·环境预警·

doi:10.3969/j.issn.1674-6732.2012.02.002

PSO和SVM混合算法确定太湖入湖河流水质主要影响因子

陈 蓓[1],徐国伟[1],王文超[2],孙小妹[2],孙培冬[3],丁彦蕊[2]
(1.无锡市滨湖区环境监测站,江苏 无锡 214072;2.江南大学物联网工程学院,江苏 无锡 214122;3.江南大学化学与材料工程学院,江苏 无锡 214122)

摘 要:以影响太湖入湖河流水质的24个因子值为研究对象,将粒子群优化算法(PSO)与支持向量机算法(SVM)相结合。PSO算法用于优化SVM算法的参数c和g,有利于快速、高效地确定c和g的全局最优值;SVM算法基于最优的c和g,分别以24,21,18,15,12,9和6个因子作为特征向量预测水质的污染程度。结果表明,当特征向量为9个影响因子时预测率最高。其参数$c=18.56$,$g=1.35$,对应的预测率为:全局预测率92.59%,重度污染水质预测率88.89%,轻度污染水质预测率94.45%。因此,通过PSO和SVM混合算法,可以确定影响太湖入湖河流水质的主要因子,利用这些主要因子对水质进行预测预警,不但可以节省时间,而且可以得到精确的结果。

关键词:粒子群优化算法;支持向量机;水体水质;影响因子
中图分类号:X11 **文献标识码**:A **文章编号**:1674-6732(2012)-02-0007-04

Study on the Key Factors Influenced the Water Quality of Rivers Flowing into Taihu Lake Using PSO and SVM Hybrid Algorithm

CHEN Bei[1], XU Guo-wei[1], WANG Wen-chao[2], SUN Xiao-mei[2], SUN Pei-dong[3], DING Yan-rui[2]
(1. Binhu District Environmental Monitoring Station, Wuxi, Jiangsu 214072, China; 2. School of IOT Engineering, Jiangnan University, Wuxi, Jiangsu 214122, China; 3. School of Chemical and Material Engineering, Jiangnan University, Wuxi, Jiangsu 214122, China)

ABSTRACT: 24 factors that influenced the water quality of rivers flowing into Taihu Lake was studied by combining the PSO and the SVM algorithm. The PSO is used to optimize the parameters c and g in SVM, so that the global optimum value of c and g could be searched efficiently and rapidly. Then we use SVM algorithm and take 24,21,18,15,12,9 and 6 influence factors as feature vectors to predict water quality based on the optimal c and g. The results showed that the prediction accuracy is the highest when 9 influence factors is the feature vector. The values of parameter c and g are 18.56 and 1.35 respectively. The corresponding prediction accuracies are computed as follows: the global prediction accuracy is 92.59%, the prediction accuracy of severe pollution water quality is 88.89%, the lightly polluted water quality is 94.45%. Therefore, warning prediction of water quality using these factors through the method of PSO and SVM hybrid algorithm is time saving and accurate.

KEY WORDS: Particle swarm optimization; supported vector machines; water quality; impact factor

自2001年以来,无锡市滨湖区环境监测站对流入太湖的多条河流进行了氨氮(NH_3-N)、总氮(TN)、高锰酸盐指数(COD_{Mn})、总磷(TP)、石油类、镉(Cd)、铅(Pb)、砷(As)、溶解氧(DO)、汞(Hg)、硒(Se)、pH值、水温、铜(Cu)、氟化物(F^-)、硫化物、电导率、5日生化需氧量(BOD_5)、化学需氧量(COD_{Cr})、锌(Zn)、挥发酚、氰化物(CN^-)、六价铬(Cr^{6+})、阴离子表面活性剂(LAS)共计24个河流水质影响因子的检测。由于河流水体的水质状况是由上述多个水质指标组成的复杂系统,在众多的因子中,有的因子是影响水体水质的主要因子,有的因子之间存在错综复杂的关系。因此,在分析过程中可能会出现因为指标间存在共线性问题而无法得到正确结论的情况。为了更好地评价水体的水质,需要确定影响水质的主要因子,并以此进行水体水质的监控和预警。

