

数据驱动的中小河流智能洪水预报方法对比研究

马凯凯, 李士进, 王继民, 余宇峰

(河海大学计算机与信息学院, 江苏南京 210098)

摘要:近几年在洪水预报中,数据驱动洪水预报模型得到了广泛的应用,并取得了良好的效果,但是数据驱动模型大都用于大流域,很少用于小流域.中小河流大多位于资料短缺的山丘区,洪水具有突发性强,汇流时间快,预见期短的特点.为此分别构建了SVM模型、BP神经网络模型、RBF网络模型、极限学习机(ELM)模型,并利用所构建的模型对昌化流域进行预报;结果表明,SVM模型和RBF网络模型在低流量区段预测较准确,而且模型预报稳定;BP神经网络模型在高流量区段较准确,但是模型预报结果不稳定;ELM模型预报误差较大,而且预报不稳定;于是采用组合模型方式:低流量区段采用SVM模型或RBF网络模型,高流量区段采用BP神经网络模型,实验结果表明组合模型预报效果更好.

关键词:数据驱动模型;中小河流洪水预报;RBF神经网络;极限学习机

中图分类号: TP 391 **文献标识码:** A doi:10.3969/j.issn.0253-2778.2016.09.009

引用格式: 马凯凯,李士进,王继民,等.数据驱动的中小河流智能洪水预报方法对比研究[J].中国科学技术大学学报,2016,46(9):774-779.

MA Kaikai, LI Shijin, WANG Jimin, et al. Comparative study of data-driven intelligent flood forecasting methods for small- and medium-sized rivers[J]. Journal of University of Science and Technology of China, 2016,46(9):774-779.

Comparative study of data-driven intelligent flood forecasting methods for small- and medium-sized rivers

MA Kaikai, LI Shijin, WANG Jimin, YU Yufeng

(School of Computer & Information, Hohai University, Nanjing 210098, China)

Abstract: In recent years, data driven flood forecasting methods have been widely used in flood forecasting, and good results have been achieved. But most data-driven models are applied to large basins, seldom in small basins. Flash floods in small- and medium-sized rivers, which are mostly located in data-poor mountainous areas in China, are featured by abruptness, rapid concentration and short forecasting time. The support-vector-machine (SVM) model, the BP neural network model, the RBF neural network model and extreme learning machine (ELM) model respectively are established and the used to forecast flash floods in Changhua basin. The results show that the SVM model and RBF network model have accurate prediction in the low flow section with simple parameters while BP network has better performance in the high flow section with less stable forecast results for the low flow section, and that the

收稿日期:2016-03-01;修回日期:2016-09-17

基金项目:公益性行业科研专项(201501022),江苏省重点研发计划(BE2015707)资助.

作者简介:马凯凯,男,1990年生,硕士生.研究方向:模式识别、数据挖掘. E-mail: 1031798928@qq.com

通讯作者:李士进,博士/教授. E-mail: lishijin@hhu.edu.cn

ELM model is not stable with large deviations. As a result, the SVM model or RBF model was adopted for the low flow section, and BP network for the high flow section. This final combination model shows better performance in experiments.

Key words: data-driven model; small- and medium-sized rivers flood forecast; the RBF neural network; extreme learning machine

0 引言

山洪是山丘区强降雨引起的中小河流突发性、暴涨暴落的洪水,我国中小河流大多位于偏远山区,水文测站较少,缺少必要的突发性洪水应急检测方法,洪水预警预报方案不健全,近年来,山丘区极易发生极端天气事件,局部强降雨、突发性暴雨时常发生,对当地居民造成生命财产的损失^[1-3],因此基于中小河流的洪水预警预报得到国家的高度重视。

由于中小河流大多位于资料短缺的山丘区,洪水具有突发性强、汇流时间快、预见期短的特点,数据驱动洪水预报模型很少用于中小河流洪水预报,本文对比利用了 SVM 模型、BP 神经网络、RBF 神经网络、ELM 模型对浙江省昌化流域进行的洪水预报,比较并选出较适用于中小河流的智能洪水预报方法。最后采用组合模型,提高预报准确率。

随着水文数据获取能力和计算机快速计算能力的发展,基于数据驱动的洪水预报模型得到了快速发展。如何利用智能算法从历史水文数据中提取洪水特征,挖掘出水文数据中蕴含的有用信息,提高对未来洪水预报的准确率,是一个重要的研究方向^[4-7]。赵钢铁等^[8]通过互信息选择输入预报因子,建立神经网络模型预报径流量。张楠等^[9]利用最小二乘支持向量方法,构建了基于多因子量化指标的径流预测模型。闫悦新等^[10]采用简单平均、最优线性组合、最优非线性组合等三种组合方法,构建了新安江模型、垂向混合产流模型和 Tank 模型相结合的组合模型,选出最稳定的组合方法。

