

相空间重构支持向量机在径流模拟中的应用研究

李代华¹, 崔东文²

(1. 云南省水文水资源局 文山分局, 云南文山 663000; 2. 云南省文山州水务局, 云南文山 663000)

摘要:将相空间重构理论引入月径流模拟中,利用 C-C 算法进行相空间重构,将一维径流时间序列拓展为多维,基于交叉验证支持向量机(CV-SVM)原理及方法,构建以相空间重构理论与支持向量机相结合的径流时间序列模拟模型,并构建传统 BP、双隐层 BP 及 GA-BP 径流时间序列模拟模型作为对比模型,以盘龙河龙潭寨月径流时间序列为例进行分析。结果显示:基于相空间重构理论的 CV-SVM 模拟模型能较好地处理复杂的径流序列,在长达 200 个月的测试样本模拟中,平均相对误差 e_{MRE} 、最大相对误差 e_{MaxRE} 分别为 0.571 7%、5.526 7%,决定系数 DC 和合格率 QR 分别为 0.999 9 和 100%。表明该模型具有较高的泛化能力和模拟精度,模拟效果明显优于传统 BP、双隐层 BP 模型,甚至优于 GA-BP 模型;表明研究建立的基于相空间重构理论的 CV-SVM 模型用于径流模拟是合理可行的,可为径流模拟提供方法和参考。

关键词:相空间重构;支持向量机;交叉验证;混沌;径流模拟

中图分类号:P338.2;TP18

文献标志码:A

文章编号:1001-5485(2013)10-0021-06

1 研究背景

河川径流是一种复杂的水文现象,由于受多种因素的影响和制约,表现出复杂、随机、多维等特性。混沌现象是非线性确定性动力系统特有的现象,是确定性和随机性的统一,其理论所研究的是揭示貌似随机的现象背后可能隐藏的规律,以求发现复杂问题普遍所遵循的共同规律^[1]。随着混沌理论和应用技术研究的不断深入,基于混沌时间序列的建模和模拟已成为混沌信息处理研究领域中的热点,成功应用于径流模拟、预测预报中^[2-4]。

本文利用 C-C 算法进行相空间重构,将一维径流时间序列拓展为多维,基于交叉验证(Cross Validation, CV)及支持向量机(Support Vector Machine, SVM)原理及方法,构建以相空间重构理论与支持向量机相结合的径流时间序列模拟模型,并构建传统 BP、双隐层 BP 及 GA-BP 径流时间序列模拟模型作为对比模型,以盘龙河龙潭寨水文站 1952 年 4 月至 2005 年 12 月月径流时间序列为例进行分析,选取决定系数 DC 等 7 个统计学指标对所构建的 CV-SVM、传统 BP、双隐层 BP 及 GA-BP 径流模拟模型的性能进行评价,为盘龙河流域水资源开发利用及径流模拟或预测预报提供方法和参考。

2 基于相空间重构原理的 SVM 径流模拟模型

2.1 混沌时间序列的相空间重构

相空间重构的目的在于在高维相空间中恢复混沌吸引子。由于吸引子的内在行为具有不规则性及混沌吸引子具有复杂的几何结构,一般来说,不同的混沌实测数据应建立不同的混沌模型。混沌模拟是在相空间中进行的,其原理就是在相空间中找到一个非线性模型去逼近系统动态特性,实现一定时期内的模拟^[5]。

设时间序列 $x(t), t=1, 2, \dots, N$, 嵌入维为 m , 时间延迟为 τ , 则重构相空间为

$$y(t) = \{x(t), x(t+\tau), x(t+2\tau), \dots, x[t+(m-1)\tau]\}, \quad t=1, 2, \dots, M. \quad (1)$$