收稿日期:2011-06-14
基金项目:江苏省环境监测科研基金项目(0902)。
作者简介:陈蓓(1966—),女,高级工程师,本科,从事环境监测工作。

目前评价水体水质的方法主要有主成分分析法和综合污染指数法。主成分分析法是判断影响水质的主要因子的方法之一[1-5]。河流水质系统是一个由多因子构成的复杂系统,其综合评价的数量化指标很多。主成分分析法是利用降维的思想,把多指标转化为少数几个综合指标,这样可以在原始数据信息量丢失最小的情况下,减少评价指标,同时客观地确定权重,减少人为干预。然而主成分分析法是一种线性降维技术,表现为其主成分是原始变量的线性组合。而在实际的水质预测中,各指标间有时存在非线性关系,主成分与原始数据之间也呈现非线性关系,线性的降维不能真实地反映出指标间的关系。综合污染指数法是另一种判断影响水质的主要因子的方法[6]。综合污染指数是各项评价指标的污染指数之和,污染分担率是其中某项指标的污染分指数占综合污染指数的比例,分担率最大的指标为首要污染物。然而该方法没有考虑因子之间的相关性。

笔者采用基于粒子群优化算法(PSO)和支持向量机算法(SVM)的混合算法,利用 PSO 优化 SVM 算法的参数 c 和 g,有利于快速、高效地确定 c 和 g 的全局最优值;接着将测定的河流水质的 24 种影响因子,随机分为含有 24,21,18,15,12,9 及 6 种影响因子的数据集,分别以这些影响因子为特征向量,对水质进行预测,通过预测率高低确定太湖入湖河流水质的主要影响因子。该方法能够准确地从影响水质的各种因子中识别出主要因子,为河流水质预警提供了有力证据。

1　数据集及方法

1.1　数据集

研究对象是无锡市滨湖区环境监测站测定的 24 条太湖入湖河流水质影响因子,时间为 2001 年 1 月—2009 年 12 月,共筛选出 256 个样本。

1.2　方法

1.2.1　数据的归一化

由于测定的 24 个因子的数值范围差别很大,所以有必要对数据进行归一化。笔者采用映射函数将所有的数据都映射到 0 ~ 1 的范围。

1.2.2　PSO 和 SVM 预测

PSO 是受人工生命研究结果的启发,通过模拟鸟群觅食过程中的迁徙和群聚行为而提出的一种基于群体智能的全局随机搜索算法,是一种通过叠代搜寻最优值的优化工具,在多目标优化、分类、模式识别等方面有广泛的应用[7-9]。

SVM 是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的,根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷,以期获得最好的推广能力的方法。在解决小样本、非线性及高维模式识别中表现出许多特有的优势[10-12]。

将 PSO 和 SVM 算法结合,利用 PSO 优化 SVM 的参数,根据 SVM 的预测率对影响水质的主要因素进行识别。

太湖入湖河流水质主要影响因子的分析过程涉及到的分类问题是一个二分类的线性不可分问题。训练集为 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,其中 $x_i \in T^m$ 为输入向量,输出向量为 $y_i \in \{0,1\}$,用一个超平面将训练集划分开,该超平面为 $W \cdot X + b = 0$,W 和 b 决定了超平面的位置,$W \cdot X$ 为两个向量的内积。为了得到最优的划分,该问题被转化为求最优化的超平面。

$$\min \Phi(W,\xi) = 1/2 \| W \|^2 + c \sum_{i=1}^{n} \xi_i \cdot c \geqslant 0 (i = 1,2,\cdots,n)$$

$$y_i[(W \cdot X_i)] \geqslant 1 - \xi_i (\xi_i \geqslant 0)$$

式中:ξ_i——松弛因子;c——对错分样本的惩罚因子。

令 $f(X) = W \cdot X + b$,上式的优化问题可转化为:

$$\min_a \frac{1}{2} \sum_{i=1}^{n} \cdot \sum_{j=1}^{n} y_i y_j a_i a_j (X_i \cdot X_j) - \sum_{i=1}^{n} a_i (i = 1,2,\cdots,n;j = 1,2,\cdots,n) \cdot \sum_{i=1}^{n} y_i a_i = 0 (0 \leqslant a_i \leqslant c)$$

