and a more accurate flow prediction result can be achieved. The mean square error between the predicted flow rate and the actual flow rate is reduced from about 150 to about 130.

Key words: support vector machines; genetic algorithm; hybrid kernel function; time series analysis;

基金项目: 国家重点研发计划(2018YFC0407900); 国家自然科学基金(51879010, 51479003)
作者简介: 王红瑞(1963—), 男, 教授, 博士, 主要从事水文与水资源研究。E-mail: henrywang@bnu.edu.cn
通信作者: 魏豪杉, 博士。E-mail: weihs.20b@igsnrr.ac.cn
引用本文: 王红瑞, 魏豪杉, 胡立堂, 等. 基于遗传算法的 SVM-AR 改进模型与应用[J]. 河海大学学报(自然科学版), 2020, 48(6): 488-497.
WANG Hongrui, WEI Haoshan, HU Litang, et al. Improved model and application of SVM-AR based on genetic algorithm[J]. Journal of Hohai University(Natural Sciences), 2020, 48(6): 488-497.

autoregressive model; runoff forecast; Weihe River Basin

水文过程的复杂性和随机性,导致其存在强烈的非线性,传统的回归、时间序列等方法难以很好地预测其变化,而支持向量机、神经网络等机器学习方法可以更好地解决这类问题^[1-5]。与神经网络等机器学习算法相比,支持向量机及其相关方法在水文预报方面具有独特的优势^[6-11]。

近年来,支持向量机(support vector machine, SVM)被广泛应用于水文时间序列的分析中且被不断地改进。部分学者将支持向量机与智能优化算法结合,通过改进智能算法的过程提升支持向量机参数的率定精度,从而提升支持向量机的预测精度,如李天宏等^[1]改进粒子群算法,结合支持向量机算法提升径流预报效果,张岩等^[12]利用构建的 PCA-PSO-SVR 算法对丹江口水库年径流进行预报。部分学者通过使用其他模型或分析方法与支持向量机耦合得到混合模型,提升预测精度,如,王红瑞等^[8]基于小波变换构建的支持向量机模型的预测精度比传统的支持向量机和神经网络模型更高;Bafitlhile 等^[10]通过进化策略(ES)优化方法对支持向量机与神经网络的参数进行优化和比较,发现支持向量机在流域水流模拟中要优于神经网络方法;Baydaroğlu 等^[13]构造了基于小波分析、奇异谱分析和混沌方法的支持向量机耦合模型预测河流流量;Meng 等^[11]改进 EMD-SVM 得到的 M-EMDSVM 比传统 SVM 方法能更准确地预测强非平稳径流;梁浩等^[4]、周婷等^[14]基于小波分析,提升支持向量机预测精度;贺玉琪等^[15]耦合支持向量机和贝叶斯岭回归的 BRR-SVR 优化模型,并将其应用于月降水量预测;朱双等^[16]利用基于灰色关联分析的模糊支持向量机方法进行径流预报。还有一部分学者通过改进支持向量机的核函数,构造混合核函数,以提升支持向量机对径流预报的预测精度,如唐奇等^[17]将多项式核与径向基核线性组合构造混合核函数提升预测精度;刘昕玥等^[18]利用改进的自适应遗传算法确定二核的混合核函数参数,进行流域的流量预报。但在确定混合核函数各个核函数的比重参数时,文献[17-18]并未给出较合理的混合核函数比重参数的确定方法。在实际问题中,各个核函数的比重参数是无法预先确定的,即使预先给定这些参数也无法保证所构造的混合核函数的预测效果会优于单一核函数,也就是说,文献[17-18]主要说明了混合核函数的可用性,但没有给出混合核函数的用法,而本文给出了一种混合核函数的具体使用方式。

针对文献[17-18]存在的问题,本文随机产生足够多的混合核函数比重参数样本,通过这些参数得到若干个流量预报结果,取中值确定最终的预报结果。本文采用的支持向量机算法的核函数包括三核混合核函数、二核混合核函数与单一核函数,采用遗传算法确定核函数参数。但无论采用何种智能算法,得到的核函数参数值都存在不确定性,针对这一问题,本文基于已有数据构造函数对核函数参数以及核函数类型进行选择,以减少这种不确定性。

1 基于 SVM 的时间序列模型构建

1.1 时间序列分解

对已知的较为复杂的时间序列 $\{X_t^{\text{complex}}\}$ 进行确定性分析,采用乘法模型化简序列:

$$X_t^{\text{complex}} = T_t S_t I_t \tag{1}$$

式中: $\{T_t\}$ 、 $\{S_t\}$ 、 $\{I_t\}$ —— $\{X_t^{\text{complex}}\}$ 的长期趋势波动的时间序列、季节性波动的时间序列、随机波动的时间序列; t ——时间序列的时间点。

$\{S_t\}$ 为有周期性的序列,需要处理的是 $\{T_t\}$ 以及 $\{I_t\}$ 。对 $\{T_t\}$ 和 $\{I_t\}$ 进行平稳性分析,若序列为非平稳序列,则通过差分等方式把序列平稳化。

1.2 SVM-AR 模型构建及平稳性分析

自回归滑动平均模型 ARMA(p, q) 为

$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (\varphi_p \neq 0, \theta_q \neq 0) \tag{2}$$

其中 $E(\varepsilon_t) = 0 \quad \text{Var}(\varepsilon_t) = \sigma_\varepsilon^2 \quad E(\varepsilon_t \varepsilon_s) = 0, s \neq t \quad E(X_s \varepsilon_t) = 0, \forall s < t$

式中: $\{X_t\}$ ——平稳的时间序列; p, q ——模型阶数; φ_p, θ_q ——模型参数; ε_t ——正态分布的随机变量; σ_ε^2 ——关于随机变量 ε_t 的方差; $E(\cdot)$ 、 $\text{Var}(\cdot)$ ——期望、方差函数。

利用 ARMA(p, q) 模型建立基于支持向量机的时间序列模型:

$$X_t = f_{\text{svm_ar}}(X_{t-1}, X_{t-2}, \cdots, X_{t-p}) + I_{q_t} \tag{3}$$

其中
$$I_{q_t} = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

式中: $f_{\text{svm_ar}}(\cdot)$ ——支持向量机预测函数。

自回归模型 $\text{AR}(p)$ 为 $\text{ARMA}(p,q)$ 中 q 不存在的特殊情况:

$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t \tag{4}$$

采用 ADF (augmented Dickey-Fuller) 检验序列的平稳性,式(4)的特征方程为 $\lambda^p - \varphi_1 \lambda^{p-1} - \cdots - \varphi_p = 0$,若其所有特征根都在单位圆内,即 $|\lambda_i| < 1 (i = 1, 2, \cdots, p)$,则序列 $\{X_t\}$ 平稳;若有一个单位根存在,则序列 $\{X_t\}$ 非平稳,且自回归系数之和恰为 1,记 $\rho = \varphi_1 + \varphi_2 + \cdots + \varphi_p - 1$,则 $\text{AR}(p)$ 过程单位根检验的假设条件为 $H_0: \rho = 0$ (序列 $\{X_t\}$ 非平稳) $\longleftrightarrow H_1: \rho = 0$ (序列 $\{X_t\}$ 平稳)。构造的 ADF 统计量 $\tau = \frac{\hat{\rho}}{S(\hat{\rho})}$,其中 $\hat{\rho}$ 为参数 ρ 的样本均值, $S(\hat{\rho})$ 为参数 ρ 的样本标准差。

1.3 模型定阶

为确定 $\text{ARMA}(p,q)$ 模型的阶数 p,q ,本文采取以样本自相关系数与偏自相关系数判别为主,AIC (Akaike's information theoretic criterion)、BIC (Bayesian information criterion) 或称 SBC (Schwartz's Bayesian criterion) 定阶为辅的模型定阶方法。

样本自相关系数为

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k} (X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2} \quad 0 < k < n \tag{5}$$

样本偏自相关系数为

$$\hat{\phi}_{kk} = \frac{\hat{D}_k}{\hat{D}} \quad 0 < k < n \tag{6}$$

其中

$$\hat{D} = \begin{vmatrix} 1 & \hat{\rho}_1 & \cdots & \hat{\rho}_{k-1} \\ \hat{\rho}_1 & 1 & \cdots & \hat{\rho}_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_{k-1} & \hat{\rho}_{k-2} & \cdots & 1 \end{vmatrix}$$
$$\hat{D}_k = \begin{vmatrix} 1 & \hat{\rho}_1 & \cdots & \hat{\rho}_1 \\ \hat{\rho}_1 & 1 & \cdots & \hat{\rho}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_{k-1} & \hat{\rho}_{k-2} & \cdots & \hat{\rho}_k \end{vmatrix}$$

式中: n ——样本容量; $\bar{X}$ —— X_t 样本均值。

当样本容量 n 充分大时,样本的自相关系数和偏自相关系数均近似服从正态分布,即

$$\hat{\rho}_k \sim N\left(0, \frac{1}{n}\right) \quad \hat{\phi}_{kk} \sim N\left(0, \frac{1}{n}\right) \tag{7}$$

由正态分布性质可得

$$\Pr\left(-\frac{2}{\sqrt{n}} \leq \hat{\rho}_k \leq \frac{2}{\sqrt{n}}\right) \geq 0.95 \quad \Pr\left(-\frac{2}{\sqrt{n}} \leq \hat{\phi}_{kk} \leq \frac{2}{\sqrt{n}}\right) \geq 0.95 \tag{8}$$

式中: $\Pr(\cdot)$ ——正态分布的分布函数。

故可以利用 2 倍标准差的范围进行判断,从而模型定阶结果见表 1。由表 1 可知,上述定阶方法仍具有一定的不确定性,若出现数种均满足模型定阶条件,或模型阶数较为模糊的情况,可以采用 AIC 或者 SBC(BIC) 定阶准则进行最终的模型定阶。AIC 准则由日本统计学家 Akaike^[19] 于 1973 年提出,SBC(BIC) 准则是为弥补 AIC 准则在样本容量趋于无穷大时,所选择模型不收敛于真实模型而提出的基于 Bayes 理论的判别标准。AIC 准则考虑的是所建模型拟合精度与未知参数个数的综合最优配置,是拟合精度与参数个数的加权函数:

$$S_{\text{AIC}} = -2 \ln M_{\text{value}} + 2 N_{\text{parameter}} \tag{9}$$

SBC 准则对 AIC 稍加改动,定义如下:

$$S_{\text{SBC}} = -2 \ln M_{\text{value}} + \ln N_{\text{parameter}} \tag{10}$$

式中: M_{value} ——模型的极大似然函数值; $N_{\text{parameter}}$ ——模型中未知参

表 1 模型定阶结果

Table 1 Model ordering results		
$\hat{\rho}_k$	$\hat{\phi}_{kk}$	模型定阶
拖尾	$\hat{p}$ 阶截尾	$p = \hat{p}, q = 0$
$\hat{q}$ 阶截尾	拖尾	$p = 0, q = \hat{q}$
拖尾	拖尾	利用 2 倍标准差范围辅助判断

注: $\hat{p}, \hat{q}$ 分别为自相关图、偏自相关图的截尾阶数,即该阶数后自相关系数或偏自相关系数趋于零,否则为拖尾。

数个数; S_{AIC} ——AIC 准则计算结果; S_{SBC} ——SBC 准则计算结果。

在合理范围内,2 个准则的取值均是越小越好,但 AIC 与 SBC 在模型定阶时可能会出现不一致的情况,本文优先选择 SBC 准则。

2 支持向量机核函数构造与参数确定

支持向量机是统计学习中常见的应用于数据分类、预测的方法,核函数是支持向量机的核心,由 Aizerman、Braverman 与 Rozoener 于 1964 年提出并在机器学习领域引入的理念^[20]。核函数方法首先通过特征空间映射,把原输入空间的数据映射到合适的特征空间,然后在该特征空间中分析、提取数据特征,进行相关运算。该方法的优势在于,以内积的形式实现非线性变换,避免了维数过大而造成的“维数灾难”。支持向量机通过引入核函数,直接计算高维特征空间中的内积,实现将输入空间升维的非线性变换,降低了计算的复杂性。核函数有多种类型,核函数类型的选择会直接影响支持向量机预测结果的好坏。