式中 $M=N-(m-1)\tau$ 为相空间中的相点数。根据 Takens 定理^[5], 选择合适的嵌入维和时间延迟重构相空间, 则重构相空间中的轨迹与原系统是动力学等价的。那么存在一个光滑映射 $F: R^m \rightarrow R^m$, 给出相空间轨迹的表达式为

$$y(t+\eta) = F[y(t)], \quad t=1, 2, \dots, N. \quad (2)$$

式中 η 为模拟步长。

2.2 支持向量机模拟原理

支持向量机 (Support Vector Machine, SVM) 是 20 世纪 90 年代中后期发展起来的基于统计学习理论构建的典型神经网络^[6], 它由 Vapnik 首先提出, 是一种通用的前馈神经网络, 用于解决模式分类和非线性映射问题。SVM 具有严谨的数学基础, 通过统计学习中的 VC 维理论 (Vapnik-Chervonenkis Dimension) 和寻求结构风险最小化原理来提高泛化能力, 有效解决了传统 BP 神经网络存在着学习收敛速度慢、易陷入局部极值以及网络结构难以确定等缺点, 能以任意精度逼近任意函数, 有着较好的泛化能力^[7]。SVM 应用于回归时, 其基本思想不再是寻找最优分类面将样本分开, 而是寻找一个最优超平面, 使得所有训练样本离该最优超平面距离最短, 这个超平面可看作拟合好的曲线。

SVM 实现回归步骤归纳如下^[8-9]。

(1) 步骤 1:

设含有 l 个训练样本的集合 $\{(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, \dots, l\}$, 其中, $\mathbf{x}_i (\mathbf{x}_i \in R^d)$ 是第 i 个训练样本的输入列向量, $\mathbf{x}_i = [\mathbf{x}_i^1, \mathbf{x}_i^2, \dots, \mathbf{x}_i^d]^T, \mathbf{y}_i \in R$ 为对应输出值。在高维特征中建立的线性回归函数为

$$f(\mathbf{x}) = w\Phi(\mathbf{x}) + b \quad (3)$$

式中 $\Phi(\mathbf{x})$ 为非线性映射函数。

(2) 步骤 2:

定义 ε 线性不敏感损失函数为

$$L(f(\mathbf{x}), \mathbf{y}, \varepsilon) = \begin{cases} 0, & |\mathbf{y} - f(\mathbf{x})| \leq \varepsilon; \\ |\mathbf{y} - f(\mathbf{x})| - \varepsilon, & |\mathbf{y} - f(\mathbf{x})| > \varepsilon. \end{cases} \quad (4)$$

式中: $f(\mathbf{x})$ 为回归函数返回的预测值; \mathbf{y} 为对应的真实值。

(3) 步骤 3:

类似于 SVM 分类情况, 引入松弛变量 ξ_i, ξ_i^* , 并将上述寻找 w, b 的问题用数学语言描述出来, 即

$$\begin{cases} \min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \\ \text{s. t. } \begin{cases} \mathbf{y}_i - w\Phi(\mathbf{x}_i) - b \leq \varepsilon + \xi_i; \\ -\mathbf{y}_i + w\Phi(\mathbf{x}_i) + b \leq \varepsilon + \xi_i^*, i = 1, 2, \dots, l; \\ \xi_i \geq 0, \xi_i^* \geq 0. \end{cases} \end{cases} \quad (5)$$

式中: C 为惩罚因子, C 越大表示对训练误差大于 ε 的样本惩罚越大, ε 规定了回归函数的误差要求, ε 越小表示回归函数的误差越小。求解式(5)时, 同时引入 Lagrange 函数, 并转换成对偶形式:

$$\begin{cases} \max \left[-\frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (a_i - a_i^*) (a_j - a_j^*) K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^l (a_i + a_i^*) \varepsilon + \sum_{i=1}^l (a_i - a_i^*) \mathbf{y}_i \right] \\ \text{s. t. } \begin{cases} \sum_{i=1}^l (a_i - a_i^*) = 0; \\ 0 \leq a_i \leq C; \\ 0 \leq a_i^* \leq C. \end{cases} \end{cases} \quad (6)$$

式中 $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \Phi(\mathbf{x}_j)$ 为核函数。

(4) 步骤 4:

设求解式(6)得到的最优解为 $\mathbf{a} = [a_1 \ a_2 \ \dots \ a_l], \mathbf{a}^* = [a_1^* \ a_2^* \ \dots \ a_l^*]$, 则有

$$w^* = \sum_{i=1}^l (a_i - a_i^*) \Phi(\mathbf{x}_i) \quad ; \quad (7)$$

$$b^* = \frac{1}{N_{\text{nsv}}} \left\{ \sum_{0 < a_i < C} [\mathbf{y}_i - \sum_{\mathbf{x}_j \in \text{SV}} (a_i - a_i^*) K(\mathbf{x}_i, \mathbf{x}_j) - \varepsilon] + \sum_{0 < a_i < C} [\mathbf{y}_i - \sum_{\mathbf{x}_j \in \text{SV}} (a_i - a_i^*) K(\mathbf{x}_i, \mathbf{x}_j) + \varepsilon] \right\} \quad (8)$$