1 方法介绍

1.1 SVM 模型和 BP 神经网络

给定 M 个水文样本数据 $(x_i, y_i), i=1, 2, \dots, M$ 。其中 $x_i = [x_i^1, \dots, x_i^d]^T$ 为洪水预报因子列向量, y_i 为预报输出值。支持向量机模型以统计学习理论的 VC 维概念和结构风险最小化原理为基础。首先利用高斯核函数将水文数据从低维非线性空间映射到高维特征空间,这样在低维非线性的洪水数据,在

高维空间中有可能变成线性,然后训练最优输入输出之间的映射函数, $y = \omega\phi(x) + t$, 其中, $\phi(x)$ 为非线性映射函数。SVM 具有模型结构简单、泛化能力强、预测精度高、适合于小样本情况且能达到全局最优的特点,被广泛应用于非线性回归问题。具体 SVM 算法步骤请参考文献^[11-14]。

BP 神经网络是一种多层前馈神经网络,神经网络水文模型结构为:

$$Q(t) = F_{BP}[Pr(t), \dots, Pr(t - x\Delta t), Q(t - y\Delta t)] \quad (1)$$

式中, t 为当前时刻, Q 为水文站点流量, Pr 为流域降雨量, Δt 为时间间隔, x, y 为整数且大于等于 1, F_{BP} 为一种非线性映射函数。

BP 神经网络模型对洪水数据进行预报过程,就是利用历史洪水数据获得非线性函数 F_{BP} 的过程。具体 BP 算法步骤请参考文献^[15-17]。

1.2 RBF 神经网络

RBF 神经网络属于前向神经网络类型^[18],基本思想为:RBF 网络输入层到隐含层利用下文公式(2)将低维空间中的非线性洪水数据映射到高维特征空间中,输出层利用公式(3)对映射到新空间的洪水数据进行线性加权^[19-21]。

RBF 网络隐含层的激活函数采用高斯函数,对于洪水时间序列输入向量 $x \in R^d$, 公式为:

$$R_i(x) = \exp\left(-\frac{1}{2\alpha^2} \|x - c_i\|^2\right) \quad (2)$$

式中, $R_i(x)$ 为第 i 个神经元的输出, x 为 d 维洪水输入变量, c_i 为高斯核函数的中心, α 为高斯核函数的方差。

RBF 网络的输出层是对映射到新特征空间的洪水数据线性加权求和,公式为:

$$y_j = \sum_{i=1}^h \omega_{ij} \exp\left(-\frac{1}{2\alpha^2} \|x - c_i\|^2\right) \quad (3)$$

式中, ω_{ij} 为隐含层到输出层的样本权重, $i=1, 2, \dots, h$ 表示隐含层共有 h 个节点, $j=1, 2, \dots, n$ 共有 n 个输出节点, y_j 为第 j 个网络输出节点的实际输出值。

RBF 神经网络相比 BP 神经网络,具有结构简

单、训练时间短、容易收敛且局部逼近能力强等特点,受到越来越多学者的关注,被广泛应用到时间序列预测、模式识别、非线性回归问题。

1.3 极限学习机模型

极限学习机模型是一种简单易用、有效的单隐层前馈神经网络 SLFNs 学习算法^[22-23],算法步骤为:水文时间序列样本 $(x_i, y_i) (i = 1, 2, \dots, M)$, M 为样本个数, $x \in R^d$ 为水文输入特征向量, d 为向量特征维数;目标值 y_i 为实际预测流量值。极限学习机模型结构公式为:

$$\sum_{i=1}^{\tilde{M}} \beta_i g(W_i \cdot x_j + b_i) = o_j, j = 1, 2, \dots, M \quad (4)$$

式中, $g(x)$ 为激活函数, $W_i = [\omega_{i,1}, \omega_{i,2}, \dots, \omega_{i,m}]^T$ 为输入层与隐含层之间的权重, β_i 为隐含层与输出层之间的权重, b_i 为第 i 个隐层神经元的偏置, $W_i \cdot x_j$ 为样本输入权重矩阵和水文时间序列矩阵相乘, \tilde{N} 为隐含层神经元节点的个数。