支持向量机的主要思想如下:给定训练样本 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$, 其中, $\mathbf{x}_i \in \mathbb{R}^n$ 是输入向量, $y_i \in \mathbb{R}$ 是输出指标 ($i = 1, 2, \dots, m$)。希望寻求一个 $\mathbb{R}^n$ 上的实值函数 $f(\mathbf{x})$, 使得对任意 $\mathbf{x}$, $f(\mathbf{x})$ 与 y 能尽可能接近。首先将原始空间中的数据映射到一个高维空间中,然后在高维空间中进行线性拟合,设映射后的回归函数为 $f(\mathbf{x}) = \boldsymbol{\omega}^T \boldsymbol{\varphi}(\mathbf{x}) + b$, 其中 $\boldsymbol{\varphi}(\cdot)$ 为 $\mathbb{R}^n$ 到高维特征空间中的非线性映射, $\boldsymbol{\omega}$ 为超平面的权值向量, b 为偏置。从而由最优化理论可以把该回归问题转化为

$$\min_{\boldsymbol{\omega}, b} \left[\frac{1}{2} \|\boldsymbol{\omega}\|^2 + C_N \sum_{i=1}^m l_{\varepsilon}(f(\mathbf{x}_i) - y_i) \right] \tag{11}$$

式中: C_N ——正则化常数; l_{ε} —— ε -不敏感损失函数。

引入松弛变量 ξ_i 和 ξ_i^* , 上述问题转化为如下形式的最小化问题:

$$\min_{\boldsymbol{\omega}, b, \xi_i, \xi_i^*} \left[\frac{1}{2} \|\boldsymbol{\omega}\|^2 + C_N \sum_{i=1}^m (\xi_i + \xi_i^*) \right] \tag{12}$$

$$\text{s. t. } \begin{cases} f(\mathbf{x}_i) - y_i \leq \varepsilon + \xi_i \\ y_i - f(\mathbf{x}_i) \leq \varepsilon + \xi_i^* \\ \xi_i \geq 0, \xi_i^* \geq 0 \end{cases} \quad (i = 1, 2, \dots, m) \tag{13}$$

引入拉格朗日乘子,可以得到上最优化问题的对偶问题:

$$\max_{\alpha_i, \alpha_i^*} \left\{ -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m [(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\mathbf{x}_i, \mathbf{x}_j)] + \sum_{i=1}^m [y_i(\alpha_i - \alpha_i^*) - \varepsilon(\alpha_i + \alpha_i^*)] \right\} \tag{14}$$

$$\text{s. t. } \begin{cases} \sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0 \\ 0 \leq \alpha_i, \alpha_i^* \leq C_N \end{cases} \tag{15}$$

式(14)中 $K(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\varphi}(\mathbf{x}_i)^T \boldsymbol{\varphi}(\mathbf{x}_j)$ 为核函数。可以发现 $K(\cdot, \cdot)$ 为对称函数,若 $K(\cdot, \cdot)$ 是定义在空间上的对称函数,那么 K 是核函数当且仅当对于任意数据 $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$, $\mathbf{x}_i \in \mathbb{R}^n$, 核矩阵 $\mathbf{K}$ 总是半正定的^[21]:

$$\mathbf{K} = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \cdots & K(\mathbf{x}_1, \mathbf{x}_m) \\ \vdots & & \vdots \\ K(\mathbf{x}_m, \mathbf{x}_1) & \cdots & K(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix}_{m \times m} \tag{16}$$

所以核函数的线性组合或直积也是核函数,表 2 给出了本文采用的 3 种常见核函数:径向基核函数、多项式核函数、Sigmoid 核函数的表达式,参数名称和参数范围等,以及通过线性组合构造的混合核函数的表达式和混合核比重参数的名称符号。并通过传统的遗传算法,确定表 2 单一核函数相关参数。

流域月径流量预测的流程如图 1 所示,首先对流域的月径流量进行分解,随后对分解得到的流域长期趋势序列、随机波动序列进行平稳性检验与处理、模型定阶,再通过 SVM 算法进行预测,最后将数据整合,得到最终的预测结果。

表2 核函数表达式、参数及范围

Table 2 Expression, parameters and range of kernel function		
核函数	表达式	待确定参数
径向基核函数	$K_1(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \ \mathbf{x}_i - \mathbf{x}_j\ ^2)$	核函数参数: γ 惩罚参数: C
多项式核函数	$K_2(\mathbf{x}_i, \mathbf{x}_j) = [\alpha(\mathbf{x}_i \cdot \mathbf{x}_j) + \beta]^{q_s}$	核函数参数: α, β , 取 $q_s = 1$ 惩罚参数: C
Sigmoid 核函数	$K_3(\mathbf{x}_i, \mathbf{x}_j) = \tanh[v(\mathbf{x}_i \cdot \mathbf{x}_j) + c]$	核函数参数: v, c 惩罚参数: C
混合核函数	$K = \frac{\sum_{i=1}^3 p_i K_i}{\sum_{i=1}^3 p_i}$	核函数参数: $\gamma, \alpha, \beta, v, c$ 混合核比重参数: p_1, p_2, p_3 惩罚参数: C

注:各参数范围为 $\gamma > 0, \alpha > 0, \beta \geq 0, v > 0, c \leq 0, 0 \leq p_1 \leq 1, 0 \leq p_2 \leq 1, 0 \leq p_3 \leq 1, C > 0$ 。

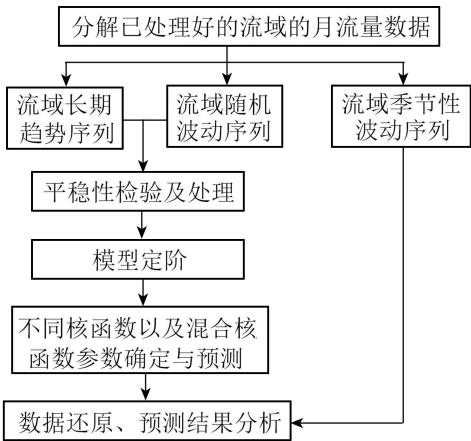


图1 SVM-AR 流程
Fig.1 SVM-AR flowchart

3 实例分析

3.1 月流量数据时间序列分析

3.1.1 流量数据分解以及平稳性分析

将渭河流域咸阳站 1960 年 1 月到 2001 年 12 月的逐月流量按近似 4 : 1 (实际为 50 : 13) 划分为训练集与检验集。首先,对逐月流量的训练集进行确定性分析,通过乘法模型对该序列进行分解,得到长期趋势序列、季节波动序列、随机波动序列(图 2),随后分别对长期趋势序列以及随机波动序列进行平稳性分析。经平稳性检验,训练集中的月流量长期趋势序列不是平稳序列,故对其进行 1 阶差分,差分后的序列经检验为平稳的时间序列,随机波动序列已经为平稳序列,不再进行其他操作,检验结果可见表 3。

