应用 GA-SVM 的渭河水质参数多光谱遥感反演

汪西莉[1], 周兆永[1,2], 延军平[3]

1. 陕西师范大学 计算机科学学院, 陕西 西安 710062；2. 西北农林科技大学网络中心, 陕西 杨凌 712100；
3. 陕西师范大学 旅游与环境学院, 陕西 西安 710062

摘 要: 建立了基于支持向量机的遥感水质参数反演模型, 构建了基于浮点数编码的遗传算法优选模型参数。以渭河为研究对象, 基于高分辨率多光谱遥感 SPOT-5 数据和水质实地监测数据, 分别建立了一元和多元经验模型进行渭河水质参数的反演。在样本数目有限的情况下, 提出的 GA-SVM 方法的反演结果比神经网络和传统的统计回归方法好, 且各方法的多元回归结果均好于一元回归的结果。SVM 具有强的非线性映射能力, 适合小样本情况, 由 GA 实现了模型参数的自动优选, 使 GA-SVM 用于解决回归问题表现出优势。将机器学习和全局优化智能计算方法引入, GA-SVM 为渭河陕西段的水环境遥感监测提供了一种新方法, 取得了较好的反演结果。

关键词: 支持向量机, 遗传算法, 水质参数, 反演, 渭河, SPOT-5

中图分类号: X832/TP79 文献标识码: A

1 引 言

渭河是黄河最大的支流, 全长 818km, 流经甘肃、宁夏、陕西 3 省区。渭河流域陕西段地处陕西关中地区, 流经西安、咸阳等 5 市集中了陕西 64%的人口、56%的耕地和 82%的工业总产值, 这里是陕西省政治、经济、文化、金融及信息中心。渭河陕西段是关中沿渭各市最重要的城市水源地, 随着该地区人口增长、经济发展、用水和排污量不断增加, 渭河水质受到严重污染, 有机污染更为严重(张玉清, 2000 ;陕西省环保厅, 2006), 其水环境质量监测意义重大。常规水质监测方法不能满足对水质的实时、大尺度的监测要求, 难以实现连续、快速的跟踪调查与分析; 基于遥感的水质监测具有大范围、快速、动态监测的优势, 可以作为常规水质监测的补充, 提供更多的信息。以往受遥感器空间分辨率的限制, 关于水环境的遥感研究主要针对海洋、湖泊等大的水体, 目前已有多种遥感器提供高空间分辨率的遥感信息, 尽管其光谱分辨率不够高, 但还是为河流的遥感水质监测提供了可能(施明伦等, 2006)。

针对内陆水体遥感监测的水质指标, 研究较多和相对比较成熟的是悬浮物和叶绿素 a (Carpenter & Carpenter, 1983；Lathrop & Lillesand, 1986；Ritchie 等, 1987)。其他指标如溶解氧(DO)、化学需氧量(COD)、5 日生化需氧量(BOD5)、总氮(TN)、总磷(TP)等也开展了研究(施明伦等, 2006；王建平等, 2003)。目前多采用传统的统计回归方法建立参数模型, 实现反演(Carpenter & Carpenter S M, 1983；Lathrop & Lillesand, 1986；Ritchie 等, 1987；王学军 & 马廷, 2000；Ana 等, 2007)。近年来非线性映射方法——人工神经网络也被用于水质遥感研究, 取得了较好的结果(石爱业等, 2006；Ana 等, 2007)。其黑箱的特征避免了指定回归函数的形式, 可以实现任意的非线性映射, 而且网络中隐含了特征提取的过程, 这些是统计回归所不及的。神经网络的问题是网络结构难以很好地确定, 其效果依赖于人的经验和样本数量。

本文采用一种新的机器学习方法——支持向量机(support vector machine)(Vapnik , 1995)进行遥感反演。SVM 建立在统计学习理论的基础上, 可以根据有限的样本信息在模型复杂性和学习能力之间寻求最佳折中, 以获得最好的推广能力(预测精度), 它在解决小样本、非线性和高维学习中表现出许多