极限学习机的学习目标是使得预测洪水流量与实际值之间的误差达到最小,可以表示为:

$$\sum_{j=1}^{\tilde{M}} \|o_j - y_j\| = 0 \quad (5)$$

即存在 β_i, b_i, W_i 使得

$$\sum_{i=1}^{\tilde{M}} \beta_i g(W_i \cdot x_j + b_i) = y_j, j = 1, 2, \dots, M \quad (6)$$

$$H(W_1, \dots, W_{\tilde{M}}, b_1, \dots, b_{\tilde{M}}, x_1, \dots, x_{\tilde{M}}) = \begin{bmatrix} g(W_1 \cdot x_1 + b_1) & \dots & g(W_1 \cdot x_1 + b_{\tilde{M}}) \\ \vdots & \dots & \vdots \\ g(W_1 \cdot x_M + b_1) & \dots & g(W_1 \cdot x_M + b_{\tilde{M}}) \end{bmatrix}_{M \times \tilde{M}} \quad (7)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\tilde{M}}^T \end{bmatrix}_{\tilde{M} \times m}; Y = \begin{bmatrix} T_1^T \\ \vdots \\ T_{\tilde{M}}^T \end{bmatrix}_{\tilde{M} \times m} \quad (8)$$

式中, H 为隐层神经元节点的输出, β 为输出权重, T 为期望输出。

一般 W_i, b_i 随机给定, $\beta = H^+ Y, H^+$ 为矩阵的广义逆。

极限学习机学习过程就是求矩阵 H 的过程,由于输入层权重和隐含层偏置随机给定,所以极限学习机具有学习速度快,不容易陷入局部最优的特点。

2 应用实例

本文选取昌化流域作为研究对象,昌化流域位

于浙江省分水江流域上游,流域地势西北高,东南低,属浙西山丘区,典型中小河流水系,上游设有 6 个雨量站可以提供雨量信息,分别为岛石坞、桃花村、龙门寺、双石、岭下、昱岭关。图 1 是昌化站及其周边地理位置示意图。



图 1 昌化站及其周边地理位置示意图

Fig. 1 The geographical position diagram of ChangHua and peripheral station

本文选取 1998 年-2010 年每年汛期场次洪水数据,数据时间间隔为 1 小时,其中 1998 年-2008 年共 6 790 个样本作为训练样本,2009 年-2010 年共 1 173 个样本作为测试样本。

2.1 预报因子的选择

采用智能化洪水预报方法对洪水进行回归预测时,预报因子的选择显得尤为重要。选取洪水预报因子要遵循两大原则:①能显著影响洪水发生的影响因子,如流域降雨量、上游流量等。②在发生洪水前容易得到这些影响因子的具体数值^[5]。分析昌化流域特点,发现影响昌化流量的因素主要是昌化上游 6 个雨量站以及自身测站的信息。由于中小河流一般流域较短、暴雨集中、汇流时间快,选择预见期为 4 h。拟选取预报因子为:岛石坞 4h-7h、桃花村 4h-7h、龙门寺 4h-7h、双石 4h-7h、岭下 4h-7h、昱岭关 4h-7h、昌化 4h-7h、昌化流量 4h-5h(其中岛石坞 4h-7h 表示预测时间点前 4h-7h 的雨量信息),总共 30 维数据作为输入,模型输出为昌化现在时刻的流量。

2.2 洪水预报方案

利用上面整理好的数据分别验证四种模型的有效性,将 SVM 模型、BP 神经网络、RBF 网络模型、ELM 模型用于昌化流域洪水预报,主要有以下步骤:

(I)数据的归一化。利用下式对训练样本和测试样本进行归一化处理,经过归一化后的数据位于 0~1 之间。

$$\hat{x} = (x - x_{\min}) / (x_{\max} - x_{\min}) \quad (9)$$

式中, \hat{x} 为归一化后的数据, x 为原始数据, x_{\max} , x_{\min} 为样本中每一维的最大值和最小值。

(II)模型率定. 使用 Matlab 软件环境实现上面四种模型, 利用归一化后的数据作为训练样本分别训练四种模型, 根据模型的特点, 通过调整参数达到模型最优, 使模型尽可能更好的刻画数据之间的映射关系。