式中 N_{nsv} 为支持向量机个数。

(5) 步骤 5:

将 w^*, b^* 代入式(3)得到回归函数为

$$f(\mathbf{x}) = \sum_{i=1}^l (a_i - a_i^*) K(\mathbf{x}_i, \mathbf{x}) + b^* \quad (9)$$

式中, 只要部分参数 $(a_i - a_i^*)$ 不为 0, 其对应的样本 \mathbf{x}_i 即为问题中的支持向量。

研究表明, 只要满足 Mercer 条件的对称函数即可作为 SVM 核函数, 常用的核函数主要类型有线性核函数 ($K(\mathbf{x}, \mathbf{x}_i) = \mathbf{x}^T \mathbf{x}_i$)、多项式核函数 ($K(\mathbf{x}, \mathbf{x}_i) = (\gamma \mathbf{x}^T \mathbf{x}_i + r)^p, \gamma > 0$)、径向基核函数 ($K(\mathbf{x}, \mathbf{x}_i) = \exp(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2), \gamma > 0$) 和两层感知核函数 ($K(\mathbf{x}, \mathbf{x}_i) = \tanh(\gamma \mathbf{x}^T \mathbf{x}_i + r)$)。核函数可以看成是实际问题的特征提取过程, 核函数的合理选取有助于提高模型精度。SVM 模型选择主要有 2 个步骤, 一是核函数的选择; 二是 SVM 本身的 2 个参数以及所选取的核函数所对应的参数。普遍认为, 基于径向基函数的 SVM 模型有着较好的模拟效果, 本文选择径向基核函数为 SVM 的核函数。

2.3 基于相空间重构的 SVM 模拟方法

基于相空间重构的 SVM 模拟模型是将原始时间序列数据进行相空间重构得到一个高维的特征空间, 获得模拟模型所需要的输入向量和输出向量。设观测到的时间序列为 $\{\mathbf{x}(t_i), i = 1, 2, \dots, N\}$, 分别计算时延 τ 与嵌入维 m , 对该时间序列进行相空间重构, 根据 Takens 定理^[2,5]有

$$\mathbf{X}_{i+T} = f(\mathbf{x}), \quad i = 1, 2, \dots, M. \quad (10)$$

式中: \mathbf{X}_i 为相空间的第 i 个相点; T 为前向预测步长; M 为 m 维相空间中的相点数。

一般情况, 我们取 $T=1$, 即前向预测一步, 则输入、输出向量 \mathbf{X} 和 \mathbf{Y} 分别为:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}(t_1) & \mathbf{x}(t_1 + \tau) & \cdots & \mathbf{x}[t_1 + (m-1)\tau] \\ \mathbf{x}(t_2) & \mathbf{x}(t_2 + \tau) & \cdots & \mathbf{x}[t_2 + (m-1)\tau] \\ & & \vdots & \\ \mathbf{x}(t_{M'}) & \mathbf{x}(t_{M'} + \tau) & \cdots & \mathbf{x}[t_{M'} + (m-1)\tau] \end{bmatrix};$$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{x}(t_1 + m\tau) \\ \mathbf{x}(t_2 + m\tau) \\ \vdots \\ \mathbf{x}(t_{M'} + m\tau) \end{bmatrix}. \quad (11)$$

式中 M' 为满足 $t_{M'} + m\tau = t_N$ 的整数, 显然 $M' < M$ 。

在重构相空间后, 即可对 SVM 进行训练, 得到 t 时刻 SVM 的进一步预测模型为

$$\hat{\mathbf{x}}_{t+1} = \sum_{i=1}^{M'} (a_i - a_i^*) K((\mathbf{x}_t, \mathbf{x}) + \mathbf{b}^* \quad (12)$$

为表述方便, 令 $\mathbf{x}_{i+1} = \mathbf{x}(t_i)$, 则对于相空间的第 $t+1$ 点, 有

$$\mathbf{x}_{i+1} = (\hat{\mathbf{x}}_{t+1}, \dots, \mathbf{x}(t - (m-2)\tau)) \quad (13)$$

