DOI:10.11705/j.issn.1672-643X.2014.06.049

遗传算法与最小二乘支持向量机在年径流预测中的应用

代兴兰
(云南省水文水资源局曲靖分局, 云南 曲靖 655000)

摘 要: 为克服最小二乘支持向量机(LSSVM)依赖人为经验选择学习参数的不足,利用遗传优化算法(GA)选择LSSVM惩罚因子 C 和核函数参数 σ^2,构建GA-LSSVM年径流预测模型,并构建LSSVM、GA-BP和传统BP模型作为对比,以云南省河边水文站年径流预测为例进行实例研究,利用实例前30 a和后22 a资料分别对各模型进行训练和预测。结果表明:GA-LSSVM模型对实例后22 a年径流预测的平均相对误差绝对值和最大相对误差绝对值分别为3.13%、8.66%,预测精度优于LSSVM、GA-BP和传统BP模型。GA算法全局寻优能力强,利用GA算法优化得到的LSSVM学习参数可有效提高LSSVM模型的预测精度和泛化能力。

关键词: 径流预测; 遗传算法; 最小二乘支持向量机; BP神经网络
中图分类号:P333.1 **文献标识码**: A **文章编号**: 1672-643X(2014)06-0231-05

Application of genetic algorithm and least squares support vector machine in prediction of annual runoff

DAI Xinglan
(*Qujing Branch Bureau of Yunnan Hydrological and Water Resources Bureau*, *Qujing* 655000, *China*)

Abstract: In order to overcome the lack of learning parameters of least square support vector machine (LSSVM) which is chosen by human experience, the paper used the genetic algorithm (GA) to select LSSVM penalty factor c and kernel parameter and construct GA-LSSVM annual runoff forecasting model, and set up LSSVM, GA-BP and traditional BP model as the comparison. Taking Yunnan province river hydrological station annual runoff prediction as a case study, it used the data before 30 years and after 22 years to train and predict every model. The results show that the absolute value of the average relative error of annual runoff and the maximum absolute value of relative error predicted by GA-LSSVM model after 22 years are 3.13%, 8.66%, the prediction accuracy of GA-LSSVM model is better than that of LSSVM, GA-BP and traditional BP model. The global optimization ability of GA algorithm is strong. The LSSVM learning parameters gotten by optimization of GA algorithm can effectively improve the prediction accuracy of LSSVM model and the generalization ability.

Key words: runoff forecast; genetic algorithm; least squares support vector machine; BP neural network

1 研究背景

提高径流预测精度对于水文预测预报具有重要意义,目前时间序列法[1]、回归分析法[2]、小波分析法[3]、集对分析法[4]、灰色预测法[5]以及组合预测法[6]等广泛应用于径流预测,取得了较好的预测效果。人工神经网络(artifical neural network, ANN)是一种非线性建模工具,常用来对输入和输出间复杂的关系进行建模,能以任意精度逼近任何非线性连续函数,在水文预测预报中具有广泛应用[7-9]。但由于传统ANN算法是基于渐近理论,仅在样本容量趋向于无穷大时其经验风险才趋近于实际风险,在实际应用中存在泛化能力差、收敛速度慢、算法容易陷入局部最优等问题[10-11]。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前馈神经网络,用于解决模式分类和非线性映射等问题。SVM通过统计学习中的VC维(vapnik-chervonenkis di-

收稿日期:2014-06-12; **修回日期**:2014-07-23
作者简介:代兴兰(1971-),女,云南会泽人,本科,高级工程师,主要从事水资源研究、水资源保护及水文情报预报工作。