对于非线性可分问题,可以通过一个映射函数(核函数),将低维的输入空间 T^m 映射到高维的特征空间 H,使线性可分。问题就可以被描述为,Ψ:$T^m \rightarrow H$ 映射到高维空间 H 中,根据泛函数的有关理论,只要一种核函数满足 Mercer 条件,它就对应某一空间中的内积,则核函数 $K(X_i,X_j) = \Psi(X_i) \cdot \Psi(X_j)$,则优化问题转化为:

$$\min_a \frac{1}{2} \sum_{i=1}^{n} \cdot \sum_{j=1}^{n} y_i y_j a_i a_j K(X_i,X_j) -$$

$$\sum_{i=1}^{n} a_i (i = 1,2,\cdots,n; j = 1,2,\cdots,n) \cdot$$

$$\sum_{i=1}^{n} y_i a_i = 0 (0 \leqslant a_i \leqslant c)$$

从上式可以看出，不需要知道 H 和 Ψ，只需要知道合适的核函数 K 和 c 就可以确定支持向量。

该处使用的核函数为径向基(RBF)核：$K(X_i, X_j) = \exp(-g \| X_i - X_j \|)^2$。RBF 核在低维、高维、小样本、大样本等情况下都是通用的，是比较理想的分类依据函数，也是 SVM 默认的核函数。

将径向基函数带入上式，最优化问题就转化为下面的最小化问题。

$$\min_a \frac{1}{2} \sum_{i=1}^{n} \cdot \sum_{j=1}^{n} y_i y_j a_i a_j \exp(-g \| X_i - X_j \|)^2) - \sum_{i=1}^{n} = a_i (i = 1,2,\cdots,n; j = 1,2,\cdots,n)$$

式中：$\sum_{i=1}^{n} y_i a_i = 0 (0 \leqslant a_i \leqslant c)$，它的最小值由参数 c 和 g 来决定。

参数 c 的作用是在确定数据的子空间中调节学习机器置信区间范围，不同数据子空间中最优化的 c 不同。核参数 g 的改变实际上隐含地改变了映射函数，从而改变样本空间分布的复杂程度，也就决定了线性分类达到的最小误差。

使用 PSO 算法的主要目的是寻找 SVM 的参数 c 和 g 的最优值，以实现整个过程的最优化。

优化过程分为如下步骤：

(1) 数据集的划分

首先是将数据集划分为训练集和预测集。将数据集分成 10 份，取其中 1 份作为预测集，剩下的作为训练集，如此循环 10 次。然后赋值给训练集和预测集以及相关标签值。

(2) PSO 算法进行参数优化

PSO 算法是基于群体的行为，根据对环境的适应度将群体中的个体移动到好的区域。具体过程如下：

① 初始化粒子群微粒的位置和速度，并初始化 SVM 的参数。

② 评价粒子群中每个微粒的适应度。

③ 对每个微粒，将它的适应度值和经历过的最好位置 pbest 作比较，选择好的作为当前最好位置 pbest。

④ 对每个微粒，将它的适应度和经历过的全局最好位置 gbest 作比较，如果它的适应度更好则重新设置 gbest。

⑤ 利用粒子速度更新方程变化微粒的位置和速度，进行多次迭代，寻找全局最优的适应度值。

⑥ 分类器评价。

算法的评价分为局部预测率和全局预测率，只有两种预测率都较高时所构建的分类器才可靠。

局部预测率：

$$LA = \frac{\sum_{i=1}^{\rho} P_i}{\rho}, \text{其中 } P_i = \frac{T_i}{n_i}$$

全局预测率：

$$TA = \frac{\sum_{i=1}^{\rho} T_i}{N}$$

式中：N——数据集中所有样本的个数；ρ——水质的种类(轻度污染和重度污染两类)；n_i——第 i 类水质样本的个数；T_i——第 i 类水质样本中成功预测的样本的个数。通过权衡局部预测率和总预测率来确定分类的条件。

2　实验结果及讨论

以太湖入湖的十几条河流水质的近千个影响因子的监测数据为依据进行处理，分别以 24，21，18，15，12，9 和 6 个因子作为特征向量，利用 PSO 算法优化参数，构建预测水质污染的分类器。不同因子作为特征向量时的预测效果见表 1。

表 1　不同因子作为特征向量时的预测效果　%

作为特征向量的不同因子	重度污染水质预测率	轻度污染水质预测率	全局预测率
24 个特征	44.07	88.85	73.31
21 个特征(去掉水温、Hg、Se)	31.07	89.14	68.71
18 个特征(再去掉 pH 值、Cu、F^-)	57.27	71.94	66.27
15 个特征(再去掉硫化物、电导率、BOD_5)	54.42	71.34	65.33
12 个特征(再去掉 COD、Zn、LAS)	55.13	64.78	61.28

续表 1　　%

作为特征向量的不同因子	重度污染水质预测率	轻度污染水质预测率	全局预测率
9 个特征(再去掉挥发酚、CN^-、Cr^{6+})	77.37	60.16	66.49
6 个特征(再去掉 DO、Pb、As)	68.11	64.01	65.52