收稿日期: 2008-08-29; 修订日期: 2008-12-05

基金项目: 国家自然科学基金(编号: 40671133)资助。

第一作者简介: 汪西莉(1969—), 女, 博士, 教授。主要从事遥感图像处理、模式识别、智能信息处理等研究。E-mail: wangxili@snnu.edu.cn。

特有的优势(Burges , 1998；Vapnik 等, 1997)。本文利用 SVM 建模, 采用遗传算法(genetic algorithm)解决 SVM 的参数自动优选问题, 针对渭河实现水质参数反演, 取得了较好的结果。

2 数据预处理及分析

样本包括 2000—2006 年间渭河陕西段的部分水质实地监测数据, 以及 11 幅 SPOT-5 遥感影像数据。其中准同步的样本数据有 13 对, 分布在 3 个代表断面, 如图 1(a)1#—林家村; (b)2#—咸阳; (c)3#—耿镇。考虑到渭河的主要污染物为有机污染物, 结合获取数据的实际情况, 选取 4 种代表性的水质参数：高锰酸盐指数(COD_{mn})、氨氮(NH_3-N)、化学需氧量(COD)、溶解氧(DO)进行研究。对遥感数据进行了预处理, 包括传感器定标和几何校正。

图 1 渭河部分河段遥感影像及 3 个采样点位置
(a) 1#—林家村; (b) 2#—咸阳; (c) 3#—耿镇

表 1、表 2 给出了 SPOT-5 数据 4 个波段间的相关性及波段与水质参数间的相关性。

表 1 SPOT-5 波段间的相关性

	$\ln R_1$	$\ln R_2$	$\ln R_3$	$\ln R_4$
$\ln R_1$	1.0000			
$\ln R_2$	0.8379	1.0000		
$\ln R_3$	0.6472	0.9122	1.0000	
$\ln R_4$	0.3491	0.7055	0.8222	1.0000

表 2 波段与水质参数间的相关性

	$\ln COD_{mn}$	$\ln NH_3$-N	lnDO	lnCOD
$\ln R_1$	–0.6444	–0.6290	0.4060	–0.7661
$\ln R_2$	–0.6862	–0.5016	0.3114	–0.7086
$\ln R_3$	–0.5427	–0.3221	0.1655	–0.4997
$\ln R_4$	–0.1311	0.0531	–0.2087	–0.1087

可见波段和波段间大多具有较强的相关性。DO 和波段的相关性较低, 其他 3 个水质参数和波段具有较好的相关性, 特别是和可见光波段(R_2、R_3)及红外波段(R_1), 而短波红外(R_4)波段与各水质参数间的相关性最差。这表明某些波段值与水质参数的变化有较强的关联, 据此可选择和各水质参数相关性最高的波段做一元回归, 考虑到综合各波段可以提供更多信息, 也将基于全部 4 个波段实现多元回归。

3 研究方法

利用遥感数据和准同步的实地监测数据, 基于 SVM 建立非线性反演模型, 并采用 GA 自动优选模型参数。

3.1 基于 SVM 的回归

设样本集合为 $X=\{(\boldsymbol{x}_i, y_i)\,|\,x\in \mathrm{R}^n, y\in \mathrm{R}, i=1, 2,\cdots,n\}$, $\boldsymbol{x}$ 和 y 存在函数依赖关系：$F=\{f:\mathrm{R}^n\to \mathrm{R}\}$。回归问题就是寻找一个最优函数 $f\in F$ 使得期望风险 $R(f)=\int L(y, f(\boldsymbol{x}))\mathrm{d}P(\boldsymbol{x}, y)$ 达到最小, 其中 $L(y, f(\boldsymbol{x}))$ 是损失函数, 若为一次ε不敏感损失函数, 其定义为：

$$L(y, f(\boldsymbol{x}))=|\,y-f(\boldsymbol{x}, \boldsymbol{w})\,|_{\varepsilon}=\begin{cases}0, & |y-f(\boldsymbol{x}, \boldsymbol{w})|\leqslant\varepsilon\\ |y-f(\boldsymbol{x}, \boldsymbol{w})|-\varepsilon, & \text{其他}\end{cases} \tag{1}$$