(III)模型验证. 模型率定后, 利用模型对测试样本进行预测, 并对结果进行反归一化, 并利用均方误差 MSE 和 DC 系数对预测结果进行评价。

(IV)为提高预报准确率, 按照昌化流量大小对原始样本进行划分, 划分为高流量区段和低流量区段. 选取昌化站警戒流量 $300\text{m}^3/\text{s}$ 为分界点, 利用 SVM 模型和 RBF 网络模型对低流量区段进行预测, 利用 BP 神经网络对高流量区段进行预测, 最后将其合并, 与单模型对比。

2.3 模型评价指标

根据《水文情报预报规范》, 洪水预报误差可以采用绝对误差、相对误差、均方误差、确定性系数等指标进行评定, 本文采用均方误差和确定性系数对预测结果进行评价。

(I)确定性系数反应了洪水预报过程与实测过程之间的吻合程度, 它的取值范围为 $[0, 1]$, 结果越接近 1.0, 预报准确率越高, 其计算公式为:

$$DC = 1 - \frac{\sum_{i=1}^n [y_c(i) - y_0(i)]^2}{\sum_{i=1}^n [y_0(i) - \bar{y}_0]^2} \quad (10)$$

式中, MSE 为均方误差, DC 为确定性系数(取 2 位小数), $y_0(i)$ 为实测值, $y_c(i)$ 为预测值, \bar{y}_0 为实测值均值, n 为样本个数。

(II)均方误差反映了预测值与真实值之间的偏差程度, 值越小越优, 其计算公式为:

$$MSE = \frac{1}{n} \sum_{i=1}^n [y_c(i) - y_0(i)]^2 \quad (11)$$

2.4 模型结果对比及分析

通过上述预报方案利用四种模型对昌化站进行预报, 结果如下表 1 所示. 由表 1 知, SVM 模型整体误差最小, 均方误差为 3 718, 其他 3 个模型误差从小到大依次为 BP 神经网络、RBF 神经网络、极限学习机模型. 由于 SVM 模型和 RBF 网络模型参数调整较简单, 而且参数相同的情况下, 每次预测结果不

变. 由于 BP 神经网络和 ELM 模型由于存在随机初始值和偏置的问题, 因此在固定隐含层节点数的情况下, 每次运行结果都变化, 对洪水预报不能得到稳定的输出, 故需要多次运行取平均值。

表 1 四种不同模型预测结果比较

Tab. 1 The Comparison of different prediction results by four models

模型	均方误差	DC 系数
SVM 模型	3 718	0.843
BP 神经网络模型	3 952	0.833
RBF 模型	4 292	0.819
极限学习机模型	4 786	0.798

利用均方误差和确定性系数只能评估模型整体预报结果, 而不能从细节对其进行分析, 而中小河流一般山高坡陡、源短流急, 所以洪水具有陡涨陡落的特点, 为了更好地分析四种模型在昌化流域的预测情况, 将昌化站 2009-2010 年的预测结果画出折线图, 如图 2-5 所示。