mension)理论和寻求结构风险最小化原理来提高泛化能力,已成为继 ANN 之后机器学习领域新的研究热点,其主要特点在于:①SVM 求解的是一个二次寻优问题,理论上得到全局最优;②SVM 拓扑结构由支持向量决定,弥补了 ANN 结构难以确定的不足;③SVM 决策函数由少数的支持向量确定,计算的复杂程度取决于支持向量的数目,避免了"维数灾"问题;④SVM 方法求解的是基于结构风险最小化原则,由经验风险和置信区间共同决定支持向量的实际风险,泛化能力优于传统 ANN。尤其在解决小样本容量时,有效解决了传统 ANN 在实际应用中存在的模型选择、过学习、高维和局部极值等问题[12-14]。但传统 SVM 不足在于使用二次规划求解问题,当训练样本数据量大和高维时,SVM 模型存在收敛速度慢、耗时长等不足。最小二乘支持向量机(least squares support vector machines,LSSVM)是 Suykens 于 1999 年提出的一种改进型 SVM,与传统 SVM 相比,LSSVM 引入最小二乘线性系统到 SVM 中,用等式约束替代不等式约束,避免了求解耗时的二次规划问题[15-16],提高了求解速度和收敛精度,而且 LSSVM 无需指定逼近精度,增加了 SVM 模型的实用性,在水文预测预报中应用广泛[14,17-18]。然而,LSSVM 模型预测效果受惩罚因子 C 和核函数参数 σ^2 的影响较大,目前 LSSVM 模型惩罚因子 C 和核函数参数 σ^2 的选取方法有经验选择法、实验试凑法和交叉验证法等,极大地制约了 LSSVM 模型预测精度与泛化能力的提高。遗传算法(genetic algorithm,GA)是近年来迅速发展起来的一种全新的随机搜索与优化算法,主要借鉴生物界自然选择和自然遗传机制进行搜索,用于 LSSVM 模型参数优选,可有效提高 LSSVM 模型的预测精度和泛化能力。

针对上述问题及不足,本文利用遗传算法(genetic algorithm, GA)优化选择 LSSVM 惩罚因子 C 和核函数参数 σ^2,构建 GA-LSSVM 年径流预测模型,并构建 LSSVM、GA-BP 和传统 BP 模型作为对比,以云南省河边水文站年径流预测为例进行实例分析。验证结果表明 GA-LSSVM 模型具有较好的预测精度和泛化能力。

2 预测模型

2.1 LSSVM 理论与算法简介

SVM 的基本思想是通过一个非线性映射,将输入空间的数据映射到一个高维特征空间,并在高维空间中求解最优回归函数,将实际问题转化为一个带不等式约束的二次规划问题。LSSVM 是基于 SVM 的改进算法,用等式约束替代不等式约束,把问题转化为一个线性矩阵求解问题,避免了求解耗时的二次规划,提高了求解速度和收敛精度,且 LSSVM 不再需要指定逼近精度 ,易于理解和操作。

LSSVM 建模原理如下[15,19-20]:

(1) 给定集合 $\{(x_i,y_i),\ i=1,2,\cdots,l\}$,其中,$x_i(x_i\in R^d)$ 为第 i 个训练样本的输入向量,$x_i=[x_i^1,x_i^2,\cdots,x_i^d]^{\mathrm{T}}$,$y_i\in R$ 为对应输出值,在高维特征中建立的线性回归函数为:

$$f(x)=w^{\mathrm{T}}\Phi(x)+b \tag{1}$$

式中:$\Phi(x)$ 为非线性映射函数;w 为权值向量;b 为偏置。

(2)利用 LSSVM 进行函数回归时优化目标为:

$$\begin{cases}\min J(w,e)=\dfrac{1}{2}w^{\mathrm{T}}w+\dfrac{1}{2}C\sum\limits_{i=1}^{l}e_i^2\\ \text{s.t. } y_i=w\Phi(x)+b+e_i,(i=1,2,\cdots,l)\end{cases} \tag{2}$$

式中:C 为惩罚因子;e_i 为误差变量。

(3)式(2)引入 Lagrange 乘子 α_i,得:

$$L(w,e,\alpha,b)=J(w,e)-\sum_{i=1}^{l}\alpha_i\{w^{\mathrm{T}}\Phi(x_i)+b+e_i-y_i\} \tag{3}$$

(4)根据 KKT(Karush-Kuhn-Tucker)条件可得:

$$\begin{cases}\dfrac{\partial L}{\partial w}=0\rightarrow w=\sum\limits_{i=1}^{l}\alpha_i\Phi(x_i)\\ \dfrac{\partial L}{\partial b}=0\rightarrow\sum\limits_{i=1}^{l}\alpha_i=0\\ \dfrac{\partial L}{\partial e_i}=0\rightarrow\alpha_i=Ce_i\\ \dfrac{\partial L}{\partial\alpha_i}=0\rightarrow w^{\mathrm{T}}\Phi(x_i)+b+e_i-y_i=0\end{cases} \tag{4}$$

