

基于 GA-LSSVR 的渭河水质参数遥感反演研究

谢屹鹏, 汪西莉

(陕西师范大学计算机科学学院, 陕西 西安 710062)

摘要:针对渭河水质参数遥感反演这一典型的非线性、小样本回归估计问题,引入最小二乘支持向量回归(LSSVR)方法来解决,它将 SVR 中的二次规划问题转化为线性方程组求解,在保证精度的同时极大地降低了计算复杂性,加快了求解速度;针对其参数难以选择的问题,利用遗传算法(GA)来优选模型参数。采用提出的方法对标准数据集进行了实验,并建模对渭河的 4 种水质参数 COD_{mn}(高锰酸盐指数)、NH₃-N(氨氮)、DO(溶解氧)、COD(化学需氧量)进行了遥感反演,结果表明 GA-LSSVR 模型可用于解决复杂的回归问题并具有较好的预测性能。

关 键 词:遗传算法;最小二乘支持向量回归;渭河;遥感;反演;水质参数

中图分类号:TP 79 **文献标志码:**A **文章编号:**1004-0323(2010)02-0257-06

1 引 言

渭河流域陕西段地处陕西中部关中地区,是陕西国民经济发展的中心。随着人口增加,国民经济的飞速发展,工农业用水和排污量日益加大,造成了沿河点、线、面污染明显增多,并对区域的可持续发展造成了严重的影响^[1]。因此开展该区域水环境的监测,为相关部门的预报、决策提供支持显得尤为紧迫。传统的水质监测一般采取实地取样分析,消耗大量的人力、物力、财力,且获得的数据呈点状分布,在时间和空间上都不连续,难以达到大范围、实时性的水质监测要求。而水质遥感监测由于具有快速、连续、动态、范围广、相对成本低等优点,在水质监测中的应用日益受到重视^[2-4],鉴于此,本文将利用遥感技术对渭河流域陕西段进行水质遥感监测的研究。

遥感反演水质是一个非线性的过程,其影响因素较多,而且要获得与遥感图像相对同步的实际水质检测数据比较困难,往往导致获取的样本点较少,因此,该反演过程成为了一个典型的小样本、非线性的回归估计问题,本文提出采用支持向量机回归理论来解决该问题。

支持向量机(Support Vector Machines, SVM)是 20 世纪 90 年代由 Vapnik 等提出的一种基于统

计学习理论的新型机器学习方法,作为统计学习理论中结构风险最小化准则的具体实现,支持向量机具有结构简单、全局最优,不需要大量样本,泛化能力较好的优点,近几年得到了广泛的研究及应用^[5,6]。近年来, Suykens^[7]提出最小二乘支持向量机 LSSVR (Least Squares Support Vector Machines),它是标准支持向量机的一种扩展,应用到模式识别和非线性回归估计(LSSVR, Least Squares Support Vector Regression)中,取得了较好的效果。和标准支持向量机一样,其性能依赖于学习机的参数,然而到目前为止,还没有指导 LSSVR 参数选择的好方法。如何确定最优参数,一直是提高 LSSVR 学习和泛化能力的主要问题之一。本文提出利用具有较强的全局搜索能力的遗传算法(Genetic Algorithm, GA)来对 LSSVR 建模过程中的重要参数进行优化,使用提出的方法首先对 3 个标准数据集进行实验,验证了模型可行,然后对渭河 4 种水质参数建模进行遥感反演,最后对结果进行了分析。