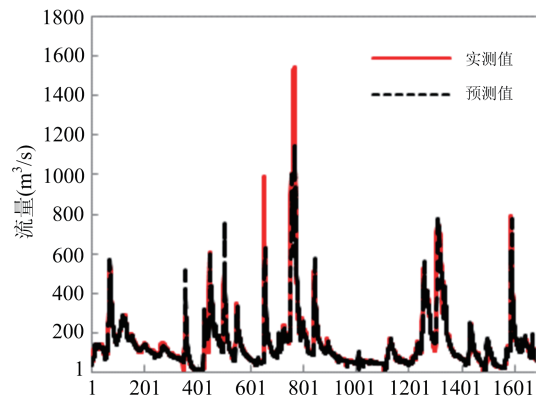


图 2 SVM 模型预测结果

Fig. 2 The prediction of SVM model

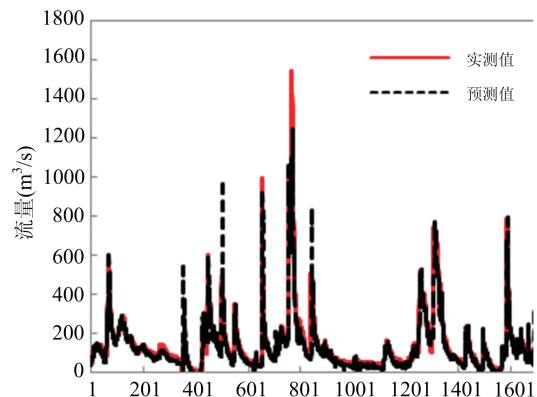


图 3 RBF 网络模型预测结果

Fig. 3 The prediction of RBF network model

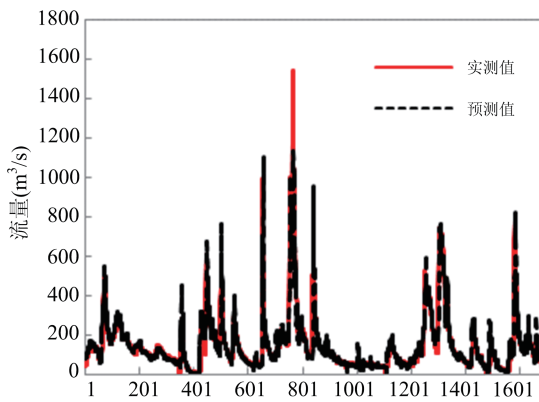


图 4 BP 网络预测结果

Fig. 4 The prediction of BP network model

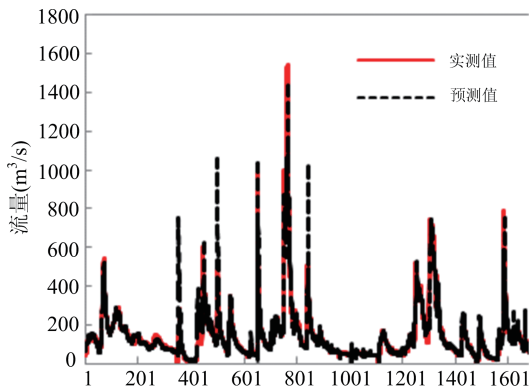


图 5 ELM 模型预测结果

Fig. 5 The prediction of ELM model

通过图 2-5 可以发现:①对于 2010 年昌化流量数据(横坐标 1000 之后的数据),由于全年流量最大达到 $800\text{m}^3/\text{s}$,没有出现特大洪峰的情况,在训练样本中存在较多这样的数据,模型训练的较充分,因此 4 种模型预报结果相差不大,预报值与真实值较为吻合。②对于 2009 年昌化流量数据(横坐标 1000 之前的数据),全年最大洪峰达到 $1500\text{m}^3/\text{s}$,另外一场洪水洪峰达到 $1000\text{m}^3/\text{s}$,对于这两次特大洪峰的预测,4 种模型预报结果与真实值都存在一定的偏差,这可能由于训练样本中存在较少这种样本,模型训练的不充分,导致预测偏差较大;但是从细节分析容易发现:极限学习机在这两次特大洪峰时刻预报结果误差最小,对于其他 3 处小的洪峰,都存在预报偏大情况;BP 网络在这两次特大洪峰时刻预报结果较好,对于其他 3 处小的洪峰,存在两处预报偏大情况;支持向量机在这两次洪峰时刻预报误差最大,但是在其他小的洪峰时刻预报较准确;RBF 神经网络在这两次洪峰时刻存在偏差,但是对于其他 3 处小的洪峰,存在一处预报偏大情况。这也验证了 BP 神

经网络和 ELM 模型在预报上的不稳定性;BP 可以在高流量时段获得较好的预报结果,而 SVM 和 RBF 在低流量时段可以获得较好的预报结果。

根据预报方案中第(IV)步,将昌化站按照警戒流量对洪水数据进行划分,低流量区段采用 SVM 模型或 RBF 网络模型预测,高流量区段采用 BP 神经网络模型预测,最后将两者组合结果如表 2 所示。

表 2 根据流量大小对模型组合

Tab. 2 The model combination according to the size of discharge

模型	均方误差	DC 系数
SVM+BP 组合模型	3 018	0.903
RBF+BP 组合模型	3 252	0.893

通过表 2 可以发现,组合模型预报效果好于单个模型。这可能由于洪水数据具有高度复杂性、非线性特点,洪水数据中间蕴含了多个模型,高流量区段与低流量区段数据之间的映射关系不同,同一个模型很难兼顾两者,将具有不同映射关系的数据分开后分别训练模型,提高了预报准确率。

3 结论

近年来,我国中小河流洪水灾害十分严重,中小河流洪水预报是中小河流防洪减灾的重要的非工程措施之一,已引起各级政府和防汛部门的高度重视和社会的广泛关注。本文用 SVM 模型、BP 神经网络、RBF 神经网络、极限学习机模型对中小河流昌化流域进行了预测并作比较。再结合模型的原理、流域特点以及预测结果分析了不同模型对中小河流流域预报的适用性,得到如下结论:

(I)数据驱动模型作为洪水预报的一种重要方法得到快速发展,本文将其用于中小河流预测,获得了较好的预报结果,而且将预见期设置为 4 小时,具有实际应用价值,可以将其用于中小河流实时预报系统中。

