支持向量机在吉林西部地下水质评价中的应用

姜振蛟,卞建民,查恩爽

(吉林大学环境与资源学院,吉林 长春　130026)

摘要:支持向量机(SVM)是基于统计学理论的线性分类器,通过寻找最优分类超平面,实现结构风险最小。对于非线性问题,引入核函数,在高维空间中构造较低 VC 维的函数集,获得较好的学习效果和泛化能力。针对以往水质评价方法存在的问题,将 SVM 方法应用于水质评价中,结合研究区的水环境特征,选择 TDS、Mn、总铁、氟、总砷、总硬度、NO_3^-、Cl^-、SO_4^{2-} 为评价因子,并对数据进行极差化处理,通过 Libsvm 软件包求模型惩罚系数 C 并实现核函数的选择,采用 Matlab 编程建立支持向量训练器和学习机,对不同层位的地下水水质状况进行评价,分析评价结果,为该区饮水安全规划提供科学依据;并将评价结果与综合水质评价法和模糊综合评判法进行比较,探讨该方法的优越性和推广前景。

关键词:地下水环境特征;支持向量机;水质评价

中图分类号:X824　　**文献标识码**:A　　**文章编号**:1004-6933(2009)S1-0089-04

在我国的三北地区,原生环境的恶劣导致地下水水质较差,由水质而引发的地方病严重影响了人们的生活质量。合理评价地方病严重区地下水水质则成为当前水资源利用的一个首要问题。随着计算机技术的迅速发展,基于数学模型的水质评价方法不断涌现,而支持向量机算法(SVM)以最小结构风险代替了传统的经验风险,通过求解二次型寻优问题,获得全局的最优解,解决了传统神经网络法无法避免的局部极值问题[1-2];此外,支持向量机的拓扑结构由支持向量所决定,避免了传统神经网络拓扑结构需要经验试算的缺陷,在解决小样本、非线性及高维模式识别问题中具有一定的优势。有关 SVM 算法已在模式识别、回归估计、概率密度函数估计等方面获得了广泛的应用,其出色的学习性能被认为是人工神经网络方法的替代方法。笔者通过对多种水质评价方法的对比分析,采用 SVM 分类算法,以吉林西部高氟、高砷区地下水水质为例,建立了水质评价的支持向量机模型,并运用 Matlab 编程实现了改进的算法。

1　区域环境概况

吉林西部位于松嫩平原的西南部,属北温带半干旱大陆性气候区,地势东、南、西三面高,北部和中部较低,地形似簸箕状[3]。研究区下伏一个巨大的含水系统,埋藏有多个含水层,包括第四系孔隙潜水和承压水含水层、新第三系大安组、泰康组孔隙裂隙含水层和白垩系下统及上统裂隙孔隙含水层,相应的埋藏有多层地下水。

研究区大部分地段矿化度小于 1 g/L,仅局部地段矿化度较大超过 3 g/L。由西部山前倾斜平原到低平原边缘再到低平原中心,地下水的化学形成作用由溶滤作用为主逐渐变为以蒸发浓缩和离子吸附交换作用为主,相应的水化学类型也由 HCO_3·Ca、Mg 型水逐渐向 HCO_3·Na 型水转变,在接近盆地中部的低平原区,由于新第三纪和第四系白土山组地层沉积时的古气候干旱、寒冷,有利于沉积物及地下水中的苏打盐渍作用的进行,加之盆地中心地下水运动滞缓,赋存于该区域的地下水甚至出现苏打水。

受特定的沉积环境和强烈蒸发等因素的影响,潜层地下水氟质量浓度普遍较高,一般在 1.5 ~ 3.0 mg/L,部分湖沼洼地处大于 3.0 mg/L,最高达 10.0 ~ 28.0 mg/L,形成了潜水高氟、第三系承压水高砷的水化学场。研究区地下水砷、氟质量浓度严重超标,以砷超标为例进行统计具体结果见表 1。

基金项目:国家自然科学基金资助项目(40572170),国土资源调查项目(1212010634701)

作者简介:姜振蛟(1986—),辽宁普兰店人,男,硕士研究生,研究方向为水环境与水生态。E-mail:jiangzhenjiao@163.com

通讯作者:卞建民,教授。E-mail:bianjianmin@126.com

表 1　研究区不同井深,砷含量及样品检出率

取样地区	井深/m	样品数/个	$\rho(As)/(mg\cdot L^{-1})$ 区段	$\rho(As)/(mg\cdot L^{-1})$ 平均	超标数/个	检出率/%
乾安	0~40	39	0.00121~0.01642	0.005282	5	12.8
	40~100	5	0.00106~0.01036	0.0053	1	20.0
	>100	31	0.0008~0.02023	0.003828	3	9.7
通榆	0~40	34	0.00228~0.17888	0.035503	22	62.1
	40~100	31	0.00189~0.15241	0.054915	27	81.95
	100~150	11	0.00158~0.33904	0.07455	8	72.7
	>150	3	0.00204~0.00881	0.00507	0	0.0