式中, ε为常数, 反映了拟和的精度。若样本点呈线性关系, 则回归函数为：$f(x)=\boldsymbol{w}\boldsymbol{x}+b$, 式中 $\boldsymbol{w}$, b 分别为线性回归函数的法向量和偏移量。为使回归函数平坦, 必须寻找一个最小的 w, 并考虑拟合误差, 则回归函数的求解可以表示成如下的约束优化问题：

$$\min R(\omega)=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)$$

$$\begin{aligned} \text{s.t.} \quad & f(\boldsymbol{x}_i)-y_i \leqslant \xi_i^*+\varepsilon \\ & y_i-f(\boldsymbol{x}_i) \leqslant \xi_i+\varepsilon \\ & \xi_i,\xi_i^* \geqslant 0 \qquad (i=1,2,\cdots,l) \end{aligned} \tag{2}$$

式中，常数 $C>0$ 决定了对大于ε的偏差的惩罚程度。上面的问题为一个凸二次优化问题，引入拉格朗日函数，并对拉格朗日函数求鞍点，可得该优化问题的对偶问题：

$$\begin{aligned} \min_{\alpha,\alpha^*} & \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)+ \\ & \varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i) \\ \text{s.t.} \ & \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \\ & 0\leqslant\alpha_i,\alpha_i^*\leqslant C, \ \ i=1,2,\cdots,l \end{aligned} \tag{3}$$

通过求解对偶问题，得到最优解 $\bar{\alpha}=(\bar{\alpha}_1,\bar{\alpha}_1^*,\cdots,\bar{\alpha}_l,\bar{\alpha}_l^*)^{\mathrm{T}}$，由此可得回归函数：$f(\boldsymbol{x})=\sum_{i=1}^{l}(\bar{\alpha}_i^*-\bar{\alpha}_i)(\boldsymbol{x}_i\cdot\boldsymbol{x})+\bar{b}$。选择位于开区间(0, C)中的 $\bar{\alpha}_i$ 或 $\bar{\alpha}_k^*$ 可求解 $\bar{b}$：

$$\begin{aligned} & \bar{b}=y_j-\sum_{i=1}^{l}(\bar{\alpha}_i^*-\bar{\alpha}_i)(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)+\varepsilon \text{ 或} \\ & \bar{b}=y_k-\sum_{i=1}^{l}(\bar{\alpha}_i^*-\bar{\alpha}_i)(\boldsymbol{x}_i\cdot\boldsymbol{x}_k)-\varepsilon \end{aligned} \tag{4}$$

若样本点呈非线性关系，可将每一个样本点用一个非线性函数Φ映射到高维特征空间，再在高维特征空间中进行线性回归，从而获得在原空间非线性回归的结果。和很多非线性方法不同的是，这里并不需要知道Φ的形式，因为只涉及高维空间中的内积运算，可以采用适当的核函数 $K(\boldsymbol{x}_i, \boldsymbol{x})$代替高维空间中的内积运算(Vapnik 等, 1997)，这时得到的结果为：$f(\boldsymbol{x})=\sum_{i=1}^{l}(\bar{\alpha}_i^*-\bar{\alpha}_i)K(\boldsymbol{x}_i,\boldsymbol{x})+\bar{b}$。

由于水质参数和光谱呈复杂的非线性关系，本文建立非线性 SVM 模型实现参数反演。

3.2 GA 优选 SVM 模型参数

采用径向基核函数来建立非线性SVM模型，其形式为：$K(\boldsymbol{x},\boldsymbol{x}_i)=\exp\{-\|\boldsymbol{x}-\boldsymbol{x}_i\|^2/\sigma^2\}$。整个模型的参数包括核函数参数$\sigma^2$，惩罚系数 C 和不敏感损失函数的宽度ε，它们决定了模型的类型、复杂程度和精度(Kwok & Tsang, 2003)。