由式(9)可得到对 $t+2$ 点的模拟为

$$\hat{\mathbf{x}}_{t+2} = \sum_{i=1}^{M'} (a_i - a_i^*) K((\mathbf{x}_t, \mathbf{x}_{i+1}) + \mathbf{b}^* \quad (14)$$

依此类推, 第 p 步的 SVM 模拟模型为

$$\hat{\mathbf{x}}_{t+p} = \sum_{i=1}^{M'} (a_i - a_i^*) K((\mathbf{x}_t, \mathbf{x}_{i+p-1}) + \mathbf{b}^* \quad (15)$$

但由于本文着重研究同期模拟, 即 $T=0$, 故模型的输入、输出向量 \mathbf{X} 和 \mathbf{Y} 分别为

$$\mathbf{X} = \begin{bmatrix} x(t_1) & x(t_1 + \tau) & \cdots & x(t_1 + m\tau) \\ x(t_2) & x(t_2 + \tau) & \cdots & x(t_2 + m\tau) \\ & & \vdots & \\ x(t_{M'}) & x(t_{M'} + \tau) & \cdots & x(t_{M'} + m\tau) \end{bmatrix};$$

$$\mathbf{Y} = \begin{bmatrix} x(t_1 + m\tau) \\ x(t_2 + m\tau) \\ \vdots \\ x(t_{M'} + m\tau) \end{bmatrix}. \quad (16)$$

3 实例应用

3.1 基本资料

龙潭寨水文站位于盘龙河上游, 设立于 1951 年

4 月, 属国家基本站, 观测项目有水位、流量、降水、蒸发、泥沙和水温, 系列均为 1951 年 4 月至 2005 年 12 月。控制流域面积 3 128 km², 流域多年平均流量 24.6 m³/s, 降水量 956.8 mm。本文研究龙潭寨水文站从 1951 年 4 月至 2005 年 12 月共 54 a 历月实测径流量, 将资料按月展开, 得到 657 个数据的月径流时间序列。

3.2 确定重建相空间滞时和嵌入维数

在重构相空间中, 延迟时间 τ 和嵌入维数 m 的选取具有十分重要的意义^[5]。关于时延 τ 与嵌入维 m 的选取, 现在主要有 2 种观点: 一种观点认为两者是互不相关的, 即 τ 和 m 的选取是独立进行的, 如求时延的自相关函数法^[10]、互信息法^[11], 求嵌入维的 G-P 算法^[12]或假最近邻法^[13]等; 另一种观点认为两者是相关的, 即 τ 和 m 的选取是互相依赖的, 如嵌入窗法^[14]、C-C 方法^[15], 可同时计算出时延和时间窗口。

实际应用中, 若时延 τ 太小, 则相空间矢量在数值上非常接近, 以至于无法区分, 从而无法提供 2 个独立的坐标分量; 若时间延迟 τ 太大, 则两坐标分量在统计意义上又是完全独立的, 混沌吸引子的轨迹在两个方向的投影毫无相关性可言。对于嵌入维数 m , 若 m 太小, 吸引子会发生折叠甚至自相交; 若 m 过大, 会在额外的维数上引入噪声, 增加计算量。因此 τ 和 m 选取的原则是既保证能准确计算各种混沌不变量, 又尽量降低计算量和噪声的影响。本文采用 C-C 法相空间重构参数, 得到延迟时间 $\tau=2$, 时间窗口 $\tau_w=8$, 根据公式 $\tau_w=(m-1)\tau$ 反推, 求得嵌入维数 $m=5$ 。

3.3 径流时间序列演变混沌特性识别

目前关于混沌时间序列识别方法主要有关联维数法、Kolmogorov 熵法、Lyapunov 指数法、功率谱方法、庞卡莱截面法等众多方法, 但尚无可以在混沌系统和随机系统之间做出准确判断的统一方法^[5,16]。本文用 Lyapunov 指数法来判断龙潭寨水文站月径流时间序列演变的混沌特性。确定 Lyapunov 指数的方法较多, 主要有 Wolf 方法、Jacobian 法、 p -范数法和小数据量法等。本文利用小数据量法求得龙潭寨水文站月径流序列的最大 Lyapunov 指数以 2 为底时等于 0.541 4, 以 e 为底时等于 0.375 2, 大于零, 根据若最大 Lyapunov 指数大于零, 则系统一定存在混沌特性的判别方法可知, 龙潭寨水文站月径流序列的演变具有混沌特性。因此可以用上述基于相空间重构理论的 CV-SVM 等模型对月径流时间序列进

