

改进的决策树支持向量机地下水水质评价

陈海洋,滕彦国,王金生

(北京师范大学 水科学研究院,北京 100875)

(chen.haiyang@hotmail.com)

摘要:基于结构风险最小原理的支持向量机(SVM)具有较强的学习泛化能力和良好的分类性能,能用来解决少样本学习的二类模式识别问题。针对具备多级类别的地下水水质评价问题,可以采用决策树 SVM 分类方法,通过对多类别水质标准的重新组合以构建类似于决策树的多个子分类器来实现。但基于决策树 SVM 分类过程中常常会出现由于正负类训练样本数据不均一导致的局部识别误差。基于二叉树原理提出了一种改进决策树 SVM 模型,通过加密数据插值和二叉分类有效避免正负类训练样本数据不均一的问题,针对地下水水质评价特点,增加了第5个子分类器以精确识别Ⅱ类水质和Ⅲ类水质。实验结果表明,改进的决策树 SVM 分类模型评价结果稳定。

关键词:支持向量机;决策树支持向量机;地下水;水质评价

中图分类号:P641.2; TP18 **文献标志码:**A

Groundwater quality evaluation based on optimized model of decision-tree-based support vector machine

CHEN Hai-yang, TENG Yan-guo, WANG Jin-sheng

(College of Water Science, Beijing Normal University, Beijing 100875, China)

Abstract: Support Vector Machine (SVM) based on the minimum of structured risk is characterized with strong ability to learn and predict and favorable classification performance, which makes it able to solve the two types of pattern recognition of fewer sample learning. In order to evaluate the groundwater quality which has five classes with SVM, the decision-tree-based way of rebuilding the classes like decision tree to create more sub two-class SVM would be used. But as a solution of classifying more classes, some defaults exist in decision-tree-based support vector machine (DTBSVM) including the local error produced by different sample mount between two classes. The authors brought forward an optimized DTBSVM model based on the principle of two cross tree to realize the evaluation for groundwater quality. The experimental results show that the optimized DTBSVM model is a good way to evaluate the groundwater quality.

Key words: Support Vector Machine (SVM); Decision-Tree-Based SVM (DTBSVM); groundwater; water quality evaluation

0 引言

地下水水质评价就是把地下水水质检测样本指标值与相应的地下水水质评价标准进行比较,通过一定的数学模型,确定检测样本的等级。地下水水质评价的目的是通过准确判断地下水水质的污染等级,为地下水资源合理开发利用和地下水污染综合防治提供科学依据。地下水水质评价常见方法有污染指数法、模糊综合评价法、灰色系统评价法、层次分析法、物元分析法、人工神经网络评价法等。这些传统的方法均存在一定的不足,如模糊综合评价法等大多数方法需要设计各评价指标对各级标准的隶属函数以及各指标权重,其评价结果容易受到人为主观因素的影响;灰色系统评价法中白化函数的选择和聚类权的确定往往因人而异,造成评价模式难以通用;而人工神经网络需要的样本数多,网络结构的优劣因人而异等。支持向量机(Support Vector Machine, SVM)是由统计学习理论发展起来的一种新型学习机器,它以结构风险最小化原理为理论基础,具有逼近复杂非线性系统、较强的学习

泛化能力和良好的分类性能,所需样本少、建模方便、计算简单、学习训练时间短、通用性强,可以用于解决属于模式识别的地下水水质评价问题^[1]。周兆永等人基于遗传算法(Genetic Algorithm, GA)优选参数建立了支持向量机水质评价模型^[2];王凯军等人运用多层次分类支持向量机对吉林省磐石市地下水水质进行了评价^[3];夏琼等人基于支持向量机对淮南市浅层地下水水质进行了评价^[4]。

但是,SVM是针对两类分类问题提出的,利用它能高效、精准实现两类问题的识别,而地下水水质有五级标准,如何构建合适的SVM分类器实现对地下水水质的准确评价是一个值得研究的问题。目前常用的SVM分类器构建方法有整体法、“一对余”分类法、“一对一”分类法和决策树分类法等。在这些方法中,决策树分类法符合SVM的二类别分类特点,能够较为准确地识别所有类别,是当前应用较为成熟的一种分类方法。但是,如果SVM的训练样本数少,且存在较大的不对称性,将导致局部识别异常问题,从而影响地下水水质的准确评价。本文在分析常用的决策树SVM分类法的基础上,

