DOI:10.11705/j.issn.1672-643X.2016.05.04

基于改进的神经网络与支持向量机的小流域日径流量预测研究

马乐宽[1], 邱 瑀[2], 赵 越[1], 李 雪[3], 王玉秋[2]
(1. 环境保护部环境规划院, 北京 100012; 2. 南开大学 环境科学与工程学院,
天津 300350;3. 天津师范大学 水资源与水环境重点实验室, 天津 300387)

摘 要: 数据驱动水文模型可以在不考虑复杂物理过程的情况下,实现对数据种类较少的小流域日径流量的准确预测。本研究基于安徽省黄山市月潭水文监测站点2009-2012年的日径流量监测数据,分别构建粒子群寻优算法改进的神经网络(PSO-BPNN)以及支持向量机(PSO-SVM)模型。通过进行不同形式的模型结果比较发现,两类模型均有较好的拟合能力及泛化能力,其中基于三日流量数据的(PSO-SVM)模型具有最优模拟结果,可以考虑用于月潭流域日径流量的预测,实现流域内水资源的合理配置以及相关灾害的预防。

关键词: 日径流量; 神经网络; 支持向量机; 粒子群寻优; 日径流量预测
中图分类号:TV121.4 **文献标识码**: A **文章编号**: 1672-643X(2016)05-0023-05

Prediction of daily runoff in a small watershed based on improved neural network and support vector machine (SVM)

MA Lekuan[1], QIU Yu[2], ZHAO Yue[1], LI Xue[3], WANG Yuqiu[2]
(1. *Chinese Academy for Environmental Planning, Ministry of Environmental Protection, Beijing* 100012, *China*;
2. *College of Environmental Science and Engineering, Nankai University, Tianjin* 300350, *China*;
3. *Tianjin Key Laboratory of Water Resources and Environment, Tianjin Normal University, Tianjin* 300387, *China*)

Abstract: Data-driven hydrological model can realize the accurate prediction for daily runoff in small watershed with less data types without considering the complex physical processes. Based on daily runoff monitoring data from Yuetan hydrological monitoring sites, in Huangshan of Anhui Province from 2009 to 2012, the paper built particle swarm optimization algorithm improved neural network (PSO-BPNN) and support vector machine (PSO-SVM) model. By comparing the results from different types of model, it discovered that both of the two models have good fitted ability and generalization ability, and PSO-SVM model based on three-day runoff data has the best simulation results and it can be used to predict daily runoff of Yuetan Basin and realize the rational allocation of water resources in the basin as well as the early prevention of related disasters.

Key words: daily runoff; neural network; support vector machine; particle swarm optimization; prediction of daily runoff

日径流量的准确预测对于水资源的有效配置及管理具有十分重要的意义,同时也是水文学研究的重要组成内容。近年来涌现出许多径流量预测方法,主要可以分为两类,基于物理过程的水文模型及数据驱动模型。一般来说,基于物理过程的模型需要将径流产生过程概化为若干复杂的数学公式,以大量的监测数据作为参数校准基础,同时需要研究者具备模型应用经验,对于数据获取种类及数量比较匮乏的流域有一定的应用限制[1-2]。而数据驱动模型不考虑复杂的物理水文过程,通过挖掘数据之

收稿日期:2016-04-13; **修回日期**:2016-06-14
基金项目:天津师范大学博士基金项目(52XB1517)
作者简介:马乐宽(1979-),男,重庆人,副研究员,博士,主要从事生态水文学、水环境规划与管理研究。
通讯作者:王玉秋(1965-),男,黑龙江齐齐哈尔人,教授,博士生导师,主要从事水环境管理决策与技术支持、流域及水源地保护战略等研究。