2 最小二乘支持向量回归(LSSVR)理论

最小二乘法 LS(Least Squares)是解决多元函

数回归的经典方法,Suykens^[7]等在 SVM 的基础上引进了 LS,将不等式约束改为等式约束,避免了求解耗时的 QP 问题,从而很大程度上降低了运算时间。设训练样本为: $(x_1, y_1), (x_2, y_2) \cdots (x_l, y_l)$, 其中 y_i 是目标值, x_i 是输入向量, 函数估计问题可以描述为求解下面的问题:

$$\min_{\omega, b, e} J(\omega, e) = \frac{1}{2} \omega^T \omega + \gamma \frac{1}{2} \sum_{i=1}^l e_i^2 \quad (1)$$

$$\text{s. t. } y_i = \Phi(x_i) \cdot \omega^T + b + e_i, i = 1, \dots, l$$

其中: $\Phi(\cdot)$ 是从输入空间 R^n 到特征空间 R^m 的映射; 权矢量 $\omega \in R^m$; b 是偏差量; 松弛变量 $e_i \in R$, 表示训练样本的预测值与实际值之间的容许误差; γ 是正则化参数, 可调, 它表示对训练样本的预测值与实际值之间误差超出 e_i 的惩罚程度。核函数可以将原始空间中的样本映射为高维特征空间中的一个向量, 以解决线性不可分的问题。用拉格朗日法求解这个优化问题:

$$L(\omega, b, e, \alpha) = J(\omega, e)$$

$$- \sum_{i=1}^l \alpha_i (\omega^T \Phi(x_i) + b + e_i - y_i) \quad (2)$$

其中: $\alpha_i (i = 1, \dots, l)$ 是拉格朗日乘子。根据优化条件:

$$\frac{\partial L}{\partial \omega} = 0 \rightarrow \omega = \sum_{i=1}^l \alpha_i \Phi(x_i) \quad (3)$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^l \alpha_i = 0 \quad (4)$$

$$\frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma e_i \quad (5)$$

$$\frac{\partial L}{\partial \alpha^i} = 0 \rightarrow \omega^T \Phi(x_i) + b + e_i - y_i = 0 \quad (6)$$

可得:

$$\begin{bmatrix} 0 & 1_v^T \\ 1_v & \mathbf{K} + \frac{1}{\gamma} 1_v 1_v^T \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (7)$$

其中: $x = [x_1, x_2, \dots, x_l]$, $y = [y_1, y_2, \dots, y_l]$, $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_l]$, $1_v = [1, \dots, 1]^T$, \mathbf{K} 为矩阵, 其第 i 行 j 列的元素为:

$$K_{ij} = K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j), i, j = 1, \dots, l$$

其中: $K(x_i, x_j)$ 为核函数, 是满足 Mercer 条件的对称函数。

最小二乘支持向量机回归估计结果为:

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

其中: α, b 由式(7)求解出。

常用的核函数有:

$$(1) \text{ 线性核: } K(x, y) = x^T \cdot y$$

$$(2) \text{ D 阶多项式核: } K(x, y) = (x^T \cdot y + 1)^d$$

(3) 径向基(RBF)核:

$$K(x, y) = \exp\left(\frac{-(x - y)^2}{2\sigma^2}\right)$$

在核函数中, 本文选择 RBF 核, 它只有一个待参数——核函数宽度 σ^2 ; 其值越大, 收敛速度越快, 但是由此得到的模型, 在预测时会使得所有的预测值趋向于某一个值, 这个值往往是取值范围的平均值。对于采用 RBF 核的 LSSVR 其主要参数是正则化参数 γ 和核函数宽度 σ^2 , γ 是权衡拟合曲线光滑度并使拟合误差最小化的参数, γ 值越大, 训练数据样本点拟合值越符合实际值, γ 减小会降低模型的复杂性, 这两个参数决定了模型的预测能力^[8]。在应用中, 正则化参数 γ 和核函数宽度 σ^2 的取值选择成为一大难题, 实际操作中也没有切实可行的参数调节方法, 多数是使用试凑法结合平时的经验, 但这种方法费时费力, 而且取得的效果也不一定很好。因此, 下文中将利用 GA(Genetic Algorithm)智能优化算法对这两个参数进行优选。

3 遗传算法(GA)优选 LSSVR 参数

3.1 遗传算法(Genetic Algorithm)