收稿日期:2010-09-21; **修回日期:**2010-11-17。 **基金项目:**国家水体污染控制与治理科技重大专项(2009ZX07419-003; 2008ZX07207-007);教育部新世纪优秀人才支持计划项目(NECT-09-0230)。

作者简介:陈海洋(1978-),男,江西石城人,博士研究生,主要研究方向:流域环境管理、环境管理信息化;滕彦国(1974-),男,黑龙江巴彦人,教授,博士生导师,主要研究方向:环境地球化学、水文地球化学;王金生(1956-),男,河南太康人,教授,博士生导师,主要研究方向:地下水数值模拟、地下水资源评价。

提出一种改进的决策树支持向量机地下水水质评价模型,并用该地下水水质评价模型对某市地下浅层潜水水质进行了评价。

1 材料和方法

1.1 训练样本数据及样品水质数据

1) 样品水质数据。选用文献[5]中的 11 份水质化验数据作为本次研究的评价目标集。选取其中的总硬度、TDS、氯化物、硫酸盐、硝酸盐、氟化物 6 项主要影响当地水质的污染物指标作为评价指标,其实测值如表 1^[3]。

表 1 某市地下浅层潜水水质数据 $\text{mg} \cdot \text{L}^{-1}$

水样号	总硬度	TDS	氯化物	硫酸盐	硝酸盐	氟化物
1	382.84	669.39	16.63	5	15.52	0.11
2	415.37	718.22	14.00	4	4.12	0.25
3	455.41	714.23	72.61	6	6.86	0.13
4	495.45	827.53	3.51	5	8.52	0.28
5	352.82	596.74	6.98	5	5.32	0.20
6	322.79	560.92	5.25	5	14.36	0.20
7	578.02	949.30	42.01	18	11.80	0.10
8	600.54	1008.76	6.10	16	7.66	0.10
9	555.50	872.17	87.50	80	223.00	0.18
10	457.91	697.61	145.25	10	125.20	0.10
11	495.45	820.69	80.51	45	80.40	0.30

2) 训练样本数据。本次研究选取《地下水质量标准》(GB/T14848—93)中与样本水质数据对应的污染物指标作为训练样本数据属性,取各级标准的上限值,通过插值方法产生训练样本数据。由于地下水质量标准包含 I ~ V 级,对第 V 级标准只有下限要求而没有上限要求,考虑到本次评价水质数据从总体上看没有严重超标的因子,故以级差取值法定义 V 级标准上限,即:

$$M_V = M_{IV} + (M_{IV} - M_{III})$$

其中 $M_i (i = 1, 2, 3, 4, 5)$ 为第 i 级标准的上限。选取结果如表 2 所示。

表 2 地下水水质评价基准表 $\text{mg} \cdot \text{L}^{-1}$

污染物	上限值				
	I 级	II 级	III 级	IV 级	V 级
总硬度	150	300	450	550	650
溶解性总固体	300	500	1000	2000	3500
氯化物	50	150	250	350	450
硫酸盐	50	150	250	350	450
硝酸盐	2	5	20	30	40
氟化物	1	1	1	2	3

尽管 SVM 支持小样本训练,但 I ~ V 共五个训练样本数据难以达到模型训练精度要求,为此采取随机插值方法生成训练样本。结合地下水水质评价标准,对 I 级水质,选取 0 值作为下线,选取水质标准中 I 级作为上线,内插 30 个样本集;对 II 级水质,选取水质标准中 II 级值作为下线,选取水质标准中 III 级值作为上线,内插 30 个样本集;依此类推,在 III 级、IV 级和 V 级上各内插 30 个样本集。共产生 150 个训练样本数据。