间的潜在关系实现径流量数据的模拟,建模过程所需数据种类较少,适用于以径流量的准确预测作为主要关注目标的研究。目前应用比较广泛的数据驱动方法包括人工神经网络(ANN)与支持向量机(SVM)[3-4]。与传统分析方法相比,这两种方法能够在大量庞杂的数据间建立准确的非线性相关关系,并具有较好的泛化能力。Nourani 等[5]利用多元小波神经网络的方法实现了降雨-径流之间关系的模拟,Taormina 等[6]利用人工神经网络的方法基于小时时间尺度对地下水位进行了模拟,Wu 等[7]应用支持向量回归机的方法对河流水位的变化进行了预测,结果表明 ANN 和 SVM 方法均可以实现水文数据的预测模拟。但是在实践中发现,应用经典 ANN 和 SVM 方法在参数确定过程中具有一定的主观性和随机性,可能造成预测结果的偏差及不稳定,因此本研究拟分别利用粒子群优化算法(PSO)改进 ANN 及 SVM 方法[8-9],针对安徽省月潭流域进行日径流量的预测,通过预测结果比较两种方法的准确性及适用性。

1 研究方法

1.1 人工神经网络

近年来人工神经网络(ANN)被广泛应用于诸多领域处理复杂的非线性数据集,针对所解决问题的不同,应用的网络类型也不尽相同,其中多层反向传播神经网络(BPNN)是应用最多的形式之一[10-11]。典型的 BPNN 包括输入层、隐含层与输出层三层结构,其中输入层神经元的个数与输入变量数量相等,输入信号经过计算和处理转化为输出信号反映在输出层上。作为有监督的学习算法,当估算的输出结果(输出信号)与正确的期望结果(监测数据)之间存在误差时,网络将通过自动的调节机制向误差减小的方向调整相应的联结强度,经过多次重复训练,最终得到与实测数据误差最小的训练结果。由于网络联结权重的初始值选择对于收敛结果有非常显著的影响,完全依靠主观判断随意赋值可能造成结果不收敛或陷入局部最小值,因此本研究拟利用粒子群优化算法以及交叉验证改进网络结构,提高模型的精度及结果稳定性。

1.2 支持向量机

支持向量机(SVM)算法可依据实现功能的不同分别用于数据分类及回归预测[12-13]。SVM 是一种基于统计学习理论的数据挖掘方法,其原理为寻找一个满足要求的最优超平面,在保证分类精度的同时使超平面两侧的边缘距离最大。给定一组训练数据 $(x_1, y_1), (x_2, y_2), \cdots, (x_i, y_i), \cdots, (x_n, y_n)$,其中 x_i 是输入向量,y_i 是与之相对应的实际观测值,n 代表样本数量,则支持向量回归机的方程形式如式(1)所示:

$$f(x) = (w,x) + b \tag{1}$$

式中:w 为与维数相同的权重向量;b 为误差项,最佳回归方程的形式通过引入不敏感损失函数 ε 并求得其极小值得到,即如式(2)所示:

$$\min\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)\right) \tag{2}$$

$$\text{满足约束条件}\begin{cases} y_i - \langle w,x_i\rangle - b \leqslant \varepsilon + \xi_i \\ \langle w,x_i\rangle + b - y_i \leqslant \varepsilon + \xi_i^* \\ \xi_i,\xi_i^* \geqslant 0 \end{cases} \tag{3}$$

式中:$i = 1, 2, \cdots, n$;ξ_i 和 ξ_i^* 分别为松弛变量的上限和下限;C 为惩罚因子,通过调节 C 可以使模型在训练误差与泛化能力之间得到平衡,该值一般由使用者直接赋值。通过式(4)所示的优化方程。

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x,x_i) + b \tag{4}$$

式中:α_i 和 α_i^* 为拉格朗日方程的待定系数;$K(\cdot)$ 为空间变换过程中应用的核函数。

$$\text{满足约束条件}\begin{cases} \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\ 0 \leqslant \alpha_i \leqslant C \\ 0 \leqslant \alpha_i^* \leqslant C \end{cases} \tag{5}$$

可以计算得到 α_i 和 α_i^*,进而确定回归方程的系数和常数项。$K(\cdot)$ 为空间变换过程中应用的核函数,有几种不同的形式可供选择。决定支持向量回归机预测能力的参数主要包括:惩罚因子 C,核函数以及核函数中涉及到的参数。本研究通过参考前人的研究成果,选择径向基函数作为核函数(式6)。

$$K(x,x_i) = \exp\left(-\frac{\|x,x_i\|^2}{c^2}\right) \tag{6}$$

式中:$K(\cdot)$ 为径向基函数;c 为径向基函数的参数。其中径向基函数中涉及到的参数 c 即为决定模型预测能力的主要因子之一。为了避免主观判断带来的误差,本研究拟利用粒子群寻优算法确定最优的惩罚因子 C 及核函数参数 c。