GA^[9]利用生物遗传学的观点, 结合了适者生存和随机信息交换的思想, 通过自然选择、交换、变异等作用机制, 实现种群的进化。在寻优过程中, GA 在解空间随机产生多个起始点并同时开始搜索, 由适应度函数来指导搜索方向, 是一种能够在复杂搜索空间快速寻求全局最优化的搜索技术。目前已在优化、机器学习和并行处理等领域得到越来越广泛的应用。

(1) 选择编码方式: 本文采用浮点数编码方式, 以避免二进制编码方式在遗传操作时进行反复译码、编码的操作; 此外, 可以克服二进制字符串长度有限的影响, 从而提高进化算法的性能和求解精度。

(2) 设计适应度函数: 由于适应度函数的作用是对个体作评价, 因此适应度函数的设计至关重要。在函数回归估计中, 比较看重的是所建模型的预测能力, 因此, 我们设计的适应度函数为:

$$F(\gamma, \sigma^2) = \frac{1}{\text{MSE}} \quad (9)$$

其中: MSE 为测试(或训练)样本的均方误差:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (10)$$

模型预测结果越好, MSE 的值就越小, 适应度值就

越大。

(3) 遗传操作:

选择操作:本文采用基于排序的选择策略,首先将群体按个体适应度好坏依次排序,然后根据一个线性或非线性函数分配选择概率。本文采用非线性的归一几何排名函数,按下式确定第 i 个个体被选择的概率 p_i :

$$p_i = r \times (1 - q)^{i-1} \tag{11}$$

其中: i 为个体排序序号; r 为排序第一的个体的选择概率。

$$r = \frac{q}{1 - (1 - q)^n} \tag{12}$$

其中: q 为最优个体被选择的概率, n 为种群规模。选择概率仅取决于个体在种群中的序位,而不是实际的适应度值。以经过转换后的概率 p_i 采取轮盘赌方法完成选择操作。

交叉操作:采用线性组合的交叉操作方式。例如以某一概率对某两个染色体 x_1 、 x_2 进行交叉操作时,可以采用如下方式:

$$x_1 = ax_1 + (1 - a)x_2 \tag{13}$$

$$x_2 = (1 - a)x_1 + ax_2 \tag{14}$$

其中: a 为 0,1 之间随机数。

变异操作:在将变异的染色体中随机选择一个变异位 j , 把它设置为一个归一化的随机数 $U(a_i, b_i)$ 。 a_i 、 b_i 为对应该变异位的上下限:

$$x_j = \begin{cases} U(a_i, b_i) & \text{if } i = j \\ x_i & \text{otherwise} \end{cases} \tag{15}$$

3.2 GA 优选模型参数的算法步骤

将GA应用于LSSVR参数优选时,算法步骤如下:

Step1 $t=0$ (t 表示进化代数);

Step2 对模型的正则化参数 γ 和核参数 σ^2 编码;

Step3 随机选择初始种群 $P(t)$;

Step4 运用训练样本集训练模型,并用测试样

本(或训练样本)进行测试得到预测结果,可计算出样本预测结果的 MSE,从而得到个体适应度函数值 $F(\gamma, \sigma^2)$;

Step5 判别 GA 的停止准则是否满足,若满足转 Step9;

Step6 $t=t+1$;

Step7 应用选择算子在 $P(t-1)$ 中选择 $P(t)$;

Step8 对 $P(t)$ 进行交叉、变异操作后转 Step4 步;