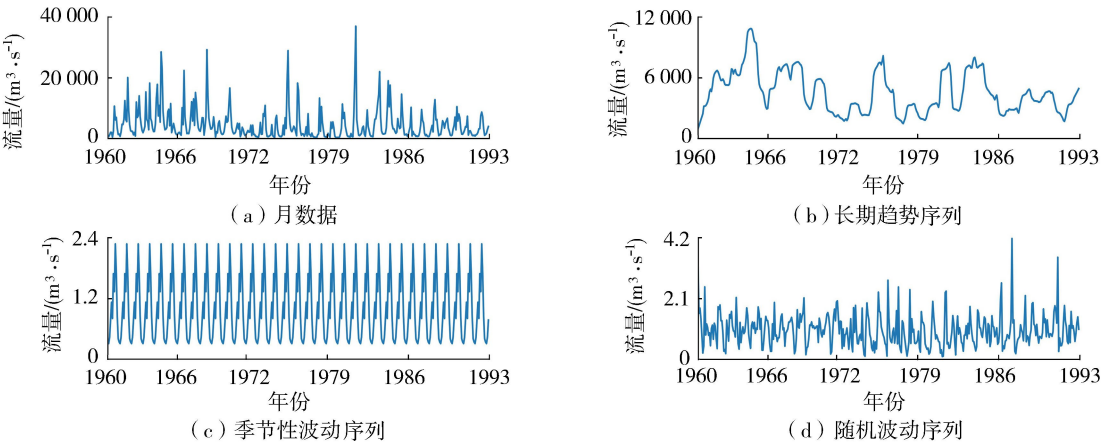


图2 渭河流域咸阳站月流量数据分解
Fig.2 Monthly data decomposition for Xianyang Station of Weihe River Basin

3.1.2 模型定阶

分别绘制渭河流域差分后的月长期趋势波动(简称月趋势)序列和随机波动序列的自相关图与偏自相关图,如图 3 所示,图中蓝色阴影为自相关图或偏自相关图的 95% 置信区间。由图 3 可以初步确定经过 1 阶差分后流域长期趋势序列的阶数大致为的 $p=4,5,6, q=4,5,6$,随机波动序列的阶数大致为 $p=1,2,3, q=2,3,4$ 。由图 4 确定具体的阶数:1 阶差分后的流域长期趋势序列阶数为 $p=5$,随机波动序列阶数为 $p=2$ 。

表3 咸阳站月流量时间序列平稳性检验结果

Table 3 Stationarity test result of monthly runoff time series for Xianyang Station					
时间序列	ADF 检验统计量	p-值	检验临界值		
			1%	5%	10%
长期趋势波动序列	-2.55	0.10	-3.44	-2.87	-2.57
差分后的月流量	-6.51	1.09×10^{-8}	-3.44	-2.87	-2.57
长期趋势波动序列	-12.54	2.32×10^{-23}	-3.44	-2.87	-2.57
随机波动序列					

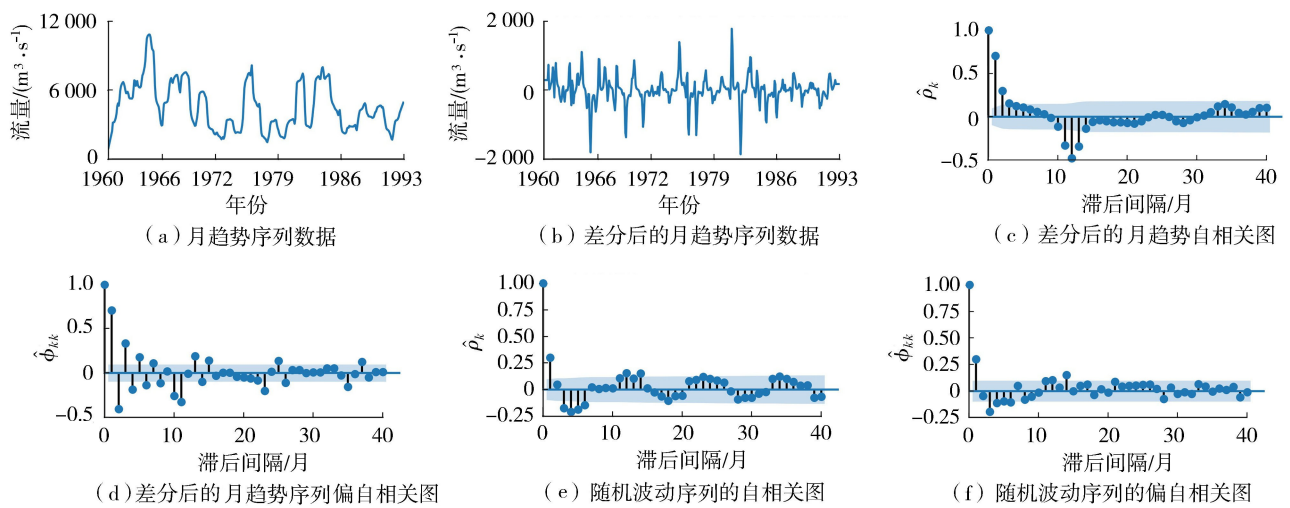


图 3 渭河流域咸阳站各时间序列的自相关和偏自相关图

Fig. 3 Autocorrelation graph and partial correlation graph of each time series for Xianyang Station of Weihe River Basin