行模拟。

3.4 径流模拟的实现

3.4.1 数据处理

数据处理的方法很多,本文采用以下方法将各指标数据无量纲化到 $[0.1, 0.9]$ 之间,有利于网络训练。公式如下:

$$\hat{x} = 0.1 + (0.9 - 0.1) \left[\frac{(x - x_{\min})}{(x_{\max} - x_{\min})} \right] \quad (17)$$

式中: \hat{x} 为经过标准化处理的数据; x 为原始数据; x_{\max} 和 x_{\min} 分别为数据序列中的最大数和最小数。

3.4.2 训练及测试样本设计

由上述可知,将龙潭寨水文站 1951 年 4 月至 2005 年 12 月的实测资料按月展开,得到 657 个月径流时间序列数据,由于延迟时间 $\tau = 2$ 、嵌入维数 $\tau = 5$,对龙潭寨水文站径流时间序列进行相空间重构,得到输入、输出样本各 647 个。为能更客观地评价模型的泛化能力,本文选取前 447 个样本作为训练样本,后 200 个样本作为测试样本。

3.4.3 SVM 设计

SVM 用于处理模式分类或非线性映射问题时,在选定径向基核函数条件下,模型中的惩罚因子 C 和核函数参数 γ 的选取对模型的模拟精度有着关键性影响^[7]。由于惩罚因子和核函数参数的选取目前尚无理论上的指导原则,最优参数的选取多凭经验、实验对比等进行搜寻,极大地制约了 SVM 模型精度^[8-9]。目前普遍采用交叉验证法(Cross Validation, CV)来搜寻合理的参数。CV 是用来验证分类器性能的一种统计分析方法,基本思想是把在某意义下将原始数据(dataset)进行分组,一部分作为训练集(train set),另一部分作为验证集(validation set),首先用训练集对分类器进行训练,再利用验证集来测试训练得到的模型(model),以此来作为评价分类器的性能指标。CV 方法可以有效地避免模型“过学习”以及“欠学习”现象的发生。

本文基于 MATLAB 环境和 libsvm 工具箱,构建基于相空间重构理论的 CV-SVM、传统 BP、双隐层 BP 及 GA-BP 月径流序列模拟模型。以相空间重构后的龙潭寨水文站径流时间序列为输入、输出向量,即输入节点数 5 个,输出节点为 1 个。本例经反复调试,在下述参数搜索范围内及设置条件下, CV-SVM、传统 BP、双隐层 BP 及 GA-BP 月径流时间序列模型具有较好的模拟效果。

CV-SVM 模型:惩罚因子 c 和核函数参数 γ 的搜索空间均设置为 $2^{-2} \sim 2^{10}$, K 取默认值 3 (K 为交叉验证参数,表示将测试集分几部分进行), γ 和 c

的步进大小均取 0.5 时,寻优得到 γ 和 c 参数最佳值分别为 4, 0.5, 此时 CV-SVM 月径流时间序列模型具有较好的模拟效果。

传统 BP 模型:由于 BP 网络各隐层神经元数的选取,目前并没有统一的计算方法,本文采用目前较为普遍的 Kolmogorv 定理^[17-18] 确定隐层单元数,最终确定 BP 模型结构为 5—11—1, 隐含层和输出层传递函数分别采用 tansig 和 purelin, 训练函数采用 traingdx (该算法具有附加动量和自适应调整算法的优点,在一定程度上能有效避免收敛速度慢、陷入局部极值), 学习速率 l_r 为 0.1, 设定期望误差为 0.000 01, 最大训练轮回为 5 000 次时,传统 BP 月径流时间序列模型具有较好的模拟效果。

双隐层 BP 模型:基于模型间的公平比较原则,在上述传统 BP 模型的基础上增加一个隐含层,结构为 5—11—7—1, 其他传递函数、期望误差、最大训练轮回次数等均与传统 BP 模型相同,此时双隐层 BP 月径流时间序列模型具有较好的模拟效果。