(II)SVM 模型和 RBF 模型在中小河流中参数调整较简单,而且易得到稳定的结果,在低流量区段预报较准确,BP 网络预报稳定性一般,但是高流量区段预报较准确,将两者结合,提高了预报准确率。

(III)中小河流流域往往资料短缺,可靠性得不到保证,因此在用这些智能化洪水预报模型时,利用数据挖掘中数据预处理方法对数据进行消噪处理可能会提高预报准确率。

参考文献(References)

- [1] 刘志雨, 侯爱中, 王秀庆. 基于分布式水文模型的中小河流洪水预报技术[J]. 水文, 2015, 35(1): 1-6.
- [2] 刘志雨, 杨大文, 胡健伟. 基于动态临界雨量的中小河流山洪预警方法及其应用[J]. 北京师范大学学报(自然科学版), 2010, 46(3): 317-321.
- [3] 叶金印, 吴勇拓, 李致家, 等. 湿润地区中小河流山洪预报方法研究与应用[J]. 河海大学学报(自然科学版), 2012, 40(6): 615-621.
- [4] 吴恒卿, 黄强, 习树峰. 基于熵权的可变模糊聚类与识别的水库洪水分类实时预报[J]. 水力发电学报, 2015, 34(2): 57-63.
- [5] 刘可新, 包为民, 阙家骏, 等. 基于主成分分析的K均值聚类法在洪水预报中的应用[J]. 武汉大学学报(工学版), 2015, 48(4): 447-450.
- [6] 李士进, 朱跃龙, 张晓花, 等. 基于BORDA计数法的多元水文时间序列相似性分析[J]. 水利学报, 2009, 40(3): 378-384.
- [7] 朱跃龙, 李士进, 范青松, 等. 基于小波神经网络的水文时间序列预测[J]. 山东大学学报(工学版), 2011, 41(4): 119-124.
- [8] 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]. 水利发电学报, 2011, 30(1): 24-30.
- [9] 张楠, 夏自强, 江红. 基于多因子量化指标的支持向量机径流预测[J]. 水利学报, 2010, 41(11): 1318-1324.
- [10] 闫月新, 包为民. 组合预报方法在洪水预报模型中的应用[J]. 水电能源科学, 2013, 31(10): 47-49.
- [11] 朱林, 蔡田. 最小二乘支持向量机建模及应用[J]. 工业控制计算机, 2013, 26(9): 60-61.
- [12] 王娜. 基于最小二乘支持向量机的北京市肉类需求量预测研究[D]. 北京交通大学, 2013.
- [13] TONG S, KOLLER D. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research, 2002, 2(1): 45-66.
- [14] 孙健, 王成华, 洪峰, 等. 基于人工鱼群优化支持向量机的模拟电路故障诊断[J]. 系统仿真学报, 2014, 26(4): 843-847.
- [15] 熊立华, 郭生练, 庞博, 等. 三种基于神经网络的洪水实时预报方案的比较研究[J]. 水文, 2003, 23(5): 1-4.
- [16] 李松, 刘力军, 解永乐. 遗传算法优化BP神经网络的短时交通流混沌预测[J]. 控制与决策, 2011, 26(10): 1581-1585.
- [17] 杨淑娥, 黄礼. 基于BP神经网络的上市公司财务预警模型[J]. 系统工程理论与实践, 2005, 25(1): 12-18.
- [18] 夏长亮, 祁温雅, 杨荣, 等. 基于RBF神经网络的超声波电机参数辨识与模型参考自适应控制[J]. 中国电机工程学报, 2004, 24(7): 117-121.
- [19] 张晓瑞, 方创琳, 王振波, 等. 基于RBF神经网络的城市建成区面积预测研究: 兼与BP神经网络和线性回归对比分析[J]. 长江流域资源与环境, 2013, 22(6): 691-697.
- [20] 党江杰. 基于模糊控制的隧道通风节能控制模型研究[D]. 长安大学, 2013.
- [21] 魏光辉. 基于RBF神经网络的河川年径流量预测[J]. 西北水电, 2014, (5): 6-9.
- [22] 孙俊, 卫爱国, 毛罕平, 等. 基于高光谱图像及ELM的生菜叶片氮素水平定性分析[J]. 农业机械学报, 2014, 45(7): 272-276.
- [23] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: theory and applications [J]. Neurocomputing, 2006, 70(1-3): 489-501.