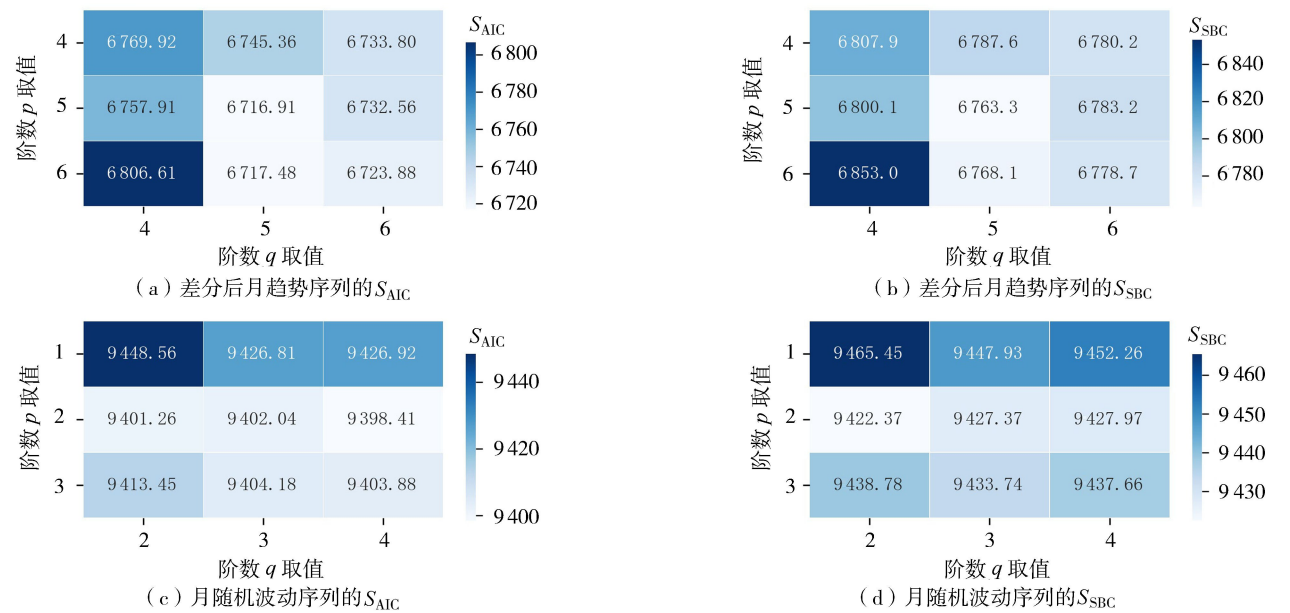


图 4 各时间序列的 S_{AIC} 、 S_{SBC} 热度

Fig. 4 S_{AIC} and S_{SBC} heat maps of each time series

3.2 支持向量机算法使用流程

采用遗传算法在测试集上进行参数确定,使目标函数最小化,目标函数为:测试集“5 折交叉验证”的均方误差最大值。每一个单一核函数的参数都用遗传算法率定一次参数,由于遗传算法具有随机性,所以每一次遗传算法率定的参数都有差异。随后构造混合核函数,由于作用在检验集之前,无法预知单一核函数或者混合核函数的预测效果,所以本文构造的混合核函数是建立在已知单一核函数的基础上的,混合核函数的核函数参数是前面已经确定的单一核函数的核函数参数,惩罚参数取参与构成混合核函数的单一核函数惩罚参数的最小值。本文不通过遗传算法对混合核函数所有参数进行重新率定,原因是参数太多,带来的不确定大大增加,且运算量很大,经检验,预测效果反而不如单一核函数。

通过遗传算法确定支持向量机核函数参数后,还需要确定混合核函数的混合核比重参数,即各个核函数的比例。由于实际问题中观测流量值无法事前得到,所以各个核函数所占的比重事先无法确定,故依据统计学相关理论,随机产生数量足够大的样本进行预测。对于三核混合核函数,随机产生 500 组混合核比重参数,得到 500 条预测流量曲线,取它们的中值作为最终的预测流量曲线。对二核混合核函数,随机产生 450 组由 2 个单一核函数构成的混合核函数比重参数(每种组合 50 组,9 种组合共 450 组,9 种组合分别是:

长期趋势数据集的3种二核混合函数与随机波动数据集的3种二核混合函数的交叉组合),得到450条预测流量曲线,取它们的中值作为二核混合核函数最终的预测流量曲线。单一核函数取9组预测流量曲线中的中值作为最终的预测流量曲线。

由于遗传算法具有一定的随机性,在没有观测流量的情况下,无法预知核函数参数的合理取值。现有数据只有遗传算法确定核函数参数后产生的若干条流量曲线(即3类核函数产生的500+450+9条流量曲线),所以分别通过三核混合核函数500条流量曲线两两之间的均方误差集合(记为 E_{MSE3}),二核混合核函数450条流量曲线两两之间的均方误差集合(记为 E_{MSE2})与单一核函数9条流量曲线两两之间的均方误差集合(记为 E_{MSE1})之间的关系对核函数参数进行选择。运行遗传算法30次,得到集合 E_{MSE1} 、 E_{MSE2} 、 E_{MSE3} 的方差,见表4。

表4 检验集30次遗传算法的运行结果

Table 4 Running results of 30 genetic algorithms on the test set

计算编号	均方误差集合的均值			$V_{MSE_{po1}}$	$V_{MSE_{po2}}$	$V_{MSE_{po3}}$	M_F	M_M	$\frac{2}{5}M_F-10$
	E_{MSE1}	E_{MSE2}	E_{MSE3}						
1	48.73	66.85	34.18	146.12	186.32	195.26	152.90	49.15	51.16
2	53.15	62.60	41.99	175.28	186.95	219.74	58.30	44.46	13.32
3	78.21	63.81	26.03	137.60	213.06	205.63	161.36	68.04	54.54
4	57.87	51.09	49.87	150.53	128.06	128.83	0.20	-21.70	-9.92
5	69.69	57.41	33.29	158.89	133.17	123.77	69.66	-35.12	17.86
6	44.30	51.15	35.92	167.45	146.76	130.33	34.85	-37.13	3.94
7	26.68	59.73	37.10	134.31	173.26	178.93	99.17	44.63	29.67
8	58.11	61.99	34.22	151.70	155.99	174.68	101.15	22.98	30.46
9	67.80	48.81	38.76	143.25	153.52	158.34	12.28	15.10	-5.09
10	142.82	53.95	31.70	154.47	153.58	151.07	41.44	-3.40	6.58
11	25.39	38.62	19.49	138.40	133.53	127.47	72.60	-10.93	19.04
12	54.86	68.63	54.10	135.02	136.03	148.36	28.47	13.35	1.39
13	61.75	57.18	44.02	139.33	147.10	140.44	22.05	1.11	-1.18
14	77.61	56.22	67.01	161.20	135.77	147.23	13.23	-13.97	-4.71
15	42.58	37.51	32.53	159.20	142.68	148.24	3.79	-10.96	-8.48
16	40.94	55.30	38.89	133.92	150.64	171.64	42.11	37.72	6.84
17	73.28	78.57	49.79	136.49	178.30	208.42	96.76	71.92	28.70
18	95.76	79.10	49.24	178.44	165.34	165.98	91.14	-12.47	26.46
19	102.60	70.69	35.62	138.79	158.38	180.23	121.48	41.44	38.59
20	47.21	86.20	45.42	135.24	177.30	219.08	242.10	83.84	86.84
21	43.41	37.29	26.36	140.96	147.45	141.13	18.12	0.17	-2.75
22	65.58	39.14	27.93	151.51	141.71	130.41	15.54	-21.10	-3.78
23	118.35	76.17	50.23	164.35	157.67	196.40	61.85	32.05	14.74
24	35.42	76.67	47.04	139.53	179.00	202.15	147.52	62.62	49.01
25	50.47	61.66	48.98	153.82	184.75	175.29	22.62	21.47	-0.95
26	44.26	56.85	60.83	146.27	143.45	138.76	2.39	-7.51	-9.04
27	100.00	78.32	54.74	137.30	178.89	221.57	55.61	84.27	12.24
28	47.24	50.39	29.00	137.59	149.89	150.01	66.54	12.42	16.62
29	49.37	58.01	36.20	149.18	160.46	148.30	67.65	-0.88	17.06
30	97.26	76.09	46.05	158.47	151.10	126.19	91.51	-32.28	26.60