SVM 模型参数的寻优是一个复杂的连续参数优化问题，这里采用遗传算法来解决。GA 基于达尔文进化论适者生存的思想，是一种能够在复杂的搜索空间并行快速寻求全局最优解的智能搜索技术(玄光男 & 程润伟, 2004)。算法通过编码来表示解(即模型参数)，初始随机产生多个解同时开始搜索，由适应度函数指导搜索方向，通过自然选择、交换、变异等作用机制实现解的进化。

3.2.1 编 码

采用浮点数编码，避免了二进制编码在遗传操作时反复编码、译码的操作，克服了二进制字符串长度对数值表示精度的局限，提高了算法的性能和求解精度。

3.2.2 适应度函数

算法根据适应度函数决定搜索的方向，这里将适应度函数定义为：

$$F(\sigma^2,C,\varepsilon)=\frac{1}{\mathrm{MAD}} \tag{5}$$

其中，$\mathrm{MAD}=\sum_{i=1}^{n}|y_i-\hat{y}_i|/n$，$y_i$, $\hat{y}_i$ 分别表示第 i 个测试样本的实际值和预测值。可见平均绝对偏差 MAD 的值越小，对应参数的适应度值就越大，这些较优的参数被遗传到下一代的可能性也越大。

3.2.3 遗传操作

通过选择、交叉、变异这 3 个遗传操作实现进化。选择操作从群体中选择出较适应环境的个体用于繁殖下一代；交叉操作对选中的个体交换相同位置的基因以产生新的个体；变异操作对选中个体中的某些位执行异向转化，以达到增加个体多样性的目的。

① 选择操作　采用基于排序的适应度分派原则，即选择概率取决于个体在种群中的排序序位，而不是实际的适应度值。

② 交叉操作　采用线性组合的交叉操作方式，如下式：

$$\begin{aligned} s_1' &= as_1+(1-a)s_2 \\ s_2' &= (1-a)s_1+as_2 \end{aligned} \tag{6}$$

式中，a 为 0, 1 之间的随机数，s_1, s_2 为父个体，s_1'，s_2' 为子个体。

③ 变异操作　在随机选中的个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，其他位不变。

综上，将GA应用于SVM参数优选并实现SVM回归的过程是：随机生成若干个体(即初始参数)并编码表示，用训练样本集训练 SVM 模型，并用测试样本测试，得到个体适应度函数值，通过选择、交叉、变异操作产生下一代个体(优化了的参数)，重复上述过程，直到得到最优参数，将最优参数带入 SVM 对样本数据得到回归结果。

4 结果与分析

分别基于 SPOT-5 和反演参数的一个最相关波段以及利用所有 4 个波段建立了一元和多元 SVM 反演模型, 利用 GA 选择模型参数。由于样本数较少, 采用 K 折交叉验证法：随机将样本分成 4 个互不相交的子集, 最后一个子集取 4 个样本, 轮流选择其中 3 个子集作为训练集, 剩余的 1 个子集做验证。SVM 利用 LIBSVM 库文件(Chang & Lin, 2006)构建, 实验运行环境及平台采用 Matlab7.0。

采用统计回归方法(Zhou & Wang, 2008)、BP 神经网络(BP-ANN)方法建立了反演模型, 并将其结果和 GA-SVM 模型的结果进行比较。一元和多元回归 BP-ANN 模型的隐层神经元个数分别为 9 个、12 个, 同样采用 K 折交叉验证法训练。

采用 MAD 和可决相关系数 R^2 来评价反演结果, 表 3—表 6 给出了以上 3 种方法对 3 个样本进行预测的结果及 MAD 和 R^2 值。

结果表明, 不论是一元还是多元回归模型, GA-SVM 表现出了明显的优势, 采用平均绝对偏差指标评价时, GA-SVM 方法的反演结果大多好于(个别相当)其他 2 种模型；采用 R^2 指标评价时, 对这 4 种水质参数, GA-SVM 的一元和多元模型得到的 R^2 均大于 0.85, 而另 2 种方法不论是一元还是多元模型达不到这个水平, 表明 GA-SVM 可以得到较准确的预测结果。