1.2 研究方法

1.2.1 SVM 分类基本原理

本次研究基于支持向量机实现地下水水质的评价。支持向量机基于结构风险最小原理,利用最大化分类边界的思想

寻求最优超平面解决线性可分情况下的模式识别问题,而对于线性不可分情况,则又通过使用非线性映射算法将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分,从而在高维特征空间采用线性分析方法进行问题求解^[6]。现对其基本原理做简要描述。设在 D 维空间有训练集 $\{X_i, Y_i\} (i = 1, 2, \dots, k, k \text{ 为样品数}; X_i \in \mathbf{R}^d, Y_i \in \{-1, 1\})$, 即如果 X_i 属于第 1 类,则 $Y_i = 1$, 如果 X_i 属于第 2 类,则 $Y_i = -1$ 。支持向量机就是要寻找一个满足要求的分割平面 $\mathbf{w} \cdot \mathbf{x} + b = 0$ (\mathbf{w} 为分割平面的法向量, b 为分割平面的偏移量), 既能把训练集中的数据正确分类, 又能使分类后的数据集的间隔尽可能最大。用数学模型描述如下:

$$\min \left(\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^k \xi_i \right) \quad (1)$$

$$\text{s. t. } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i; i = 1, 2, \dots, k, \xi_i \geq 0$$

其中 C 为惩罚因子,它控制对错分样本惩罚的程度。这是一个典型的二次凸规划问题,通过构建拉格朗日函数,可得该目标函数的对偶形式:

$$\max Q(\alpha) = \sum_{i=1}^k \alpha_i - \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \quad (2)$$

$$\text{s. t. } \sum_{i=1}^k y_i \alpha_i = 0; 0 \leq \alpha_i \leq C, i = 1, 2, \dots, k$$

对于非线性 SVM,则可根据泛函的有关理论,设计一种满足 Mercer 条件的核函数,将原始线性不可分样品数据映射到某一高维特征空间,在特征空间运用内积函数实现线性不可分数据集的分类。假设利用核函数 $K(\mathbf{x}_i, \mathbf{y}_j) = (\Phi(\mathbf{x}_i), \Phi(\mathbf{y}_j))$ 把原始数据映射到高维特征空间,那么特征空间的核函数 SVM 为:

$$\max Q(\alpha) = \sum_{i=1}^k \alpha_i - \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i \cdot \mathbf{x}_j) \quad (3)$$

$$\text{s. t. } \sum_{i=1}^k y_i \alpha_i = 0; 0 \leq \alpha_i \leq C, i = 1, 2, \dots, k$$

通过对式(2)或式(3)问题的求解,可分别得到各自的最终判别函数为:

$$f(\mathbf{x}) = \text{sgn} \left\{ \sum_{i=1}^k \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}) + b^* \right\} \quad (4)$$

$$f(\mathbf{x}) = \text{sgn} \left\{ \sum_{i=1}^n \alpha_i^* y_i K(\mathbf{x}_i \cdot \mathbf{x}) + b^* \right\} \quad (5)$$

1.2.2 基于决策树的 SVM 多类别分类方法

决策树 SVM 分类方法是一种应用较为广泛的 SVM 多类别分类方法,它结合了 SVM 的二类别分类特点,将多类别分类的各个类别重新组合,构建类似于决策树的多个子分类器实现多类别的分类,能够较为准确地识别所有类别^[7]。对于 k 类分类问题,决策树 SVM 分类法只需构造 $k-1$ 个 SVM 子分类器,以地下水水质评价为例,构建方法见图 1(简称为 α 模型)。从该图可以看出,要分类出具有 5 个级别的地下水水质,只需要构造 4 个 SVM 子分类器,进行评价时,只要从根节点开始计算决策函数,根据值的正负决定下一节点,如此下去,直到到达某一叶子节点为止,此叶子节点所代表的类别就是测试样本的所属级别^[8]。

1.2.3 改进的决策树 SVM 多类别分类方法

以图 1 所示的决策树实现地下水水质评价时,将由于样本数据的不均一导致识别误差。在训练分类器 1 时,属于正类的 X_i 与属于负类的 X_j 样本数相差较大,从而导致识别精度降低。同样,在训练分类器 2、分类器 3 时也存在同样的问题。为解决训练样本的均一性所带来的识别误差问题,有关研究提

出在惩罚因子上增加一个样本权重来解决,即在目标函数的惩罚因子前增加一项 m_i 作为样本权值,数学表达如下:

$$\min \left(\frac{1}{2} \| \mathbf{w} \|^2 + m_i C \sum_{i=1}^k \xi_i \right)$$

$$\text{s. t. } y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i; i = 1, 2, \dots, k, \xi_i \geq 0 \quad (6)$$