注: $V_{MSE_{po3}}$ 、 $V_{MSE_{po2}}$ 、 $V_{MSE_{po1}}$ 分别为三核混合核函数、二核混合函数、单一核函数的最终预测流量曲线与实际观测流量曲线的均方误差。

由表4可知,多数情况下 $V_{MSE_{po2}}$ 比 $V_{MSE_{po1}}$ 和 $V_{MSE_{po3}}$ 大,或者介于两者之间; $V_{MSE_{po3}}$ 数次达到130以下,这是其他2种情况达不到的; $V_{MSE_{po1}}$ 保持在130~180之间,未出现超过200的情况,比其他2种更稳定,所以优先考虑均方误差能达到最小值的三核混合核函数和稳定性最好的单一核函数。构造如下形式的函数 M_F 、 M_M :

$$M_F = \frac{(E(E_{MSE3}) - E(E_{MSE2}))^2}{\sqrt{E(E_{MSE1})}}$$

(17)

$$M_M = V_{MSE_{po3}} - V_{MSE_{po1}}$$

(18)

运行遗传算法30次后得到的 M_F 与 M_M 间有强线性关系,相关系数为0.63,拟合曲线为

$$M_M = \frac{2}{5} M_F - 10$$

(19)

拟合曲线参数经T检验得到的 p -值为 $6.06 \times 10^{-5} \ll 0.05$,说明拟合效果良好。这说明 M_F 越小, M_M 越小,

故而在先验条件下,期望 M_F 数值尽量小, M_F 足够小时, M_M 会小于零,即 $V_{MSE_{p03}} < V_{MSE_{p01}}$ 。考虑到 $V_{MSE_{p01}}$ 稳定性较好,此时 $V_{MSE_{p03}}$ 会达到一个较低水平,可以采用三核混合核函数进行预测。 M_M 的最小值不是在 M_F 最小时取得,但是 M_F 取最小值,可以保证 M_M 处于一个较低水平,尽管可能不是最小值,但是可以保证预测的精度处在较高水平。

核函数参数选取与流量预测流程:(a)遗传算法确定若干组核函数参数;(b)确定每组的三核混合核函数、二核混合核函数和单一核函数的均方误差集合,并据此计算 M_F ,找到 M_F 最小的一组,记为组 1;(c)通过组 1 参数产生若干条三核混合核函数(此处是 500 条)的预测流量曲线,取中值得到最终的预测流量曲线。

3.3 结果与分析

通过遗传算法另产生 10 组参数进行验证,得到 10 组 M_F ,按照 M_F 从小到大排列得到图 5,每张子图箱线图上方为每类核函数的最终预测流量曲线与实际观测流量曲线之间的均方误差。可以看到 M_F 最低时,虽然三核混合核函数得到的最终预测流量曲线与实际观测流量曲线的均方误差不是 10 次中的最小值,但也是达到了 132 这一较低值,图 5 中 M_F 较低的组 1~3 三核混合核函数的预测结果均保持较低值:132、138、129。尽管图 5(a)中结果不是最好的(最好的结果为图 5(g)),但这在无实际观测流量的前提下,可以保证最终的预测结果精度处在较高水平(即均方误差处在较低水平),大大降低了核函数参数确定过程中的不确定性。依照 3.2 中的流程,采用图 5(a)中的支持向量机核函数参数(即使 M_F 最小的核函数参数)进行后续工作。

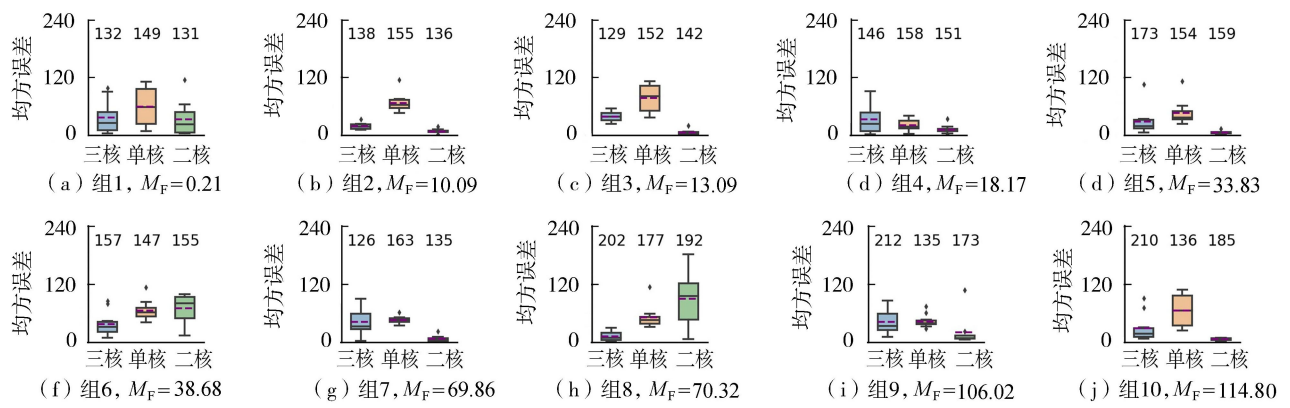


图 5 检验过程中集合 E_{MSE1} 、 E_{MSE2} 、 E_{MSE3} 的箱线图

Fig. 5 Box plot of E_{MSE1} 、 E_{MSE2} 、 E_{MSE3} in inspection process

首先考察该参数下,3 类核函数的最终预测流量曲线差异。三类核函数的最终预测流量曲线如图 6 所示。除均方误差外,再引入 2 个额外指标:相关系数与相对误差的绝对值(简称相对误差),相对误差取绝对值是方便比较。