GA-BP 模型:同样基于公平原则,GA-BP 模型的网络结构、隐含层和输出层传递函数、期望误差、最大训练轮回数等均与上述传统 BP 模型相同,在此条件下运用 GA 来优化 BP 神经网络的初始权值和阈值。GA 参数设置为:种群规模为 50, 进化次数为 50 次,交叉概率为 0.1, 变异概率为 0.05, 此时 GA-BP 月径流时间序列模型具有较好的模拟效果。

3.4.4 性能评价

依据水文情报预报规范^[19] 及文献^[20], 选取平均相对误差 e_{MRE} 、最大相对误差 e_{MaxRE} 、决定系数 DC、合格率 QR、平均绝对误差 e_{MAE} 、均方根绝对误差 e_{RMSE} 和均方根相对误差 e_{RMAPE} 7 个统计学指标作为模型的评价指标。决定系数 DC 范围在 $[0, 1]$ 内, 愈接近 1, 表明模型的性能越好; 合格率 QR 是以相对误差小于 20% 为合格标准, 计算相对误差小于 20% 的个数占总体的百分比; 其他评价指标越小, 表明模型的性能越好。

评价指标计算公式如下:

$$e_{\text{MRE}} = \frac{1}{n} \sum_{i=1}^n \frac{|\hat{y}_i - y_i|}{y_i} \times 100\% \quad ; \quad (18)$$

$$e_{\text{MaxRE}} = \max_{1 \leq i \leq n} \frac{|\hat{y}_i - y_i|}{y_i} \quad ; \quad (19)$$

$$\text{DC} = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad ; \quad (20)$$

$$\text{QR} = \frac{k}{n} \times 100\% \quad ; \quad (21)$$

$$e_{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad ; \quad (22)$$

$$e_{RMSE} = \frac{1}{n} \sqrt{\sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad ; \quad (23)$$

$$e_{RMAPE} = \frac{1}{n} \sqrt{\sum_{i=1}^n \left[\frac{|\hat{y}_i - y_i|}{y_i} \right]^2}, \quad i = 1, 2, \dots, n. \quad (24)$$

式中： \hat{y}_i 为第 i 个样本模拟值； y_i 为第 i 个样本实测值； \bar{y} 为实测值的均值； k 为相对误差小于 20% 的模拟样本数； n 为模拟样本数。

3.5 模拟结果及分析

利用上述训练好的基于相空间重构理论的 CV-SVM、传统 BP、双隐层 BP 及 GA-BP 月径流序列模拟模型对龙潭寨水文站后 200 组月径流进行模

表 1 CV-SVM、传统 BP、双隐层 BP 及 GA-BP 模型模拟效果评价

Table 1 Evaluation of the simulation results of CV-SVM, traditional BP, double hidden layers of BP and GA-BP model

模拟模型	$e_{MRE}/\%$	$e_{MaxRE}/\%$	DC	QR/%	e_{MAE}	e_{RMSE}	e_{RMAPE}
CV-SVM	0.571 7	5.526 7	0.999 9	100	0.079 5	0.009 9	0.000 8
传统 BP	10.404 4	53.934 0	0.981 9	84.5	1.803 3	0.168 7	0.010 3
双隐层 BP	6.199 0	28.710 8	0.992 7	96.5	1.079 6	0.107 1	0.006 0
GA-BP	1.128 8	6.803 5	0.999 8	100	0.183 1	0.016 9	0.001 2

分析图 2 和表 1 可以得出以下结论：

(1) 基于相空间重构理论的 CV-SVM 径流模拟模型能较好地处理复杂的水文序列,在长达 200 个月的测试样本模拟中, CV-SVM 模拟模型的 DC 和 QR 均是 4 种模型中最大的(除 GA-BP 模型的合格率同样为 100% 外),其他 5 个评价指标均为最小,其平均相对误差 e_{MRE} 、最大相对误差 e_{MaxRE} 、决定系数 DC 和合格率 QR 分别为 0.571 7%、5.526 7%、0.999 9 和 100%,表明该模型具有较高的泛化能力和模拟精度,模拟效果明显优于传统 BP、双隐层 BP 网络模型,甚至优于 GA-BP 模拟模型,表明研究建立的基于相空间重构理论的 CV-SVM 模型用于径流模拟是合理可行的,可为径流模拟预报提供方法和参考。