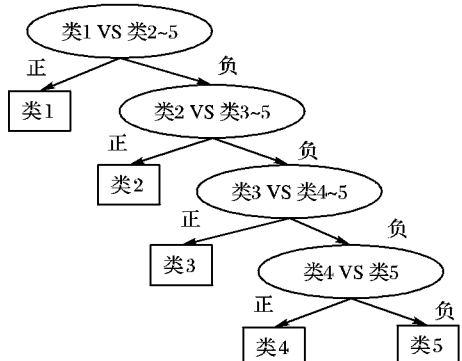


图1 决策树 SVM 地下水水质分类模型——α 模型

使用该方法,可以有效平衡训练样本的不均一性问题。但 m_i 的选择缺乏实质性的理论指导,只能是通过反复实验,人工选取参数,这不仅要求操作人员有丰富的实际经验,而且需要付出较高的时间代价。为此,本文提出两种改进方案以解决样本不均一所带来的识别误差问题。

1) 样本加密插值法。即对数量偏少的正类样本加密插值,使之与负类样本基本保持均等。以图1中的分类器1为例,处于正类的样本数为30个,处于负类的样本数则有多达120个,样本极不均一,为此,对正类样本加密插值到120个,最后以加密插值后的样本进行训练。同样,对分类器2、分类器3也采用同样的方法。

2) 改进决策树分类模型。作为α模型的改进,建立图2所示的类二叉树决策树分类模型(简称为β模型),该模型基于二叉树原理,在α模型的基础上做了三个方面的改进:首先使各子分类器识别数量均等的正负类样本,以保证识别精度,减少识别误差;其次,在使用分类器3辨识作为正值的类3和作为负值的类4、类5时,对类3的样本数据进行加密插值,使其数量等于类4、类5数量之和;第三,增加第5个子分类器,用来解决类3经过第1分类器后部分样本偏差问题。

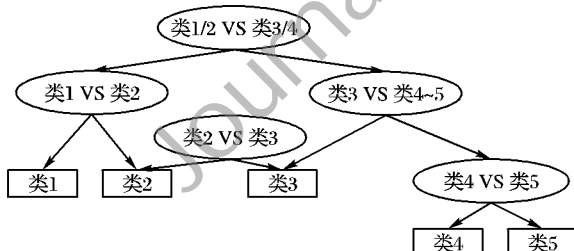


图2 改进的决策树 SVM 地下水水质分类模型——β 模型

2 评价结果与讨论

2.1 基于决策树支持向量机地下水水质评价结果

为比较正负样本均一性对水质评价结果的影响,本文首先选用α模型作为水质评价多分类模型,以未经过加密插值的水质标准样本数据进行训练,并用训练好的模型对样品水质数据进行评价。作为比较,再以加密插值的水质标准样本数据进行训练并进行样品水质数据评价。同样,选用改进后的β模型做同样的训练和评价。数据插值样本数及评价结果如表3、4所示。其中:1) 污染损失率指数法、模糊综合评价

法、F-法的评价结果来自文献[5];2) 基于决策树 SVM 评价采用 Matlab 7.9.01 编制程序,在 CPU 为 P8700 2.53 GHz、内存为 2.5 GB 的微型计算机上运行通过。

表3 基于决策树 SVM 的地下水水质评价训练数据表

子分类器	α 模型				β 模型			
	未加密插值训练	加密插值训练	未加密插值训练	加密插值训练	未加密插值训练	加密插值训练	未加密插值训练	加密插值训练
I	正类 30	负类 120	正类 120	负类 120	正类 60	负类 60	正类 60	负类 60
II	30	90	90	90	30	30	30	30
III	30	60	60	60	30	60	60	60
IV	30	30	30	30	30	30	30	30
V	—	—	—	—	30	30	30	30

表4 基于决策树 SVM 的某市地下浅层潜水水质评价结果

测点	污染损失率指数法	模糊综合评价法	F-法	α 模型评价结果		β 模型评价结果	
				未加密插值训练	加密插值训练	未加密插值训练	加密插值训练
1	III	II	II	II	III	III	III
2	II	III	II	II	III	III	III
3	III	II	IV	III	III	II	II
4	III	II	IV	III	III	III	III
5	II	II	II	II	II	III	III
6	II	II	II	II	III	III	III
7	III	IV	IV	IV	IV	IV	IV
8	III	IV	IV	IV	IV	IV	IV
9	IV	V	IV	V	V	V	V
10	IV	IV	IV	IV	IV	IV	IV
11	IV	IV	IV	IV	IV	IV	IV