Step9 结束,输出最佳的正则化参数 γ 和核参数 σ^2 ,然后代入模型对测试样本数据进行训练以获得精确的预测结果。

4 标准数据集实验及分析

为了验证模型的可行性,本文采用了 3 种标准数据集^[10]进行实验,其数据生成函数如表 1 所示。为了体现方法适于小样本情况,这里每一种数据集均生成 25 个样本,前 15 个样本为训练样本,接下来的 5 个样本为测试样本 A(下文简称:TestA),最后 5 个样本为测试样本 B(下文简称:TestB)。实验使用了 3 种模型,分别为最小二乘支持向量回归模型 LSSVR (Least Squares Support Vector Regression)、基于遗传算法(GA)优选参数的标准支持向量回归模型 GA-SVR (Genetic Algorithm-Support Vector Regression)和基于遗传算法(GA)优选参数的最小二乘支持向量回归模型 GA-LSSVR(Genetic Algorithm-Least Squares Support Vector Regression)。对于 LSSVR 用留一法和网格搜索相结合的方法来优选参数,对于 GA-SVR 和 GA-LSSVR 中的 GA 采用 GAOT 工具箱,在适应度函数的设计上采用了两种形式,一是采用计算训练样本的均方误差(Mean Square Error, MSR)的倒数,二是采用计算测试样本 A 的 MSR 的倒数,两者中的 MSE 的计算公式如式(10)所示。

表 1 实验用标准数据集
Table 1 The standard data sets for experiment

数据集名称	函数表达式	变量取值范围	数据集大小
2-d Mexican Hat	$y = \text{sinc} x = \frac{\sin x }{ x }$	$x \in U[-2\pi, 2\pi]$	25
Multi	$y = 0.79 + 1.27x_1x_2 + 1.56x_1x_4 + 3.42x_2x_5 + 2.06x_3x_4x_5$	$x_i \in U[0, 1], i = 1, 2, \dots, 5$	25
Polynomial	$y = 1 + 2x + 3x^2 + 4x^3 + 5x^4$	$x_i \in U[0, 1]$	25

表 2 为 GA 算法中采用 15 个训练样本的 MSE 的倒数为适应度值,在模型建立后,对 TestA 和

TestB 进行预测的结果;表 3 显示的为 GA 算法中采用 TestA 的 MSE 的倒数为适应度值,在模型建

立后,对 TestA 和 TestB 进行预测的结果。表 2 和表 3 中的 R^2 表示样本的可决系数,利用公式(16)计算,其中, y_i 表示样本的原始值, \hat{y}_i 表示样本的预测值, \bar{y} 表示样本原始值的均值。

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (16)$$

表 2 GA 中采用训练样本的 MSE 的倒数为适应度值
Table 2 The reciprocal of training samples' MSE value adopted to GA's fitness value

测试样本		LSSVR		GA-SVR		GA-LSSVR	
		MSE	R^2	MSE	R^2	MSE	R^2
2-d Mexican Hat	TestA	3.000000E-06	9.999970E-01	2.000000E-06	9.999960E-01	6.000000E-08	9.999999E-01
	TestB	1.000000E-08	9.999999E-01	4.000000E-10	1.000000	1.000000E-10	1.000000
Multi	TestA	0.042817	0.963178	0.037109	0.967311	0.026303	0.973345
	TestB	0.078274	0.862561	0.067086	0.864872	0.062269	0.874304
Polynomial	TestA	0.001742	9.999430E-01	6.540000E-04	9.999898E-01	3.000000E-06	9.999990E-01
	TestB	0.001969	9.999080E-01	1.650000E-04	9.999903E-01	5.700000E-05	9.999930E-01
平均运行时间/s		25.579		108.141		25.391	

表 3 GA 中采用 TestA 的 MSE 的倒数为适应度值
Table 3 The reciprocal of TestA's MSE value adopted to GA's fitness value

测试样本		LSSVR		GA-SVR		GA-LSSVR	
		MSE	R^2	MSE	R^2	MSE	R^2
2-d Mexican Hat	TestA	3.000000E-06	9.999970E-01	2.000000E-09	9.99999E-01	5.000000E-10	1.000000
	TestB	1.000000E-08	9.999999E-01	8.000000E-11	1.000000	3.000000E-11	1.000000
Multi	TestA	0.042817	0.963178	0.025924	0.979430	0.024610	0.981028
	TestB	0.078274	0.862561	0.109477	0.747876	0.105564	0.880786
Polynomial	TestA	0.001742	9.999430E-01	4.180000E-04	9.999850E-01	2.600000E-05	9.999990E-01
	TestB	0.001969	9.999080E-01	4.350000E-04	9.999820E-01	1.610000E-04	9.999840E-01
平均运行时间/s		25.781		112.776		26.672	