(5)消去变量 w、e_i,得线性方程组:

$$\begin{bmatrix}b\\ \alpha\end{bmatrix}=\begin{bmatrix}0 & 1_v^{\mathrm{T}}\\ 1_v & \Omega+C^{-1}I\end{bmatrix}^{-1}\begin{bmatrix}0\\ y\end{bmatrix} \tag{5}$$

式中:$y=[y_1,y_2,\cdots y_l]^{\mathrm{T}}$;$1_v=[1,1,\cdots,1]^{\mathrm{T}}$;$\alpha=[\alpha_1,\alpha_2,\cdots\alpha_l]^{\mathrm{T}}$;$I$ 为 1 阶单位矩阵;Ω 为 $l\times l$ 非负定矩阵;$\Omega_{im}=\Phi(x_i)^{\mathrm{T}}\Phi(x_m),i,m=1,2,\cdots,l$;根据 Mercer 条件,存在映射函数 Φ 和核函数 $K(\cdot,\cdot)$,使得:

$$K(x_i,x_l)=\Phi(x_i)^{\mathrm{T}}\Phi(x_l) \tag{6}$$

式中:对于径向基核函数,则有:

$$K(x_i,x_l)=\exp\left\{-\frac{\|x-x_i\|}{2\sigma^2}\right\} \tag{7}$$

式中:σ 是 RBF 核函数宽度。

(6)利用最小二乘法求解回归系数 α 和偏置 b,

从而得到 LSSVM 预测函数:

$$y(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i) + b \tag{8}$$

根据上述 LSSVM 算法原理,在给定样本和 RBF 核函数条件下,LSSVM 的性能主要受惩罚因子 C 和核函数参数 σ^2 的影响。惩罚因子 C 用于控制函数的拟合误差,C 值越大,拟合误差越小,训练时间增加,但过大的 C 值会导致模型过拟合;核函数参数 σ^2 代表 RBF 带宽,σ^2 值越小,拟合误差越小,训练时间延长,但过小的 σ^2 值同样会导致模型过拟合。

2.2　GA – LSSVM 预测模型

遗传算法(GA)是由美国 Michigan 大学教授 JohnHolland 于 20 世纪 60 年代模拟生物在自然环境中的遗传和进化过程而提出的自适应全局优化概率搜索算法。GA 算法的基本思想是将问题的求解表示成"染色体",并置于问题的"环境"中,根据适者生存原则,从中选择出适应环境的"染色体"进行复制,并通过交叉(crossover)、变异(mutation)两种基因操作产生出新一代更适合环境的"染色体"群,这样不断重复,最后收敛到一个最适合环境的个体上,从而求得问题的最优解或次优解。由于在给定样本和 RBF 核函数条件下,LSSVM 的性能受惩罚因子 C 和核函数参数 σ^2 的影响较大,因此,本文利用 GA 算法优良的全局寻优能力,选择 LSSVM 模型中的惩罚因子 C 和核函数参数 σ^2。

定义适应度(目标)函数:

$$\begin{cases} \min \ f(C,\sigma^2) = \dfrac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y})^2 \\ \text{s.t} \quad C \in [C_{\min}, C_{\max}],\ \sigma^2 \in [\sigma^2_{\min}, \sigma^2_{\max}] \end{cases} \tag{9}$$

式中:y_i 为第 i 个样本实测值;$\hat{y}_i$ 为第 i 个样本模拟值,可由(8)计算得到。LSSVM 模型参数优化的思想就是通过迭代算法搜寻一组参数 (C,σ^2),通过 LSSVM 学习使适应度函数(9)达到最小。

本文选取公式(9)作为适应度函数,利用 GA 算法对参数 (C,σ^2) 进行寻优。具体算法步骤描述如下[17-18]:

(1)选取 LSSVM 模型的训练样本和检验样本,设定惩罚因子 C 和核函数参数 σ^2 的搜寻范围,对 LSSVM 惩罚因子 C 和核函数参数 σ^2 进行染色体基因编码,产生 LSSVM 学习参数初始种群。