以预测率的高低来判断影响因子集合与水质优劣的相关性。

当输入特征向量是 9 个因子(NH_3-N、TN、COD_{Mn}、TP、石油类、Cd、Pb、As、DO)时局部预测率和全局预测率都相对较好。因此,以这 9 个因子作为 SVM 的特征向量,利用 PSO 算法优化 SVM 的参数 c 和 g,结果见表 2。

表 2　9 个因子作为特征向量时的参数和对应预测率

参数 c	参数 g	全局预测率/%	重度污染水质预测率/%	轻度污染水质预测率/%
25.65	1.41	84.62	77.78	88.24
25.94	1.50	81.48	100.00	70.59
1.73	11.31	82.14	80.00	83.33
18.56	1.35	92.59	88.89	94.45
31.29	0.01	84.61	88.89	82.35
26.08	1.83	80.77	88.89	76.47
12.72	2.09	77.78	77.78	77.78

由表 2 可以看出,c 和 g 不同时,所得到的预测率也不同,全局预测率最高为 92.59%,污染水质预测率最高为 100%,未污染水质预测率最高为 94.45%,考虑到全局预测率和局部预测率的平衡,选择参数 $c = 18.56$,$g = 1.35$,对应的预测率为:全局预测率 92.59%,污染水质预测率 88.89%,未污染水质预测率 94.45%。

从预测率可知,构建的支持向量机分类器对轻度污染水质和重度污染水质都有较好的预测能力。该支持向量机分类器的最终决策函数由少数的支持向量确定,其优点是:①方法使用简单,具有较好的鲁棒性;②复杂度低,运行速度快;③具有较好的推广能力;④该方法不同于其他的机器学习算法,它需要的先验干预很少。因此该分类器适合于环境监测站对水质状况进行监测,并为实时作出预警提供有力支持。

3　结语

笔者将 PSO 和 SVM 算法结合,利用 PSO 优化了 SVM 的参数 c 和 g,接下来 SVM 利用优化的参数,分别以 24,21,18,15,12,9 和 6 个影响水质的因子为特征向量,分别构建了分类器。发现以不同影响因子个数作为特征向量时,预测率有很大差别,这表明,并不是影响因子越多,预测率越高,相反,只要几个关键因子就可以将重度污染水质和轻度污染水质分类,当然,当因子过少时,有的重要因子会丢失,同样会产生不高的预测率。因此通过 PSO 和 SVM 的混合算法,可以确定影响太湖入湖河流水质的主要因子,利用这些主要因子对水质进行预测预警,不但可以节省时间,而且可以得到精确的结果。在以后的河流监控过程中,也只需要对这些主要因子进行重点监控即可。

[参考文献]

[1] 刘德林,刘贤赵. 主成分分析在河流水质综合评价中的应用[J]. 水土保持研究,2006,13(3):124-128.

[2] 伊元荣,海米提·依米提,王涛,等. 主成分分析法在城市河流水质评价中的应用[J]. 干旱区研究,2008,25(4):497-501.

[3] 左一鸣,崔广柏,顾令宇. 太湖水质指标因子分析[J]. 辽宁工程技术大学学报:自然科学版,2006,25(2):312-314.

[4] 鲁斐,李磊. 主成分分析法在辽河水质评价中的应用[J]. 水利科技与经济,2006,10(10):660-662.

[5] 王晓鹏. 河流水质综合评价之主成分分析方法[J]. 数理统计与管理,2000,31(3):49-52.

[6] 刘春燕. 深圳市河流水质评价指标筛选方案探讨[J]. 干旱环境监测,2010,24(1):47-50.

[7] KENNEDY J, EBERHART R. Particle Swarm Optimization [C]. Proceedings of IEEE International Conference on Neural Networks. IV. 1995:1942-1948.

[8] SHEN Q, MEI Z, YE B X. Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification[J] . Computers in Biology and Medicine, 2009, 39(7):646-649.

[9] PEDERSEN M E H, CHIPPERFIELD A J. Simplifying particle swarm optimization[J]. Applied Soft Computing, 2010, 10(2): 618-628.

[10] VAPNIK V. The nature of statistical learning theory[M]. New York: Springer, 1995.

[11] HUA S, SUN Z. Support vector machine approach for protein subcellular localization prediction[J]. Bioinformatics, 2001, 17(8): 721-728.

[12] PARK K J, KANEHISA M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs[J]. Bioinformatics, 2003, 19 (13): 1656-1663.