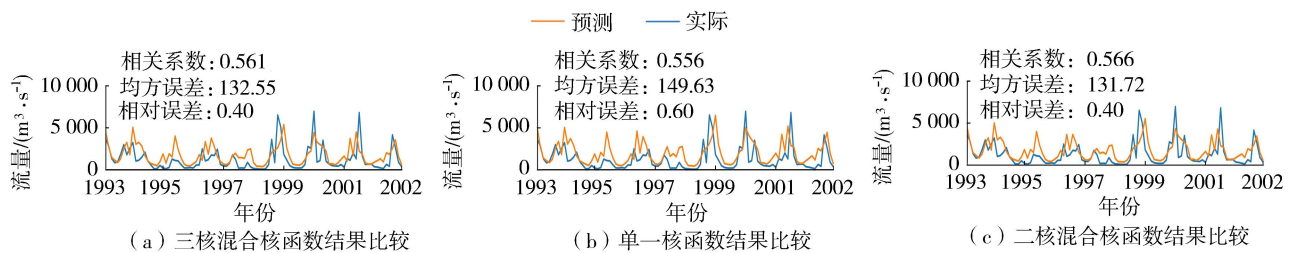


图 6 检验集上不同核函数预测结果趋势

Fig. 6 Trend chart of prediction results of different kernel functions on the test set

由图 6 可知,三核和二核混合核函数在此时基本处于同一水平,均高于单一核函数预测精度。在流量较低的年份,单一核函数的预测值较高,混合核函数相比单一核函数的预测结果更接近实测流量值;流量较高的年份,混合核函数能一定程度上降低流量峰值错峰带来的误差。三核混合核函数预测结果的均方误差能够将单一核函数预测结果的均方误差从 150 左右降低到 130 左右,使精度得以提升。根据 3.2 中流程,图 6(a)中预测流量变化曲线为最终的预测结果。

由图 7 可知,在随机的混合核比重参数下,三核混合核和二核混合核的均方误差与相对误差都可能取到远

比单一核函数差或者远比单一核函数好的情况,这说明混合核函数的混合核比重参数值无法事先确定,因为给定的混合核比重参数,可能会使预测效果变好,同样也可能会使预测效果变差,但此时混合核得到的均值和中值都好于单一核函数,与最终的预测结果基本相当。3类核函数的均值和方差的相关系数基本处于同一水平,都达到0.5以上,说明3类核函数的预测流量与实际观测流量均具有较强的相关性,预测效果均良好。

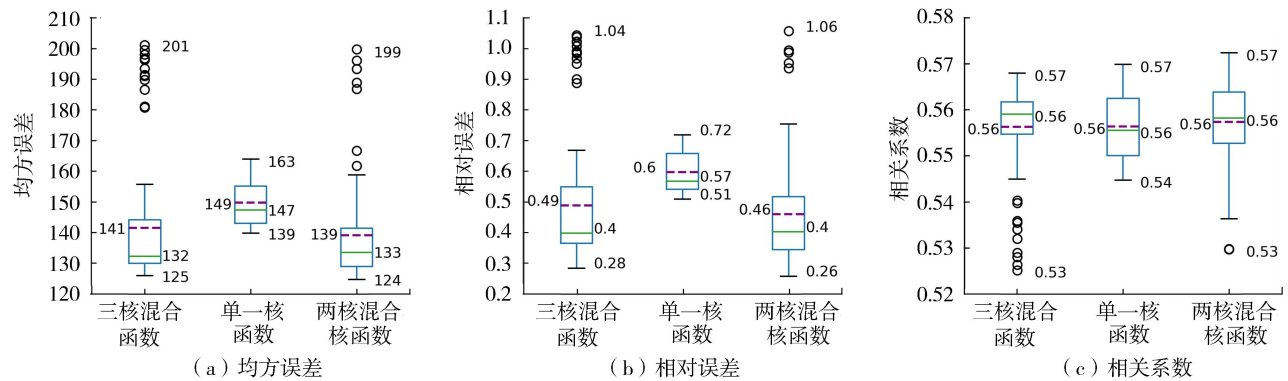


图7 不同类型核函数预测效果的均方误差、相对误差、相关系数箱形图

Fig.7 Mean square error, relative error and correlation coefficient box plots of different types of kernel function prediction effects

4 结 语

本文给出确定核函数参数的方法:通过遗传算法(或其他优化算法),给定若干组核函数参数,通过计算函数 M_F ,取 M_F 最小值时的核函数参数进行后续的预测工作。同样,为确定实际问题中混合核函数的混合核比重参数,通过随机产生足够多的混合核比重参数,得到足够多的预测流量曲线,并取这些预测流量曲线的中值作为最终的流量预测值。

多数情况下二核混合核函数的预测结果都介于三核混合核函数与单一核函数之间,尽管二核混合核函数也可能出现略好于三核混合核函数和单一核函数的情况,但这种“略好”并不明显,总是与三核混合核函数或者单一核函数较好的一方相当,所以本文主要考虑了三核混合核函数与单一核函数之间的关系,得到了核函数参数选取方法,最终的预测结果也推荐采用三核混合核函数的预测结果。

采用本文构造的方法,能够得到较好的预测结果。尽管这种预测结果可能不是最优的,但是它在流域未来流量未知的情况下,能够保证预测的精度处在所有可能预测结果中的较高水平,大大降低了核函数参数选取时带来的不确定性。

采用和构造的混合核函数是在单一核函数上产生的,能够一定程度上提升单一核函数的预测精度,但如果单一核函数出现“错峰”问题(即洪峰错位的情况),混合核函数也会出现,如何能够解决“错峰”问题,是日后亟须解决的问题。另外,本文采用的是线性组合的方式构造混合和函数,还有其他的混合函数构造方法,如直积等,也有很大的研究与发展空间。

参考文献:

[1] 李天宏,曾现进. 基于粒子群算法优化支持向量机的延河流域水沙模拟[J]. 应用基础与工程科学学报,2015,23(增刊1):79-87. (LI Tianhong, ZENG Xianjin. Water and sediment simulation of Yanhe River Basin based on particle swarm optimization optimization support vector machine[J]. Journal of Applied Basic Science and Engineering, 2015, 23(Sup1): 79-87. (in Chinese))

[2] 林剑艺,程春田. 支持向量机在中长期径流预报中的应用[J]. 水利学报,2006,37(6):681-686. (LIN Jianyi, CHENG Chuntian. Application of support vector machine in medium and long-term runoff forecasting[J]. Journal of Hydraulic Engineering, 2006,37(6): 681-686. (in Chinese))

[3] LIN Jianyi, CHENG Chuntian, CHAU K W. Using support vector machines for long-term discharge prediction [J]. Hydrological Sciences Journal,2006,51(4):599-612.