针对研究区地下水环境特征,笔者采用支持向量机方法进行了水质评价,为保证研究区用水安全提供依据。

2　研究方法

支持向量机是 Vapnik 等[1]提出来的基于统计学理论的小样本学习方法,基本原理:二值分类问题通常用一个实值函数 $f:X\subseteq R^n\rightarrow R$ 按照这样的方式操作:当 $f(x)\geqslant 0$ 时,输入 $X=(x_1,x_2\cdots x_n)^T$ 赋给正类,否则赋给负类。在 $f(x)$(其中 $x\in X$)是线性函数的情况下,函数可以写成为:

$$f(x)=<w\cdot x>+b=\sum_{i=1}^{n}w_ix_i+b \tag{1}$$

式中:w_i 为权重系数;b 为偏置系数;$(w,b)\in R^n\times R$ 是控制 h 函数的参数,决策规则由 $\mathrm{sgn}(f(x))$ 给出,按照惯例 $\mathrm{sgn}(0)=1$。在这类假设中,$<w\cdot x>+b=0$ 被定义为超平面,将输入空间 X 分为两部分[1-2]。

在试验过程中,固定函数间隔等于 1(见式(2)、图 1)。

$$\begin{aligned}<w\cdot x^+>+b&=+1\\<w\cdot x^->+b&=-1\end{aligned} \tag{2}$$

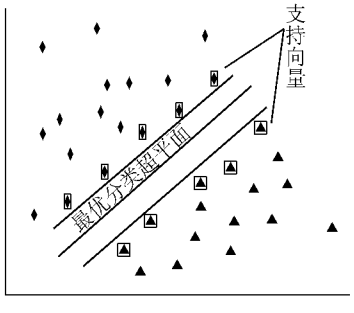

图 1　线性可分情况下的最优超平面及支持向量

将支持向量所在平面的函数间隔 γ 定义如下:

$$\gamma=\frac{1}{2}\left(<\frac{w}{\|w\|_2}\cdot x^+>-<\frac{w}{\|w\|_2}\cdot x^->\right)=\frac{1}{\|w\|_2} \tag{3}$$

其中 $\|w\|_2$ 为权重系数的 2 阶范数,因此,最好的分类效果对应着支持向量所在平面几何间隔的最大化,即权重向量的范数的最小化。

对于非线性问题,支持向量机理论中,引入核函数,把输入样本 X,通过映射 $\varphi:X\rightarrow H$ 将输入向量 X 映射到高维特征空间 H 中。当在特征空间中构造最优超平面时,训练算法仅使用空间中的点积,只要核函数 $k(x_i,x_j)=\varphi(x_i)\varphi(x_j)$ 满足 Mercer 条件,没有必要知道 φ 的形式,这种内积运算是可以用原空间中的函数来实现的,这一点提供了可能导致的"维数灾难"问题解决方法[2]。

非线性问题的二次优化的目标函数可写成

$$Q(\alpha)=\frac{1}{2}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_jy_iy_jK(x_i,y_i)=c^TI-\frac{1}{2}c^THc \tag{4}$$

式中:α_i 为拉格朗日算子;y_i 为对应于一组输入 X 的一次实现,$c=(\alpha_1,\alpha_2,\cdots,\alpha_n)^T$ $H_{ij}=y_iy_jK(x_i,x_j)$,并满足条件:$0<\alpha_i<C$(C 为错误分类的惩罚因子),$\sum_{i=1}^{n}\alpha_iy_i=0$ 对式(4)进行求解,得到训练集的最优分类函数为:

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{l}y_i\alpha_iK(x_i,x)+b\right) \tag{5}$$

对于最优分类平面中的偏置 b 则按下述公式进行推求:

$$b=\frac{1}{N_{sv}}\sum_{x_i^*\in S_V}\left(y_i-\sum_{x_i^*\in S_V}\alpha_iy_iK(x_i^*,x_j)\right) \tag{6}$$