3 种方法对各水质参数的多元反演结果好于对应的一元反演结果, 表明没有某一个波段能为水质

表 3　3 种方法反演 COD_{mn} 的结果

测试样本	实际值	统计回归预测值		BP-ANN 预测值		GA-SVM 预测值	
		一元	多元	一元	多元	一元	多元
1	19.5	27.44	31.21	78.5	17.08	24.94	19.94
2	79.1	24.68	56.16	73.16	42.63	24.34	78.7
3	26.9	23.29	38.59	6.24	26.55	23.84	51.18
MAD		21.99	15.45	28.53	13.08	21.09	8.37
R^2		0.47	0.76	0.25	0.78	0.85	0.90

表 4　3 种方法反演 NH_3-N 的结果

测试样本	实际值	统计回归预测值		BP-ANN 预测值		GA-SVM 预测值	
		一元	多元	一元	多元	一元	多元
1	7.34	9.14	8.98	322.34	6.32	1.01	7.39
2	12.6	3.02	12.17	13.92	95.16	12.76	12.51
3	3.14	2.67	7.49	13.74	80.33	1.14	6.85
MAD		3.95	2.14	108.97	53.59	2.83	1.29
R^2		0.40	0.65	0.30	0.50	0.91	0.94

表 5　3 种方法反演 COD 的结果

测试样本	实际值	统计回归预测值		BP-ANN 预测值		GA-SVM 预测值	
		一元	多元	一元	多元	一元	多元
1	109	204.33	176.07	50.24	117.37	78.92	177.17
2	252	79.32	254.66	303.57	0.67	242.92	227.50
3	183	71.26	168.16	677.57	0.36	185.15	184.47
MAD		126.58	28.19	201.64	146	13.77	31.07
R^2		0.59	0.84	0.38	0.56	0.87	0.91

表 6　3 种方法反演 DO 的结果

测试样本	实际值	统计回归预测值		BP-ANN 预测值		GA-SVM 预测值	
		一元	多元	一元	多元	一元	多元
1	1.57	1.66	1.65	5.75	1.57	2.51	1.57
2	0.3	1.94	0.93	3.9	0.01	2.91	1.07
3	3.6	2.82	1.89	0.42	0.01	4.11	1.24
MAD		0.84	0.81	3.65	1.29	1.35	1.04
R^2		0.21	0.76	0.15	0.60	0.85	0.86

参数反演提供充足的信息，将各波段提供的信息综合起来可以得到更精确的反演结果，采用多元模型要优于一元模型。

GA-SVM 和 BP-ANN 是非线性模型，更适于这里复杂的非线性回归问题。从结果看，神经网络没有比统计回归更优越，其主要原因在于这里的样本集小。若训练样本数太少(至少要达到网络连接边数)，采用神经网络在理论上就存在不确定问题，得到的模型不可信。对于 SVM 模型，即使样本数少，但模型参数合适，仍然有较好的预测精度。这里采用 GA 在参数空间进行全局搜索寻找最佳参数，取得了较好的结果，并且非线性映射的优势体现得很明显，其一元模型的结果比其他两种方法一元模型的结果明显好很多，多元模型结果又比一元模型结果更好。

5 结 论

本文以渭河陕西段为研究对象，采用 SPOT-5 遥感数据进行水质参数遥感反演。在方法上，采用非线性的SVM建模，得到的模型可以方便地实现隐式的非线性映射，不需要大量的样本，推广性好，此外模型在高维情况下不增加训练算法的复杂度，适合推广到高光谱遥感应用中。SVM 模型的参数选择影响到整个模型的性能，本文采用 GA 实现了 SVM 模型参数的自动全局优选。将提出的 GA-SVM 方法用于渭河水质参数反演取得了较好的预测结果，较常用的统计回归和神经网络方法具有明显的优势，为水质遥感监测提供了一种新方法。实验中的样本数量还偏少，希望未来可以增加样本数量，以区分不同的季节，做更多的时空分析及验证。