1.3 粒子群寻优算法

粒子群寻优(PSO)是 Eberhart 和 Kennedy 在 1995 年提出的一种优化算法。在 PSO 中,每个粒子最初都被随机赋予一个速度和位置,下一时刻的位置则是根据当前种群最优粒子的位置以及自身遍历的

最优位置决定的。在待解决的问题域中,每一个粒子的最优位置都通过追溯其上一个最优位置获得,记为 p_{best}。其可能的飞行方向则通过 p_{best} 和其临近粒子的最优位置 g_{best} 共同决定。问题域的维数定义为 D,粒子总数为 n,第 i 个粒子的位置可以用向量 $X_i(x_{i1},x_{i2},\cdots,x_{iD})$ 表示,移动速度表示为 $v_i=(v_{i1},v_{i2},\cdots,v_{iD})$,则整个寻优过程可以表示为式(7)。

$$\begin{cases} v_{id}(k+1)=\omega v_{id}(k)+c_1\text{rand}\cdot[p_{bestid}-x_{id}(k)]+ \\ \qquad c_2\text{rand}\cdot[g_{bestid}-x_{id}(k)] \\ x_{id}(k+1)=x_{id}(k)+v_{id}(k+1) \end{cases} \tag{7}$$

式中:k 和 $k+1$ 为迭代次数;c_1 为自身因子;c_2 为全局因子,二者均为常数,rand 为[0,1] 之间的随机数;ω 为粒子的当前速度受其上一时刻速度影响的惯性因子。PSO 可以作为寻优方法结合应用到 BPNN 和 SVM 中,通过计算适应函数确定最优参数,如式(8)所示。

$$f(x_i)=\sum_{k=1}^{S}[t_k-p_k(x_i)] \tag{8}$$

式中:$f()$ 为适应函数值;S 为训练样本总数;t_k 为实际观测值;p_k 为预测值。在 PSO-BPNN 中,x_i 为输入层与隐含层之间以及隐含层与输出层之间的连接权重矩阵,在 PSO-SVM 中,x_i 为惩罚因子 C 以及径向基核函数中的参数 c,优化的主要目标即为搜寻到一组参数可以得到最准确的预测结果。

2　模型构建

本研究选择安徽省黄山市的月潭流域作为研究区域,该流域为新安江流域的一部分,流域面积约为 950 km^2,月潭流量监测站点位于流域出口位置(图 1)[14]。

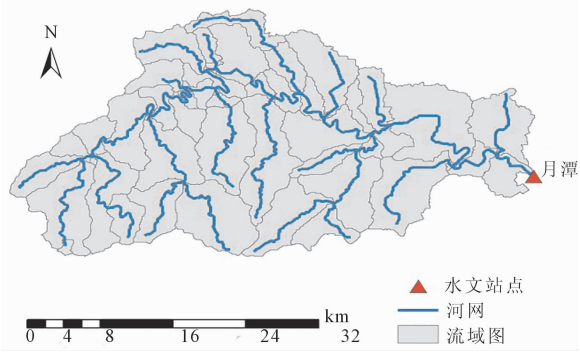

图 1　月潭流域位置示意图

该地区为典型的亚热带季风气候,多年平均气温约为 16℃,夏季最高气温可达 37℃,最低气温一般出现在 12 月、1 月或 2 月。受地形影响,降雨以暴雨居多,主要集中在 5-8 月,多年平均降雨量为 1 762 mm,蒸发量为 677.5 mm[15]。研究拟分别利用改进的 BPNN 与 SVM 进行月潭流域日径流量的预测,并依据预测结果比较模型的适用性。选取月潭水文站点 2009-2012 年的日流量监测数据作为研究对象,共包含 1 461 天的流量数据,其中模型参数校准阶段基于 2009-2010 年的数据完成,2011-2012 年的数据用于进行模型验证。