观察实验结果,我们不难发现,针对衡量预测精度的指标 MSE 和 R^2 ,对于这 3 种标准数据,3 种模型的预测结果都比较精确。但是横向比较,会发现 GA-LSSVE 的预测结果要好于 LSSVR 和 GA-SVR;针对 GA-LSSVR 进行纵向比较,不难看出,GA 中采用 TestA 的 MSE 的倒数为适应度值的算法要好于 GA 中采用训练样本的 MSE 的倒数为适应度值的算法;观察 3 种模型的运行时间,发现使用 LSSVR 和 GA-LSSVR 的运行时间要远远少于 GA-SVR,这与 LSSVR 将 SVR 中的二次规划问题转化为线性方程组求解,加快了求解速度有关。因此,我们可以采用 GA-LSSVR 算法来对遥感水质反演进行实验,其中 GA 采用测试样本的 MSE 的倒数为适应度值。

5 遥感数据实验及分析

5.1 数据及预处理

首先,我们获取了准同步的 SPOT5 1A 级数据

和实地监测数据,包括 11 幅 SPOT5 数据影像(时间为:2003 年 9 月 14 日,2003 年 12 月 16 日,2004 年 1 月 27 日,2004 年 3 月 9 日,2005 年 3 月 24 日,2005 年 4 月 13 日,2005 年 5 月 10 日,2005 年 12 月 19 日,2006 年 6 月 8 日,2006 年 6 月 9 日)。实地监测数据来源于渭河不同时间的 3 个监测断面,数据有 13 对。根据现行水质标准,结合近几年陕西省环境公报公布的渭河水质状况,以有机污染为主,并参考监测站的监测数据,本文选取了 CODm(高锰酸盐指数)、NH₃—N(氨氮)、DO(溶解氧)、COD(化学需氧量)4 个水质参数进行反演分析。

对遥感数据进行了辐射定标、大气校正、几何校正等预处理。

5.1.1 辐射定标和大气校正

辐射定标是将传感器记录的遥感影像灰度值(即 DN 值)转换成传感器的入瞳辐亮度的处理过程,并且,由于遥感影像在成像过程中,受到大气散射、折射及吸收等一系列因素的影响,造成了卫星传

感器接收到的地面辐射信息存在一定的辐射失真,因此在进行遥感反演的过程中进行大气校正也是非常必要的。本文根据蒋赛^[11]等得出的结论,采用辐射定标后各波段归零的方法。校正公式为:

$$\text{辐射值} = \frac{\text{DC}}{\text{GAIN}} + \text{BIAS} \tag{17}$$

其中:DC 为校正后的像元灰度值 $\text{DC} = \text{DN} - L_n$, DN 为卫星测得的灰度值, L_n 为近红外波段的最小值;GAIN 为传感器增益,BLAS 为传感器偏置,遥感图像的辅助数据文件 METADATA.DIM 中以 XML 的形式记录了遥感数据的各种参数,GAIN 和 NLAS 以标签 PHYSICAL_GAIN 和 PHYSICAL_BIAS 给出。

5.1.2 几何校正

本文在 ERDAS 8.7 运行环境下,以 1 : 50 000 的地形图为参考图,通过地形图和遥感图像上选取同名地物控制点来对遥感影像图进行几何校正。选取 16 个控制点,采用二次多项式校正模型,重采样采用最近邻插值法,最终均误差控制在 0.35 个像元以内。