式中:x_i^* 为支持向量;x 为未知向量;N_{sv} 为标准支持向量数;S_V 为支持向量集合。

式(5)称为支持向量机。

3　模型应用过程及结果讨论

3.1　评价因子及样本集数据处理

基于对研究区水质特征的研究,参照地下水环境质量标准,考察研究区多年水质评价过程,实验性地构造随机数据(这里用 Matlab 中的 Rand()函数来实现),选择合乎要求的数据作为训练样本。结合吉林西部高氟高砷地区的水质特点以及地下水水质标准,在水质评价中选择 TDS、Mn、总铁、F^-、总砷、总硬度、NO_3^-、Cl^-、SO_4^{2-} 为评价因子。

由于水质评价中各项指标的量级不同,必须在

评价之前进行数据的归一化处理，笔者采用极差化处理方法，其表达式为：

$$x'_i = (x_i - x_{\min}) / x_{\max} - x_{\min}$$

式中：$x_{\min}$、$x_{\max}$分别为各项指标实测数据中的最大值和最小值。

采用二叉树多值分类算法，分步进行水质评价，对于第 k 级水质评价过程中，使用第 k 个支持向量机，将 k 级训练样本取为正，大于 k 级的训练样本取为负值，以此类推，训练得到 5 级水质评价的四组支持向量，对待评价的样本进行水质评价。

3.2 评价模型

支持向量机水质评价按以下基本步骤进行。

a. 核函数及惩罚系数 C 的选择：

笔者使用 LibSVM 软件包[5-7]来实现核函数、惩罚系数 C 等的选择，参数选择时首先将训练数据写成如下格式：

$-1 \quad 1\ 0.0473 \quad 2\ 0.1282 \cdots\cdots 9\ 0.3820$

$+1 \quad 1\ 0.0798 \quad 2\ 0.0054 \cdots\cdots 9\ 0.9714$

第一列表示评价对象的类别，第二列及以后类似列表示评价因子编号，第三列则表示经过极差化处理的评价因子的值。DOS 下运行 svmtrain data data.model 和 svmpredict data data.model data.out 后，通过 svmtrain-c n1-t n2 data.model 进行调参。固定 t 后，改变 c 得到图 2 类似的关系曲线。

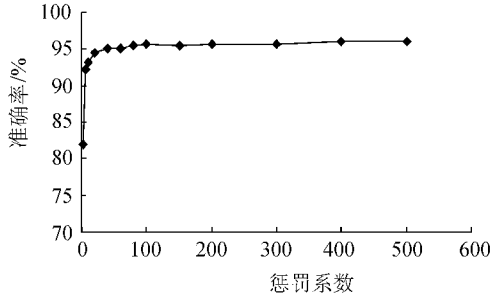

图 2 任意核函数下，惩罚系数与准确率的关系

结合图 2，选择惩罚系数在 100 ~ 600，与核函数组合，训练结果的准确率见表 2。吻合率最高为 96.8661%，对应的惩罚系数为 500 核函数选择为高斯核函数，即：

$$K(x_i, x) = \exp(-\gamma \| x_i - x \|^2), \gamma = 1$$

b. 将训练样本 x_i 和所对应的分类值 y_i 带入式(4)，并求解最大值。求出非零并小于 C 的拉格朗日算子对应的样本作为支持向量。

c. 将支持向量重新组合并记录，同样使用高斯核函数，将待预测的数据输入评价模型中，得到评价结果。

3.3 模型应用及结果分析

以 Matlab 为平台，建立支持向量训练机和水质评价程序，以研究区 2007 年测试数据为例进行实证研究，程序运行过程中各分类器所实现的最大分类间隔值及分类值见表 3。对研究区水质进行评价，评价结果见表 4。从表 4 可以看出，本区地下水水质由于受高氟、高砷的影响，普遍较差，潜水Ⅳ级水占全部评价点的 55%，第四系承压水中Ⅳ级水占全部评价点的 36.36%，特别值得关注的是本区第三系地下水一直被认为是研究区的后备水源，在本次采样评价结果中Ⅳ级水仅占全部评价点的 50%，应引起有关部门的关注。

表 3 支持向量机运行过程及三种方法的水质评价部分结果比较(共 60 个点)

支持向量机法模拟过程				结果对比		
模型机-1	模型机-2	模型机-3	模型机-4	支持向量机法	水质综合评价法	模糊综合评判法
间隔 = 0.0274	间隔 = 0.0245	间隔 = 0.0408	间隔 = 0.3857			
-1	-1	-1	1	4	4	4
-1	-1	1		3	4	3
-1	-1	-1	1	4	4	4
-1	-1	1	1	4	4	4
-1	-1	1		3	4	4
-1	-1	-1		3	4	3
-1	-1	-1	1	4	4	4
-1	-1	-1	1	4	4	3
-1	-1	-1	1	4	4	4
-1	-1	1	1	4	3	4
-1	-1	1		3	4	3
1				1	3	1
-1	-1	1		3	4	3
-1	-1	1		3	3	4
-1	-1	1		3	3	4
-1	-1	-1	1	4	4	4
-1	-1	-1	1	3	4	4
-1	-1	-1	1	4	4	4
-1	-1	1		3	4	3
-1	-1	-1	1	4	2	4