REFERENCES

Ana C T, Fernando V G and Hernni G. 2007. Retrieving TSM concentration from multispectral satellite data by multiple regression and artificial neural networks, *IEEE Trans on Geoscience and Remote Sensing*, **45**(5), 1342—1350

Burges C J G. 1998. A tutorial on support vector machine for pattern recognition. *Data Mining and Knowledge Discovery*, **2**(2): 121—167

Carpenter D J and Carpenter S M. 1983. Modeling inland water quality using Landsat data. *Remote Sensing of Environment*, **13**(44): 345—352

Chang C C and Lin C J. 2006. http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html (2006-12-01)

Kwok J T and Tsang I W. 2003. Linear Dependency between epsilon and the input noise in support vector regression. *IEEE Trans on Neural Networks*, **14**(3): 544—553

Lathrop R G Jr and Lillesand T M. 1986. Use of thematic mapper data to assess water quality in green bay and central Lake Michigan. *Photogrammetric Engineering and Remote Sensing*, **52**(5): 671—680

Ritchie J C, Cooper C M and Jiang Y Q. 1987. Using Landsat multispectral scanner data to estimate suspended sediments in Moon Lake, Mississippi. *Remote Sensing of Environment*, **23**(1): 65—81

Shaanxi environment protection bureau. 2006. General situation about Weihe. http://www.snepb.gov.cn/admin/pub_newsshow.asp?id=1004630&chid=100243(2007-02-01)

Shi A Y, Xu L Z, Yang X Y and Huang F C. 2006. A neural network model for water quality retrievals using knowledge and remote-sensed image. *Journal of Image and Graphics*, **11**(4): 521—528

Shi M L, You B S, Wan T Z, Luo W Y and Zhang W Z. 2006. Effect of atmospheric correction on stream water quality monitoring by using spot satellite remote sensing images. *Journal of Remote Sensing*, **10**(4): 548—558

Vapnik V. 1995. The Nature of Statistical Learning. New York: Springer.

Vapnik V, Golowich S and Smola A. 1997. Support vector method for function approximation, regression estimation, and signal processing. Advances in Neural Information Processing Systems 9, Cambridge, MA, MIT Press

Wang J P, Cheng S T, Jia H F, Wang Z S and Deng N H. 2003. An artificial neural network model for lake color inversion using TM imagery. *Environmental Science*, **24**(2): 73—76

Wang X J and Ma T. 2000. The application of remote sensing technology in monitoring the water quality of Taihu lake. *Environmental Science*, **21**(6): 65—68

Xuan G N and Cheng R W. 2004. Genetic Algorithm and Engineering Optimizing, Beijing: Tsinghua University Press

Zhang Y Q. 2000. A rational analysis of water pollution factors in Wei river valley and the prevention and control measures. *Journal of Xi'an United University*, **3**(2): 78—82

Zhou Z Y and Wang X L. 2008. Quantitative remote sensing research about water quality of Weihe River based on SPOT-5 imagery. Proceedings of International Conference on Informational Technology and Environmental System Science. Beijing: Publishing House of Electronics Industry

附中文参考文献

陕西省环境保护厅. 2006.渭河概况. http://www.snepb.gov.cn/admin/pub_newsshow.asp?id=1004630&chid=100243(2007-02-01)

石爱业，徐立中，杨先一，黄凤辰. 2006. 基于知识和遥感图像的神经网络水质反演模型. 中国图象图形学报, **11**(4): 521—528

施明伦，游保杉，万腾州，罗文忆，张伟智. 2006. 大气校正对SPOT 卫星遥测水质的影响. 遥感学报, **10**(4): 548—558

王建平，程声通，贾海峰，王志石，邓宁华. 2003. 用 TM 像进行湖泊水色反演研究的人工神经网络模型. 环境科学，**24**(2): 73—76

王学军，马廷. 2000. 应用遥感技术监测和评价太湖水质状况. 环境科学, **21**(6): 65—68

玄光男，程润伟. 2004. 遗传算法与工程优化. 北京：清华大学出版社

张玉清. 2000. 渭河流域水污染成因的探析及防治对策.西安联合大学学报(自然科学版), **3**(2): 78—82