2.2 正负样本均一性对评价结果影响分析

从评价结果可以看出,不采用加密插值训练的α模型评价结果与模糊综合评价法相对吻合,比污染损失率指数法和F-法结果偏低。而经过加密插值训练的α模型消除了正负样本不均一带来的偏差问题,整体评价结果与β模型评价结果极为吻合。我们知道,地下水水质评价指标共有三十几项,且各指标之间存在复杂的非线性关系,虽然 SVM 可以从小样本中寻求表征样品分类特征的支持向量以保证其良好的泛化能力,但是,基于二类识别的 SVM 要求正负训练样本数保持合适比例,过少的正类样本将导致“欠学习”现象,而过少的负类样本将导致“过学习”现象,这都影响模型的推广能力。以上实验表明,通过对过少的正类样本或负类样本进行加密插值,可以优化边缘样本的训练性能,获得较为良好的识别效果。

2.3 改进的决策树 SVM 评价模型评价效果分析

基于α模型的地下水水质评价,由于正负样本不够均一,使得评价结果容易出现局部偏差,特别是对于边缘样本,容易引发局部识别偏低的现象。对过少的正类样本或负类样本采取加密插值可以优化识别效果,减少局部偏差,但也会出现细粒度数据的异泛化问题。β模型是从 SVM 二类识别的本质出发建立的基于二叉树多类别分类方法,既能避免正负类样本数不均一的问题,也能减少分类周期,提高评价效率。针对特定的地下水水质评价问题,特别增加的子分类器5可以进一步识别出Ⅱ类水质和Ⅲ类水质,避免部分处于Ⅱ类水质和Ⅲ类水质中间,而又接近Ⅲ类水质的样品被识别为Ⅱ类水质,进一步提高了识别精准度。通过反复实验,利用β模型的交叉验

(下转第855页)

并将矩阵分段读取,因而提高了改进实现的速度。当矩阵规模为 3000 时 GPU 改进实现的加速比达到 9.2。

对于 Laplace 算法,本文同样测试了在不同矩阵规模下 CPU 程序、GPU 基本实现和 GPU 优化实现的性能对比。结果如表 2 和图 9 所示。从图中 9 可以看出,Laplace 算法的加速比和 LU 分解具有相同的变化趋势,即随着矩阵规模的增大而增大。同样在矩阵规模较小的情况下 GPU 程序性能不如 CPU 程序,这是由于计算量较小时,数据拷贝的开销相对较大,占据了 GPU 程序的运行时间。改进后的 GPU 程序与基本实现相比主要体现在消除了分支语句,从而使得线程执行时效率提高。

表 2 Laplace 算法的执行时间

实现方式	矩阵规模 N				
	256	512	1024	2000	3000
CPU 实现	0.548	3.14	13.81	56.22	128.93
GPU 基本实现	0.612	1.22	3.25	10.53	18.23
GPU 改进实现	0.603	0.83	1.76	6.51	11.07

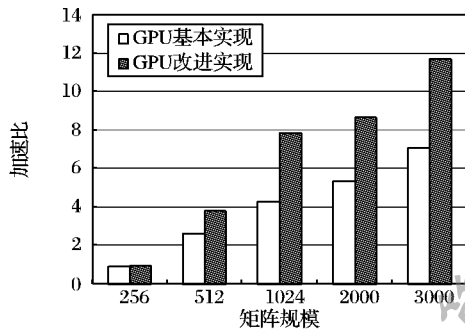


图 9 Laplace 算法加速比

5 结语

早期的 GPU 是为了分担 CPU 对图形图像处理的任务, GPU 适合的是计算密集型运算,且可并行计算,在算法相同的情况下,数据量越大,其处理效率越高。针对如何在 GPU

上实现 LU 分解和 Laplace 算法,本文提出了 CPU 与 GPU 任务划分,对 GPU 程序进行分支消除,使用共享存储器和矩阵分段读取的方法,优化 GPU 程序。最后实验结果表明,GPU 优化实现算法随着矩阵规模的增大加速比也同时增大。

参考文献:

- [1] OWENS J D, LUEBKE D, GOVINDARAJU N, *et al.* A survey of general purpose computation on graphics hardware [J]. Computer Graphics Forum, 2007, 26(1): 80-113.
- [2] Nvidia Corporation. NVIDIA CUDA programming guide V3.0 [EB/OL]. [2010-08-20]. <http://www.nvidia.com/cuda>.
- [3] LUEBKE D, HARRIS M, KRUGER J, *et al.* GPGPU: General-purpose computation on graphics hardware [C]// Proceedings of the 2006 ACM/IEEE Conference on Supercomputing. Washington, DC: IEEE Computer Society, 2006: 10-16.
- [4] HUSBANDS P, YELICK K. Multi-threading and one-sided communication in parallel LU factorization [C]// Proceedings of the 2007 ACM/IEEE Conference on Supercomputing. Washington, DC: IEEE Computer Society, 2007: 1-10.
- [5] DAVIES B, MARTIN B. Numerical inversion of the Laplace transform: A survey and comparison of methods [J]. Journal of Computational Physics, 1979, 3(1): 1-32.
- [6] MANAVSKI S A. CUDA compatible GPU as an efficient hardware accelerator for AES cryptography [C]// Proceedings of IEEE International Conference on Signal Processing and Communication. Washington, DC: IEEE Computer Society, 2007: 65-68.
- [7] SZERWINSKI R, GUNEYSU T. Exploiting the power of GPUs for asymmetric cryptography [C]// CHES 2008: Proceedings of the 10th International Workshop on Cryptographic Hardware and Embedded Systems, LNCS 5154. Berlin: Springer, 2008: 79-99.
- [8] 张舒,褚艳利. GPU 高性能运算之 CUDA[M]. 北京:中国水利水电出版社,2009.
- [9] DONGARRA J, HAMMARLING S, WALKER D. Key concepts for parallel out-of-core LU factorization [J]. Computers and Mathematics with Applications, 1998, 35(7): 13-31.

(上接第 850 页)

证、随机插值验证准确率均达到 100%,对文献[5]中的某市地下浅层潜水水质评价结果非常稳定。

3 结语

1) 基于结构风险最小原理的支持向量机具有较强的学习泛化能力和良好的分类性能,具有所需样本少、建模方便、计算简单、学习训练时间短、通用性强等特点,可以用于解决属于模式识别的地下水水质评价问题。

2) 决策树 SVM 分类方法是一种应用较为成熟的 SVM 多类别分类方法,它结合了 SVM 的二类别分类特点,通过对多类别的重新组合以构建类似于决策树的多个子分类器实现多类别的准确识别,可以有效应用于具有多级水质类别的地下水水质评价。

3) 鉴于决策树 SVM 分类过程中正负类训练样本数据不均一的问题,本文基于二叉树原理提出了一种改进决策树 SVM 模型,通过加密数据插值和二叉分类有效避免正负类训练样本数据不均一的问题,还针对地下水水质评价特点增加了第 5 个子分类器以精确识别第 II 类水质和 III 类水质。实验

结果表明,改进的决策树 SVM 分类模型评价结果稳定。

参考文献:

- [1] 祁亨年. 支持向量机及其应用研究综述[J]. 计算机工程, 2004, 30(10): 6-8.
- [2] 周兆永,汪西莉,曹艳龙. 基于 GA 优选参数的 SVM 水质评价方法研究[J]. 计算机工程与应用, 2008, 44(4): 190-193.
- [3] 王凯军,曹剑峰,李升. 多层次分类支持向量机在水质评价中的应用[J]. 水资源保护, 2009, 25(2): 30-33.
- [4] 夏琼,钱家忠,陈舟. 基于支持向量机的淮南市浅层地下水水质评价[J]. 水文地质工程地质, 2009, 36(1): 56-59.
- [5] 邓颂霖,梁秀娟,肖长来,等. 基于 GA 优化的地下水水质评价法[J]. 东北水利水电, 2009, 27(9): 55-57.
- [6] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42.
- [7] 厉小润,赵光宙,赵辽英. 决策树支持向量机多分类器设计的向量投影法[J]. 控制与决策, 2008, 28(7): 745-747.
- [8] WANG X D, SHI Z W, WU C M, *et al.* An improved algorithm for decision-tree-based SVM [C]// Proceedings of the 6th World Congress on Intelligent Control and Automation. Dalian: [s. n.], 2006: 4234-4237.