表 4 SVM 模型运行结果

层位	Ⅰ级	Ⅲ级	Ⅳ级	超标(Ⅳ)水点所占的比例/%
潜水	4	5	10	55
第四系承压水	1	6	2	36.36
第三系承压水	2	4	6	50

表 2 吻合率与惩罚系数及核函数的关系

	函数名称	100	200	300	400	500	600
0	线性核函数	95.6205	95.7265	95.7265	96.0114	96.0114	95.7265
1	多阶多项式核函数	55.8405	55.8405	78.3476	64.1026	52.8291	59.5442
2	高斯核函数	95.7265	96.5812	96.2963	96.2963	96.8661	96.5812
3	sigmoid 核函数	95.1567	95.4416	95.4416	95.7265	95.7265	95.7265

为了说明支持向量机评价结果的合理性以及这种方法在水质评价中应用的可行性,将其与模糊综合评判法及综合水质评价法进行了比较(表 3)。

从表 3 的结果对比可知,支持向量机的评价结果与综合水质评价结果有 67% 相同,与模糊综合评判法有 80% 相同。综合分析原因如下:

综合水质评价法是根据地下水水质标准来确定水质评价指标的,对检测项目个数要求较全。若检测项目偏少,会严重影响评价结果。而实际情况下,部分离子或化合物要求较高的检测技术和费用;对部分地区地下水的研究,无需严格按照水质标准来测试多项指标浓度。因此,综合水质评价法受到了限制。

模糊综合评判法[7]是以隶属度来描述模糊的水质分级界限的,能较好地刻画环境质量客观存在的状况。但是,Ⅰ类和Ⅴ类的隶属度为 1 时,所控制的范围过大,会造成评价结果趋向两端;在给定污染因子权重时,并没有理论上的依据;对于浓度很小,但对水质评价结果影响很大的因子考虑不多。

支持向量机方法巧妙地运用了结构风险最小化的理论,通过寻找最优超平面实现最大分类间隔,解决了水质评价过程中的样本数量少、影响因素复杂、评价指标有限等问题。在评价过程中,只要已知水质标准,可用 Matlab 中的 rand()函数随机地产生足够的训练数据,训练出支持向量后,对评价的数据进行评价。即使只监测了两种离子,也只需根据地下水水质标准,随机地产生这两种离子在各等级限值范围内的一系列数据作为训练样本,通过给定的支持向量学习机,训练得到 n 个二维的支持向量,利用式(4)所示的模型来评价水质。

4 结 论

a. 笔者建立了地下水水质量评价的支持向量机模型,基于 Matlab 程序设计,对研究区 2007 采样点进行了水质评价,评价结果表明,该区地下水水质普遍较差,特别是以往被作为备用水源地的第三系承压含水超标也比较严重。

b. 将支持向量机模型的评价结果与综合水质评价、模糊综合评价方法进行了应用对比,结果有 67% 相同,与模糊综合评判法有 80% 相同。对比分析表明,支持向量机通过寻找最优超平面实现结构风险最小,引入核函数来避免维数灾难,较好地解决了其他数学方法在水质评价中遇到的小样本、非线性、过学习、高维数、易产生局部极小点等实际问题,可应用于水质评价中且具有较好的推广前景。

通过本次评价也发现,在利用支持向量机进行训练的过程中,对训练数据的选择没有统一的规则,而且还存在一定的随机性,这对最后的训练结果产生一定的影响,有待进一步改进。

参考文献:

[1] VAPNIK V. Statistical learning theory[M]. New York:John Wiley & Sons, 1998:7-10.

[2] BURGESS C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2):121-167.

[3] 卞建民,林年丰,汤洁.吉林西部向海湿地环境退化及驱动机制研究[J].吉林大学学报:地球科学版,2004,34(3):441-444,458.

[4] 卢敏.基于支持向量机的水资源可持续利用评价[J].水电能源学报,2000,23(5):18-22.

[5] 郑一华.基于支持向量机的水质评价与预测研究[D].南京:河海大学,2006.

[6] 李亚伟,陈守煜,韩小军. 基于支持向量机 SVR 的黄河凌汛预报方法[J].大连理工大学学报,2006,46(2):272-275.

[7] 朱雷,陈威.模糊综合评判法在水质评价中的应用[J].武汉理工大学学报,2001(8):63-67.

(收稿日期:2008-10-20 编辑:徐娟)