通过以上步骤,可以将遥感数据转化为在实验中使用的各参数对应的水质数据。

5.2 实验结果及分析

结合赵玉芹^[12]等得出的相关性分析的结果,本文采用多元线性回归、LSSVR 和 GA-LSSVR 3 种方法进行了反演。由于样本只有 13 个,我们采用了

K—折交叉验证的思想(这里 K 取 4),并考虑到数据的大小不一致性,为了使每一组测试样本和训练样本的数据分布较为均匀,在 4 个分组上我们在保证样本独立性的基础上(即保证 4 个分组中的数据相互独立),尽量使每一组训练样本中的数据都能够包含较大、中等及较小的数据,而测试样本也符合这种分布。每次试验中我们取其中的 1 个分组作为测试样本,其余的 3 个分组作为训练样本,试验完成后,我们取预测结果最好的分组来进行比较。对于 3 种方法反演结果的评价我们采用测试样本的平均绝对偏差(MAD)、均方根误差(RMSE)和样本整体的可决系数(R^2)3 种指标进行评价。其中 MAD 利用公式(19)计算,RMSE 为 MSE 的平方根,利用公式(10)计算, R^2 利用公式(16)计算。实验结果如表 4 所示。

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \tag{18}$$

表 4 中给出了线性回归模型、LSSVR 模型和 GA-LSSVR 模型对遥感水质参数反演的结果,从实验结果中可以看出,对于 CODmn、NH₃—N、DO、COD 4 种水质参数,使用 LSSVR 和 GA-LSSVR 进行反演得到的各水质参数的测试样本的 MAD、RMSE 和样本整体的 R^2 要优于多元线性回归的结果,这与支持向量机适用于小样本的理论相吻合;而 GA-LSSVR 的结果要优于 LSSVR,这与 GA 算法能够很好地找到模型的参数有关,从另一方面也说明了模型的参数对模型的预测能力具有决定作用。

表 4 3 种方法反演水质参数结果比较

Table 4 The regression results by three methods for water quality parameters to compare

水质变量	多元线性回归			LSSVR			GA-LSSVR		
	MAD	RMSE	R^2	MAD	RMSE	R^2	MAD	RMSE	R^2
CODmn	4.6967	7.2283	0.8574	4.9374	6.0108	0.9782	3.3915	5.4979	0.9822
NH ₃ —N	3.0118	4.5733	0.4833	2.0983	3.1445	0.9319	1.8497	3.1248	0.9362
DO	0.7258	0.9186	0.8132	0.5113	0.5407	0.9840	0.4900	0.5172	0.9853
COD	30.4505	41.7274	0.7895	20.671226	22.197244	0.9813	13.5643	17.8914	0.9879

6 结 语

本文利用最小二乘支持向量机回归理论来解决遥感水质反演问题,并且针对模型参数难以选择的问题引入了遗传算法,通过标准数据集的实验,得出 GA-LSSVR 的预测精度有了一定的提高,而求解速度与 GA-LSSVR 相比提高将近 3 倍。于是,利用该方法对渭河的 CODmn、NH₃—N、DO、COD 4 种水质参数进行了反演,并通过与多元线性回归和 LSS-

VR 的对比,得出利用 GA-LSSVR 模型可以用于遥感水质参数的反演,能够获得较为理想的结果。

参考文献:

[1]

Zhang Rongzhen,Zhang Xing. The Evolution of the Ecological Environment in Shaanxi Part of Weihe River Basin in Recent 50 Years[J]. Journal of Arid Land Resources and Environment, 2008,22(2):37-42. [张蓉珍,张幸. 渭河流域陕西段近 50 年生态环境演变[J]. 干旱区资源与环境,2008,22(2):37-42.]