(2)确定式(9)为优化目标适应度函数。

(3)设定 GA 算法的变异概率、交叉概率以及种群规模、进化迭代次数等。

(4)进行 LSSVM 训练,按式(9)计算优化目标适应度函数值,若满足优化目标训练停止准则(达到训练误差或迭代次数),则转至步骤(6),否则转至步骤(5)。

(5)在不满足优化目标训练停止准则条件下,依据 GA 算法设定的交叉及变异概率进行交叉及变异操作,重新计算优化目标适应度函数值,判断是否满足优化目标训练停止准则。

(6)结束训练,获得最佳惩罚因子 C 和核函数参数 σ^2 向量,将学习参数向量赋予 LSSVM 模型,利用 LSSVM 模型对检验样本进行预测。

3　实例应用

3.1　数据来源与分析

黄泥河河边水文站设立于 1958 年 8 月,位于云南省曲靖市富源县营上镇河边村,属珠江流域西江水系,南盘江一级支流。水文站控制径流面积 1 536 km^2,观测项目有水位、流量、降雨和蒸发,是国家基本水文站、省级报汛站和国家水质监测站。本文以云南省河边水文站 1959 – 2010 年 52 a 的实测资料为例进行分析。利用 SPSS 软件分析年径流与 1 – 10 月月均流量的相关性,分析结果见下表 1。

表 1　年径流与 1 – 10 月月均流量的相关系数

月　份	1	2	3	4	5	6	7	8	9	10
年径流	– 0.001	0.101	0.144	– 0.010	0.152	0.469**	0.589**	0.669**	0.682**	0.725**

注:"**"表示在 0.01 水平(双侧)上显著相关。

从表 1 可以看出,年径流与 1、4 月月均流量呈弱负相关,与 2、3 月,5 – 10 月月均流量呈正相关关系,其中与 6 – 10 月月均流量均呈显著正相关,相关系数在 – 0.001 ~ 0.725 之间,相性并不显著。

本文为充分挖掘输入变量间的有用信息,并兼顾输入变量的空间维度,选取河边站 2、3 月及 5 – 10 月月均流量作为年径流预测的影响因子,并以 1959 – 1988 年 30 年的实测资料作为训练样本,1989 – 2010 年 22 年的实测资料作为检验样本。

3.2　应用

将 GA – LSSVM、LSSVM 及 GA – BP、BP 模型应用于河边站年径流预测的步骤分为:样本的选取及

预处理、模型的构建、模型训练及预测、预测结果分析4个步骤。

(1)样本的选取及预处理。选取龙潭站1959－1988年2、3月及5－10月月均流量作为年径流量预测的影响因子,其中以1959－1988年实测资料作为训练样本,1989－2010年作为检验样本,并利用式(10)对各径流序列进行归一化处理:

$$\hat{x} = (x - x_{\min})/(x_{\max} - x_{\min}) \tag{10}$$

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。

(2)模型的构建。本文基于MatlabR2011b软件环境和LSSVMlab 1.5工具箱、谢菲尔德(Sheffield)遗传算法工具箱编程构建各预测模型,以2、3月及5－10月月均流量作为模型输入向量,年径流作为输出向量,创建8输入1输出的GA－LSSVM、LSSVM及GA－BP、BP年径流预测模型。

(3)模型的训练及预测。利用所选取的训练样本和预测样本对GA－LSSVM、LSSVM及GA－BP、BP模型进行训练及预测,并选取平均相对误差绝对值 MRE 和最大相对误差绝对值 MaxRE 和对各模型的预测效果进行评价,公式如下:

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i} \times 100 \tag{11}$$

$$\text{Max}RE = \max_{1\leqslant i\leqslant n}\frac{|\hat{y}_i - y_i|}{y_i} \times 100 \tag{12}$$

式中:$\hat{y}_i$ 为第 i 个样本预测值;y_i 为第 i 个样本实测值;$i = 1,2,\cdots,n$,n 为预测样本数。

本文模型中的参数设置如下:

GA－LSSVM模型:基于谢菲尔德(Sheffield)遗传算法工具箱和LSSVMlab 1.5工具箱,选择径向基核函数作为核函数。惩罚因子 C 和核函数参数 σ^2 的搜索空间均设为0.01～1000;最大迭代次数设为1000,期望误差 10^{-5};种群规模设为30;交叉概率、变异概率分别设置为0.90、0.10(其他参数采用工具箱默认值)。$C = 747.4197$;$\sigma^2 = 785.4799$。