考虑到输入数据存在若干种可能的组合形式,首先利用偏自相关分析(PACF)对流量滞时进行识别。偏相关分析主要用来反映当前序列与之前某一序列的相关程度,即随机剔除了 $x(t-1)$,$x(t-2)$,$\cdots$,$x(t-k+1)$,共 $k-1$ 个随机变量的干扰之后,$x(t-k)$ 对 $x(t)$ 影响的相关程度。日径流量数据的偏自相关分析结果如图 2 所示,滞时为 1 d 和 3 d 时对 $Q(t)$ 的影响最为显著,因此选择前三天的流量数据作为输入数据,分别以 $Q(t-3)$,$Q(t-2)$,$Q(t-1)$ 表示,模型建立过程中分别基于三种不同形式的输入数据(1)$Q(t-3)$,$Q(t-2)$,$Q(t-1)$,(2)$Q(t-3)$,$Q(t-1)$,(3)$Q(t-1)$ 进行第 t 天的日径流量预测并进行结果比较。在数据前处理过程中,对流量数据进行了如式(9)所示的对数变换,Shanker 等人的研究表明经过变换的数据在进行网络训练的过程中更易达到收敛状态,训练结果也较原始数据更优。

$$q=\log_{10}(Q+\alpha) \tag{9}$$

式中:q 为变换后的流量数据,m^3/s;Q 为实际流量监测数据,m^3/s;α 为任意常数,以保证对数的底数不为 0,在本研究中 α 取值为 1。

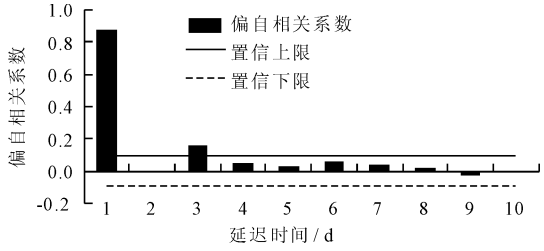

图 2　日径流量偏自相关分析图

本研究主要基于预测结果与实测数据间的相关系数(R)以及均方根误差($RMSE$)比较模型的拟合及泛化能力。其中 R 的计算过程如式(10)所示:

$$R=\frac{\sum_{i=1}^{n}(Q_{io}-\overline{Q_o})(Q_{ip}-\overline{Q_p})}{\sqrt{\sum_{i=1}^{n}(Q_{io}-\overline{Q_o})^2\sum_{i=1}^{n}(Q_{ip}-\overline{Q_p})^2}} \tag{10}$$

均方根误差 $RMSE$ 的计算公式如式(11) 所示:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_{io} - Q_{ip})^2} \quad (11)$$

式中: n 为样本总数; Q_{io} 为第 i 个样本的实际观测值; $\overline{Q_o}$ 为 n 个观测数据的平均值; Q_{ip} 为第 i 个样本的预测值; $\overline{Q_p}$ 为 n 个预测数据的平均值。当预测结果与实测数据完全吻合时应该满足 $R = 1$, $RMSE = 0$。

由于共有三种形式的输入数据,因此需分别建立输入层神经元个数分别为 1,2,3 的 BPNN,对应的输出层神经元个数都为 1,即第 t d 的流量数据。网络的初始权值利用 PSO 方法寻优确定,隐含层的神经元个数利用试错法计算得到。支持向量回归机同样拥有三种形式的输入数据,输出数据为第 t 天的流量数据,模型的惩罚因子以及核函数的参数利用 PSO 方法寻优得到。