(2) 双隐层 BP 模型的 7 个评价指标均优于传统 BP,表明增加隐含层数可以提高 BP 网络的泛化能力和模拟精度,但难点是如何确定最佳网络结构和各类参数,本例中双隐层 BP 网络所确定的结构和参数理论上仅仅是较佳,而非最佳。

(3) 在相同网络结构、传递函数、训练函数及期望误差等条件下, GA-BP 模型的 7 个评价指标均优于传统 BP,同样优于双隐层 BP 模型,表明 GA 能有效优化 BP 网络初始权值和阈值。

拟,模拟结果及评价见图 1 和表 1。

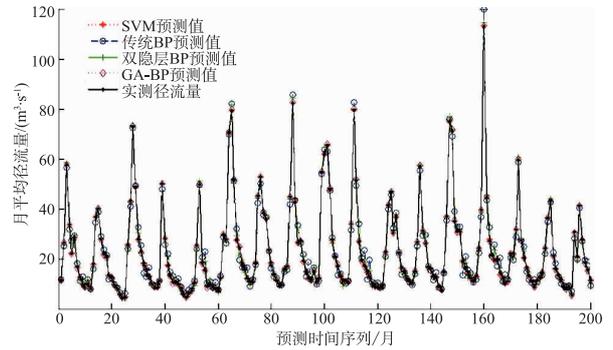


图 1 CV-SVM、传统 BP、双隐层 BP 及 GA-BP 模型模拟结果比较

Fig. 1 Comparison of the simulation results among CV-SVM, traditional BP, double hidden layers of BP and GA-BP model

4 结 语

(1) 利用相空间重构理论,将龙潭寨水文站一维径流时间序列拓展为多维序列,构造 447 组训练样本和 200 组测试样本,采用 CV 方法优化 SVM 模型的惩罚因子和核函数参数,构建 CV-SVM、传统 BP、双隐层 BP 及 GA-BP 月径流序列模拟模型,结果表明基于相空间重构理论的 CV-SVM 月径流时间序列模拟模型具有适应性强、全局优化、训练时间短等特点,模型有着较高的模拟精度和泛化能力。

(2) 作为对比模型,本文基于公平比较的原则,采用循环算法确定最佳 BP 神经网络结构,在训练次数及期望误差等尽可能相同条件下,采用增加 BP 隐层数和利用 GA 优化 BP 网络初始权值及阈值的方法,构建双隐层 BP 及 GA-BP 径流模拟模型,模拟结果表明增加隐含层数可以提高 BP 网络的泛化能力和模拟精度,GA 能有效优化 BP 网络初始权值和阈值。

(3) 由于 GA 算法可能存在早熟收敛、易陷入局部极值等缺点和不足,针对多维、复杂的月径流时间序列 GA-BP 模拟模型,利用改进的 GA 算法可以有效避免其早熟收敛、易陷入局部极值等现象,如多

种群遗传算法、量子遗传算法和免疫遗传算法等,以期进一步提高 GA-BP 模型的泛化能力和模拟精度。

参考文献:

- [1] 吕金虎,陆君安,陈士华.混沌时间序列分析及其应用[M].武汉:武汉大学出版社,2002.(LV Jin-hu, LU Jun-an, CHEN Shi-hua. Chaotic Time Series Analysis and Its Applications [M]. Wuhan: Wuhan University Press, 2002. (in Chinese))
- [2] 于国荣,夏自强.混沌时间序列支持向量机模型及其在径流模拟中应用[J].水科学进展,2008,19(1):116-122.(YU Guo-rong, XIA Zi-qiang. Prediction Model of Chaotic Time Series Based on Support Vector Machine and Its Application to Runoff[J]. Advances in Water Science, 2008, 19(1): 116-122. (in Chinese))
- [3] 陈南祥,黄强,曹连海,等.径流序列的相空间重构神经网络预测模型[J].河海大学学报(自然科学版),2005,33(5):490-493.(CHEN Nan-xiang, HUANG Qiang, CAO Lian-hai, et al. Neural Network Forecasting Model for Phase Space Reconstruction of Runoff Series [J]. Journal of Hohai University (Natural Sciences), 2005, 33(5): 490-493. (in Chinese))
- [4] 战国隆,马孝义,刘继龙,等.泾河月径流量的混沌特征识别及模拟[J].水利水电科技进展,2010,30(1):7-9.(ZHAN Guo-long, MA Xiao-yi, LIU Ji-long, et al. Chaotic Feature Recognition and Simulation of Monthly Runoff of Jinghe River[J]. Advances in Science and Technology of Water Resources, 2010, 30(1): 7-9. (in Chinese))
- [5] 邵东国,刘丙军,阳书敏,等.水资源繁杂系统理论[M].北京:科学出版社,2012.(SHAO Dong-guo, LIU Bing-jun, YANG Shu-min, et al. Theory of Complex Water Systems[M]. Beijing: Science Press, 2012. (in Chinese))
- [6] Vladimir N. Vapnik. 统计学习理论的本质[M].张学工译.北京:清华大学出版社,2000.(VAPNIK V N. The Nature of Statistical Learning Theory [M]. Translated by ZHANG Xue-gong. Beijing: Tsinghua University Press, 2000. (in Chinese))
- [7] 田景文,高美娟.人工神经网络算法研究及应用[M].北京:北京理工大学出版社,2006.(TIAN Jing-wen, GAO Mei-juan. Artificial Neural Network Algorithm Research and Application[M]. Beijing: Beijing Institute of Technology Press, 2006. (in Chinese))
- [8] MATLAB 中文论坛. MATLAB 神经网络 30 个案例分析[M].北京:北京航空航天大学出版社,2010.(MATLAB Chinese Forum. 30 Case Analysis of MATLAB Neural Network [M]. Beijing: Beihang University Press, 2010. (in Chinese))
- [9] 崔东文.支持向量机在湖库营养状态识别中的应用研究[J].水资源保护,2013,29(4):26-30.(CUI Dong-wen. Application of Support Vector Machine to Lake and Reservoir Trophic Status Recognition [J]. Water Resources Protection, 2013, 29(4): 26-30. (in Chinese))
- [10] KANTZ H, SCHREIBER T. Nonlinear Time Series Analysis [M]. Cambridge: Cambridge University Press, 1997.
- [11] FRASER A M, SWINNEY H L. Independent Coordinates for Strange Attractors from Time Series [J]. Physical Review A, 1986, 33(2): 1134-1140.
- [12] GRASSBERGER P, PROCACCIA I. Measuring the Strangeness of Strange Attractors [J]. Physica D, 1983, 9(1/2): 189-208.
- [13] KENNEL M B, BROWN R, ABARBANEL H D I. Determining Embedding Dimension for Phase-Space Reconstruction Using a Geometrical Construction [J]. Physical Review A, 1992, 45(6): 3403-3407.
- [14] KUGIURMTZIS D. State Space Reconstruction Parameters in the Analysis of Chaotic Time Series-the Role of the Time Window Length [J]. Physica D, 1996, 95(1): 13-28.
- [15] KIM H S, EYKHOLT R, SALAS J D. Nonlinear Dynamics, Delay Times and Embedding Windows [J]. Physica D, 1999, 127(1/2): 48-60.
- [16] 吴学文.考虑生态多目标水电站水库混沌优化调度研究[M].北京:中国水利水电出版社,2012.(WU Xue-wen. Multi-objective Optimization Scheduling of Hydro-power Station in Consideration of Ecological Chaos [M]. Beijing: China Water Power Press, 2012. (in Chinese))
- [17] 张德丰. MATLAB 神经网络应用设计 [M]. 北京:机械工业出版社,2009.(ZHANG De-feng. MATLAB Neural Network Design [M]. Beijing: Mechanical Industry Press, 2009. (in Chinese))
- [18] 张良均,曹晶,蒋世忠.神经网络实用教程[M].北京:机械工业出版社,2008.(ZHANG Liang-jun, CAO Jing, JIANG Shi-zhong. Practical Guide to Neural Networks [M]. Beijing: Mechanical Industry Press, 2008. (in Chinese))
- [19] SL250—2000,水文情报预报规范[S].(SL250—2000, Standard for Hydrological Information and Hydrological Forecasting [S]. (in Chinese))
- [20] 崔东文.基于极限学习机的长江流域水资源开发利用综合评价[J].水利水电科技进展,2013,33(2):14-19.(CUI Dong-wen. Comprehensive Evaluation of Water Resources Development and Utilization in Yangtze River Basin Based on Extreme Learning Machine [J]. Advances in Science and Technology of Water Resources, 2013, 33(2): 14-19. (in Chinese))

(编辑:刘运飞)

(下转第 31 页)