LSSVM模型:基于LSSVMlab 1.5工具箱,选择径向基核函数作为核函数;采用实验试凑法确定LSSVM模型最佳惩罚因子 C 和核函数参数 σ^2 分别为200、500。

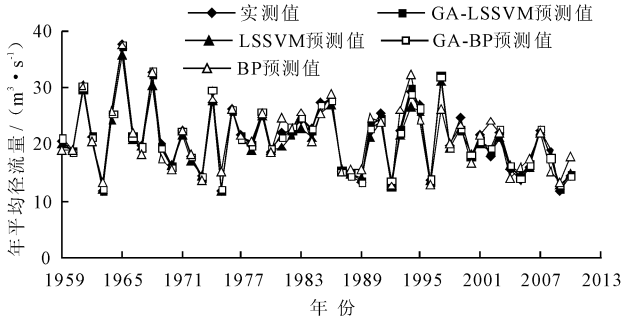
GA－BP模型:依据Kolmogorv定理得出一个初始神经元数,然后利用逐步增长或逐步修剪法确定GA－BP模型中BP部分的最佳神经元数,经高调试,最终确定GA－BP模型中BP模型结构为8－15－1,隐含层和输出层传递函数分别采用logsig和purelin,训练函数采用traingdx,学习速率lr为0.05,期望误差为0.0001,最大训练轮回为500次。在此条件下运用GA来优化BP神经网络的初始权值和阈值,GA参数设置为:种群规模为100,进化次数为200次,其他参数采用工具箱默认值。

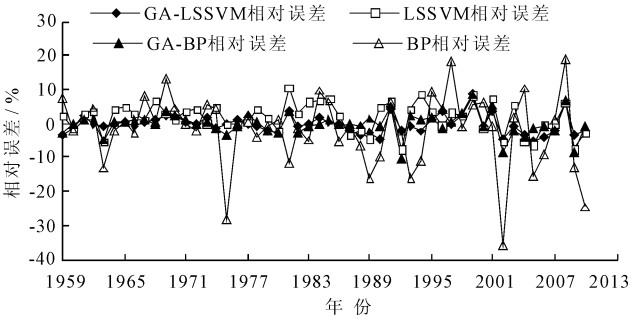
传统BP模型:依据Kolmogorv定理,然后利用逐步增长或逐步修剪法确定最佳神经元数。最终确定该实例BP模型结构为8－12－1,隐含层和输出层传递函数分别采用logsig和purelin,训练函数采用traingdx,学习速率lr为0.05,设定期望误差为0.01,最大训练轮回为1 000次。

利用训练好的GA－LSSVM、LSSVM及GA－BP、BP模型对河边站年径流进行拟合及预测,见表2及图1、图2。

表2 河边站年径流拟合、预测结果及其比较

样本	误差/%	GA－LSSVM模型	LSSVM模型	GA－BP模型	BP模型
训练样本	MRE	1.31	3.42	1.50	5.06
	maxRE	3.54	10.84	4.91	28.11
预测样本	MRE	3.13	4.67	3.52	10.73
	maxRE	8.66	8.91	10.35	35.82





河边站1959－2010年(前30年为训练样本,后22年为检验样本)

图1 河边站年径流拟合及预测值比较

图2 河边站年径流拟合及预测相对误差

(4)预测结果分析。从表 2 及图 1 ~ 2 可以看出:①GA - LSSVM 模型对河边站后 22 a 年径流预测的 *MRE* 、Max*RE* 分别为 3.13%、8.66%,预测精度和泛化能力均优于 LSSVM 和 GA - BP,优于 BP 模型,表明 GA 算法具有较好的全局寻优能力,利用 GA 算法优化得到的 LSSVM 惩罚因子 C 和核函数参数 σ^2 可有效提高 LSSVM 模型的预测精度和泛化能力。②从图 1 ~ 2 来看,GA - LSSVM、LSSVM 和 GA - BP 模型对河边站长达 52 a 的年径流拟合及预测中,其相对误差在 -10.35% ~ 10.84% 之间变化,表明 GA - LSSVM、LSSVM 和 GA - BP 模型均具有较好的拟合及预测效果,其中 GA - LSSVM 模型的拟合及预测相对误差在 -4.97% ~ 8.66% 之间变化,具有更高预测精度。③从 GA - BP 及 BP 模型的预测结果来看,由于 BP 算法是基于渐近理论,仅在样本容量趋向于无穷大时其经验风险才趋近于实际风险,在实际应用中存在泛化能力差、易陷入局部最优等问题,采用 GA 算法能有效优化 BP 网络初始权值和阈值,提高 BP 模型的预测精度和泛能能力。