[2]

Giardino C,Pepe M. Detecting Chlorophyll,Secchi Disk Depth

- and Surface Temperature in a Sub-alpine Lake Using Landsat Imagery[J]. The Science of the Total Environment, 2001, 268:19-29.
- [3] Dall'Olmo G, Gitelson A A, Rundquist D C, *et al.* A Assessing the Potential of SeaWiFS and MODIS for Estimating Chlorophyll Concentration in Turbid Productive Waters Using Red and Near-infrared Bands[J]. Remote Sensing of Environment, 2005, 96:176-187.
- [4] Gu Liang, Zhang Yuchao, Qian Xin, *et al.* Study on Retrieval of Chlorophyll-a Concentration by Remote Sensing in Taihu Lake[J]. Science and Management, 2007, 32(6): 5-29. [顾亮, 张玉超, 钱新, 等. 太湖水域叶绿素 a 浓度的遥感反演研究[J]. 环境科学与管理, 2007, 32(6): 25-29.]
- [5] Cherkassky V, Ma Y Q. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression[J]. Neural Networks(S0893-6080), 2004, 17(1):113-126.
- [6] Chang Yuqing, Zou Wei, Wang Fuli, *et al.* Research on Soft Sensing Method Based on Support Vector Machines[J]. Control and Decision, 2005, 20(11):1307-1310. [常玉清, 邹伟, 王福利, 等. 基于支持向量机的软测量方法研究[J]. 控制与决策, 2005, 20(11):1307-1310.]
- [7] Suykens J A K, Vandewall J. Least Squares Support Vector Machine Classifiers[J]. Neural Processing Letters, 1999, 9(3):293-300.
- [8] Han Liwei, Li Zongkun, Wang Jianyou, *et al.* Study and Application of Inverse Method for Determining Parameters in Least Squares Support Vector Machine[J]. Advances in Science and Technology of Water Resources, 2009, 29(1):8-11. [韩立伟, 李宗坤, 王建有, 等. 最小二乘支持向量机参数反演方法及其应用[J]. 水利水电科技进展, 2009, 29(1):8-11.]
- [9] Cristianini N, Shawe-taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M]. Beijing: Publishing House of Electronics Industry, 2004.
- [10] Zhou ZhiHua, Wu Jianxin, Tang Wei, *et al.* Ensembling Neural Networks: Many Could Be Better than All [J]. Artificial Intelligence, 2002, 137(1-2):239-263.
- [11] Jiang Sai, Wang Xili, Zhao Yuqin, *et al.* Study on the Effect of Atmospheric Correction in the Weihe River Quality Quantitative Retrieving of Remote Sensing[J]. Remote Sensing Technology and Application, 2009, 24(2):204-209. [蒋赛, 汪西莉, 赵玉琴, 等. 渭河定量遥感水质反演中的大气校正作用研究[J]. 遥感技术与应用, 2009, 24(2):204-209.]
- [12] Zhao Yuqin, Wang Xili, Jiang Sai, *et al.* Study on Neural Network Model for Weihe River Water Quality Retrieving Using Remote Sensing Image[J]. Remote Sensing Technology and Application, 2009, 24(1):63-67. [赵玉芹, 汪西莉, 蒋赛, 等. 渭河水质遥感反演的人工神经网络模型研究[J]. 遥感技术与应用, 2009, 24(1):63-67.]

Study on GA-LSSVR for Weihe River Quality Parameters Retrieving by Remote Sensing

XIE Yi-peng, WANG Xi-li

(School of Computer Sciences, Shannxi Normal University, Xi'an 710062, China)

Abstract: In order to solve Weihe River water quality retrieving by remote sensing, which is regression estimation problem characterized by non-linear and small sample, least squares support vector regression (LSSVR) is introduced in this paper. It transforms the quadratic programming problems into linear equations, which reduce the computational complexity greatly, increase the speed of computing, and meanwhile assure the accuracy. To overcome difficulties in selecting the parameters of the model, genetic algorithm (GA) is used to optimize the parameters of the model. The proposed methods are used to carry on experiments to the standard data sets, and retrieve four water quality parameters COD_{mn} (potassium permanganate index), $\text{NH}_3 - \text{N}$ (ammonia nitrogen), DO (dissolved oxygen), COD (chemical oxygen demand) for Weihe River. The results show that GA-LSSVR model can be used to solve complex regression problems and has better prediction performance.

Key words: Genetic algorithm; Least squares support vector regression; Weihe river; Retrieve; Remote sensing; Water quality parameters