3 结果与讨论

利用 PSO 优化算法改进后的 BPNN 以及 SVM 模型虽然准确性有所提高,但是得到的结果依然存在不稳定性,本研究利用十折交叉验证的方法,将训练数据随机分为 10 份,并轮流将其中 9 份作为训练数据,1 份作为验证数据进行模型训练,通过将 10 次的结果进行平均来评价每种模型的预测精度。基于不同输入数据的 PSO - BPNN 及 PSO - SVM 在训练阶段和验证阶段的结果如表 1 所示,可以看出二者在两个阶段的模拟精度相似。在 PSO - BPNN 中,经过试错法的分析,输入神经元个数分别为 1,2,3 时,均在隐藏神经元个数为 5 时获得最好的模拟结果,其中以 $Q(t-1)$ 作为输入数据的模型模拟结果劣于其他两种输入形式,但是基于 $Q(t-3)$, $Q(t-1)$ 和 $Q(t-3)$, $Q(t-2)$, $Q(t-1)$ 的模拟结果非常接近,无显著改善。

在 PSO - SVM 中,无论训练阶段还是验证阶段,以前三天连续流量数据作为输入数据的模拟结果均为最优,以前一天流量数据作为输入数据的模拟结果最差。

综合分析两种方法的预测结果,最大的相关系数以及最小的均方根误差均出现在以 $Q(t-3)$, $Q(t-2)$, $Q(t-1)$ 作为输入数据的 PSO - SVM 模型中,证明基于前三天连续流量数据并利用 PSO - SVM 方法预测日径流量在月潭流域效果最佳。在最佳模型中,训练阶段与验证阶段的评价统计量相差很小,证实了模型具有较强的泛化能力,利用训练阶段数据校准得到的参数可以用来预测其他不同时间段的日径流量。

表 1 BPNN 及 SVM 在训练阶段和验证阶段的结果比较

输入数据	PSO - BPNN				PSO - SVM			
	训练		验证		训练		验证	
	R	$RMSE$	R	$RMSE$	R	$RMSE$	R	$RMSE$
$Q(t-3), Q(t-2), Q(t-1)$	0.885	0.239	0.889	0.233	0.904	0.220	0.897	0.225
$Q(t-3), Q(t-1)$	0.885	0.240	0.883	0.239	0.893	0.233	0.886	0.241
$Q(t-1)$	0.876	0.248	0.893	0.230	0.872	0.253	0.889	0.240

对两种模型的整体预测能力进行评价的同时,还需要分析模型针对每一个最小时间间隔的预测效果。图 3 和图 4 分别表示 PSO - SVM 模型和 PSO - BPNN 模型在验证阶段对月潭站点每日径流量的预测效果,二者均在流量较小的时段具有较高的预测精度,而在流量达到峰值时出现低估现象。在 727 d 的预测周期内,PSO - SVM 模型的结果中有 26% 的预测误差低于 10%,7% 的预测误差高于 100%,PSO - BPNN 模型的结果中有 22% 的预测误差低于 10%,7% 的预测误差高于 100%,其中误差较大的点主要出现于达到流量峰值的日期,二者估计的偏差程度相近,说明两种方法均不能很好的解决流量峰值估算问题,需要在今后的研究中结合其他方法进一步纠正和改进。总体来说,PSO - SVM 模型的预测准确率略高于 PSO - BPNN 模型,误差分布相对平稳,更加适宜进行月潭流域日径流量的估计及预测。