[4] 梁浩,黄生志,孟二浩,等. 基于多种混合模型的径流预测研究[J]. 水利学报,2020,51(1):112-125. (LIANG Hao,

- HUANG Shengzhi, MENG Erhao, et al. Research on runoff prediction based on multiple mixed models[J]. Journal of Hydraulic Engineering, 2020, 51(1): 112-125. (in Chinese))
- [5] NIU Wenjing, FENG Zhongkai, FENG Baofei, et al. Comparison of multiple linear regression, artificial neural network, extreme learning machine, and support vector machine in deriving operation rule of hydropower reservoir[J]. Water, 2019, 11(1): 88-112.
- [6] ATEEQ U R, GHUMMAN A R, AHMAD S, et al. Performance assessment of artificial neural networks and support vector regression models for stream flow predictions[J]. Environmental Monitoring and Assessment, 2018, 190(12): 1-20.
- [7] ADNAN R M, LIANG Zhongmi, HEDDAM S, et al. Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs[J]. Journal of Hydrology, 2020, 586: 124371.
- [8] 王红瑞, 刘晓红, 唐奇, 等. 基于小波变换的支持向量机水文过程预测[J]. 清华大学学报(自然科学版), 2010, 50(9): 1378-1382. (WANG Hongrui, LIU Xiaohong, TANG Qi, et al. Support vector machine hydrological process prediction based on wavelet transform[J]. Journal of Tsinghua University (Natural Science Edition), 2010, 50(9): 1378-1382. (in Chinese))
- [9] HE Zhibin, WEN Xiaohu, LIU Hu, et al. A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region[J]. Journal of Hydrology, 2014, 509: 379-386.
- [10] BAFITLHILE T M, LI Zhijia. Applicability of ϵ -support vector machine and artificial neural network for flood forecasting in humid, semi-humid and semi-arid basins in China[J]. Water, 2019, 11(1): 58-99.
- [11] MENG Erhao, HUANG Shengzhi, HUANG Qiang, et al. A robust method for non-stationary streamflow prediction based on improved EMD-SVM model[J]. Journal of Hydrology, 2019, 568: 462-478.
- [12] 张岩, 杨明祥, 雷晓辉, 等. 基于 PCA-PSO-SVR 的丹江口水库年径流预报研究[J]. 南水北调与水利科技, 2018, 16(5): 35-40. (ZHANG Yan, YANG Mingxiang, LEI Xiaohui, et al. Research on Danjiangkou Reservoir annual runoff forecast based on PCA-PSO-SVR[J]. South-to-North Water Diversion and Water Conservancy Science and Technology, 2018, 16(5): 35-40. (in Chinese))
- [13] BAYDAROGLU Ö, KOCAK K, DURAN K. River flow prediction using hybrid models of support vector regression with the wavelet transform, singular spectrum analysis and chaotic approach[J]. Meteorology and Atmospheric Physics, 2018, 130(3): 349-359.
- [14] 周婷, 金菊良, 李荣波, 等. 基于小波支持向量机的径流预测性能优化分析[J]. 水力发电学报, 2017, 36(10): 45-55. (ZHOU Ting, JIN Juliang, LI Rongbo, et al. Optimization analysis of runoff prediction performance based on wavelet support vector machine[J]. Journal of Hydroelectric Engineering, 2017, 36(10): 45-55. (in Chinese))
- [15] 贺玉琪, 王栋, 王远坤. BRR-SVR 月降水量预测优化模型[J]. 水利学报, 2019, 50(12): 1529-1537. (HE Yuqi, WANG Dong, WANG Yuankun. BRR-SVR monthly precipitation prediction optimization model[J]. Journal of Hydraulic Engineering, 2019, 50(12): 1529-1537. (in Chinese))
- [16] 朱双, 周建中, 孟长青, 等. 基于灰色关联分析的模糊支持向量机方法在径流预报中的应用研究[J]. 水力发电学报, 2015, 34(6): 1-6. (ZHU Shuang, ZHOU Jianzhong, MENG Changqing, et al. Application of fuzzy support vector machine method based on grey relational analysis in runoff forecasting[J]. Journal of Hydroelectric Power, 2015, 34(6): 1-6. (in Chinese))
- [17] 唐奇, 王红瑞, 许新宜, 等. 基于混合核函数 SVM 水文时序模型及其应用[J]. 系统工程理论与实践, 2014, 34(2): 521-529. (TANG Qi, WANG Hongrui, XU Xinyi, et al. SVM hydrological time series model based on hybrid kernel function and its application[J]. System Engineering Theory and Practice, 2014, 34(2): 521-529. (in Chinese))
- [18] 刘昕玥, 姚建国, 万定生. 基于组合核函数 SVM 的中小流域流量预测研究[J]. 计算机仿真, 2019, 36(11): 454-457. (LIU Xinyue, YAO Jianguo, WAN Dingsheng. Study on flow prediction of small and medium watershed based on combined kernel function SVM[J]. Computer Simulation, 2019, 36(11): 454-457. (in Chinese))
- [19] AKAIKE H. Stochastic theory of minimal realization [J]. IEEE Transactions on Automatic Control, 1974, 19(6): 667-674.
- [20] STITSON M O, WESTON J A E, GAMMERMAN A, et al. Theory of support vector machines[M]. London: Royal Holloway College, University of London, 1996.
- [21] SCHOLKOPF B, SMOLA A J. Learning with Kernels: support vector machines, regularization, optimization and beyond[M]. Cambridge, MA: The MIT Press, 2001.

(收稿日期: 2020-09-29 编辑: 刘晓艳)