4　结　语

(1)针对 LSSVM 模型性能受惩罚因子 C 和核函数参数 σ^2 的影响较大以及采用试凑法等方法确定惩罚因子 C 和核函数参数 σ^2 的不足,利用 GA 算法强大的寻优能力与 LSSVM 融合,实现了 LSSVM 模型惩罚因子 C 和核函数参数 σ^2 的自动确定,提高了 LSSVM 模型的预测精度及泛化能力。

(2)从应用实例来看,GA - LSSVM 预测模型各项性能评价指标大幅优于 BP 模型,优于 LSSVM 和 GA - BP 预测模型,表明 GA 优化算法具有较强的全局寻优能力。利用该算法优化 LSSVM 学习参数得到的模型具有预测精度高、泛化能力强以及收敛速度快等优点。

参考文献:

[1] 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.

[2]李 琪,马建斌,刘洪吉. 基于逐步回归的投影寻踪水文预报模型研究[J]. 水电能源科学,2011,29(2):10 - 12.

[3]汪丽娜,陈晓宏,李 艳. 不同径流尺度的小波神经网络预测[J]. 华南师范大学学报(自然科学版),2013,45(2):104 - 106.

[4]王延亭,王建群,张玉杰. 基于加权秩次集对分析法的年径流预报模型[J]. 水电能源科学,2012,30(3):17 - 19 + 67.

[5] 张 霞,陈亚宁,黄海龙,等. 河川径流的灰色拓扑预测研究——以新疆阿克苏河为例[J]. 第四纪研究,2010,30(1):216 - 223.

[6]崔东文. 多重组合神经网络模型在年径流预测中的应用[J]. 水利水电科技进展, 2014,34(2): 59 - 63.

[7]崔东文. 多隐层 BP 神经网络模型在径流预测中的应用[J]. 水文,2013,33(1):68 - 73.

[8]崔东文. 改进 Elman 神经网络在径流预测中的应用[J]. 水利水运工程学报,2013(2):71 - 77.

[9]邓 霞,董晓华,薄会娟. 基于 BP 网络的河道径流预报方法与应用[J]. 人民长江,2010,41(2):56 - 59.

[10] 田雨波. 混合神经网络技术[M]. 北京:科学出版社,2009.

[11] 王 雷. 支持向量机在汽轮机状态监测中的应用[M]. 北京:北京师范大学出版社,2012.

[12]崔东文. 支持向量机在湖库营养状态识别中的应用[J]. 水资源保护,2013,29(4):26 - 30.

[13]崔东文. 支持向量机在水资源类综合评价中的应用——以全国 31 个省级行政区水资源合理性配置为例[J]. 水资源保护,2013,29(5):20 - 27.

[14]李彦彬,尤 凤,黄 强,等. 多元变量径流预测的最小二乘支持向量机模型[J]. 水力发电学报,2010,29(3):28 - 33.

[15]张豪,罗亦泳,张立亭,等. 基于遗传算法最小二乘支持向量机的耕地变化预测[J]. 农业工程学报,2009,25(7):226 - 231.

[16]周 竹,李小昱,李培武,等. 基于 GA - LSSVM 和近红外傅里叶变换的霉变板栗识别[J]. 农业工程学报,2011,27(3):331 - 335.

[17]李 佳,王 黎,马光文,等. LS - SVM 在径流预测中的应用[J]. 中国农村水利水电,2008(5):8 - 10 + 14.

[18]崔东文. 基于多元变量组合的回归支持向量机集成模型及其应用[J]. 水利水运工程学报,2014(2):66 - 73.

[19]曾 鸣,吕春泉,田 廊,等. 基于细菌群落趋药性优化的最小二乘支持向量机短期负荷预测方法[J]. 中国电机工程学报,2011,31(34):93 - 99.

[20]刘吉臻,吕 游,杨婷婷. 基于变量选择的锅炉 NO_x 排放的最小二乘支持向量机建模[J]. 中国电机工程学报,2012,32(20):102 - 107.