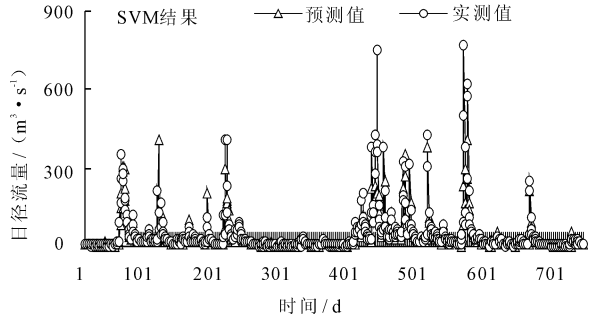

图 3 2011-01 - 2012-12 验证阶段 PSO - SVM 方法的预测值与实测值比较

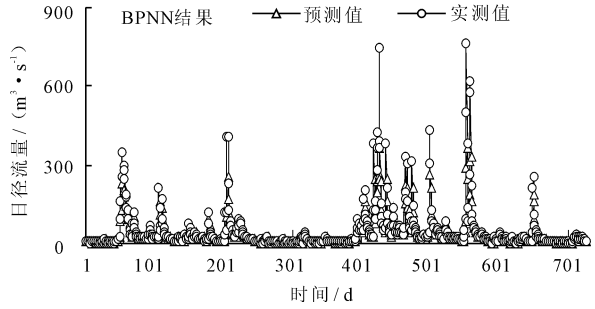

图 4　2011-01 – 2012-12 验证阶段 PSO – BPNN 方法的预测值与实测值比较

4　结　论

本研究基于月潭流域 2009 – 2012 年的日径流量监测数据分别构建了 BPNN 及 SVM 模型,为了获取尽量准确的预测结果,两个模型均利用 PSO 算法进行优化,并应用十折交叉验证提高模型结果的稳定性。分别基于训练数据(2009 – 2010)和验证数据(2011 – 2012)进行模型结果比较,发现每种方法在两个阶段的统计量都没有明显差异,表现出较好的拟合及泛化能力。其中将连续三天流量数据作为输入数据的 PSO – SVM 模型的模拟及预测精度高于其他模型,*R* 值和 *RMSE* 结果均为最优,可以认为是月潭流域日径流量估算的最优模型。进一步对模型的误差分布进行分析发现,模型对于流量峰值的预测精度较差,在流量峰值点容易出现低估现象,造成较大误差。因此虽然 PSO – SVM 模型可以用来预测小流域内的日径流量,但是在今后的研究中需要进一步优化算法,改进模型在流量峰值处的预测精度,通过准确的径流量模拟,实现流域内可能发生的水文灾害预测以及水资源的优化配置。

参考文献:

[1] 陈军锋,陈秀万. SWAT 模型的水量平衡及其在梭磨河流域的应用[J]. 北京大学学报(自然科学版),2004,40(2):265 – 270.

[2] 白晓燕,丁华龙,陈晓宏. 基于 HSPF 模型的东江流域土地利用变化对径流影响研究[J]. 灌溉排水学报,2014,33(2):58 – 63.

[3] 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(529):379 – 386.

[4] 马细霞,穆浩泽. 基于小波分析的支持向量机径流预测模型及应用[J]. 灌溉排水学报, 2008,27(3):79 – 81.

[5] Nourani V, Komasi M, Mano A. A multivariate ANN – wavelet approach for rainfall – runoff modeling[J]. Water resources management, 2009, 23(14):2877 – 2894.

[6] Taormina R, Chau K W, Sethi R. Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon[J]. Engineering Applications of Artificial Intelligence, 2012, 25(8):1670 – 1676.

[7] Wu C L, Chau K W, Li Y S. River stage prediction based on a distributed support vector regression[J]. Journal of Hydrology,2008,358(1):96 – 111.

[8] Chau K W. Particle swarm optimization training algorithm for ANNs in stage prediction of Shing Mun River[J]. Journal of Hydrology,2006, 329(3):363 – 367.

[9] Lin Shihwei, Ying Kuoching, Chen Shihchieh, et al. Particle swarm optimization for parameter determination and feature selection of support vector machines[J]. Expert Systems with Applications, 2008, 35(4):1817 – 1824.

[10] 李云良,张 奇,李 淼,等. 基于 BP 神经网络的鄱阳湖水位模拟[J]. 长江流域资源与环境,2015,24(2):233 – 240.

[11] Goh A T C. Back – propagation neural networks for modeling complex systems[J]. Artificial Intelligence in Engineering, 1995, 9(3):143 – 151.

[12] Dixon B, Candade N. Multispectral landuse classification using neural networks and support vector machines:one or the other, or both? [J]. International Journal of Remote Sensing, 2008, 29(4):1185 – 1206.

[13] Yu P S, Chen S T, Chang I F. Support vector regression for real – time flood stage forecasting[J]. Journal of Hydrology, 2006, 328(3):704 – 716.

[14] 辛朋磊,李致家,汤嘉辉,等. 新安江模型参数全局优化——以月潭流域为例[J]. 湖泊科学,2011,23(4):626 – 634.

[15] 李 雪,曹芳芳,陈先春,等. 敏感区域目标污染物空间溯源分析[J]. 中国环境科学,2013,33(9):1714 – 1720.