

多层次分类支持向量机在水质评价中的应用

王凯军^{1,2}, 曹剑峰¹, 李 升¹, 曹东平³

(1. 吉林大学环境与资源学院, 吉林 长春 130026 ; 2. 吉林省水文水资源局, 吉林 长春 130022 ; 3. 中国石化集团上海海洋石油局, 上海 200120)

摘要 在介绍支持向量机(SVM)原理的基础上,设计采用多层次分类器改进支持向量机的分类算法,并利用该方法评价磐石市的地下水水质,同时与模糊综合评判、灰色聚类 and 神经网络评价结果对比。结果表明,该方法可以用于水资源学科的分类评价计算。

关键词 支持向量机 水质评价 多层次分类器

中图分类号 TP641.7 ;O234 **文献标识码** :A **文章编号** :1004-6933(2009)02-0030-03

Application of multiple classified support vector machine in water quality assessment

WANG Kai-jun^{1, 2}, CAO Jian-feng¹, LI Sheng¹, CAO Dong-ping³

(1. College of Environment and Resources, Jilin University, Changchun 130026, China ; 2. Hydrology and Water Resources Bureau of Jilin Province, Changchun 130022, China ; 3. SINOPEC Shanghai Offshore Petroleum Bureau, Shanghai 200120, China)

Abstract Based on the principle of support vector machine (SVM), an improved SVM method combined with a multiple classifier was developed. The improved SVM method was applied to water quality assessment in Panshi City and compared with the results of the neural network method, fuzzy mathematic method, and grey clustering method. The results show that the method is reliable in the classification and evaluation of water resources.

Key words support vector machine ; water quality assessment ; multiple classifier

磐石市位于吉林省中南部,吉林市南部,主要地下水化学类型为 $\text{Cl}-\text{HCO}_3$ 型、 HCO_3-Cl 型和 HCO_3-SO_4 型。近年来随着工农业经济的持续发展,工业规模不断扩大,其水质也受到一定程度的污染,因此,对磐石地区的地下水水质进行评价,可以为磐石市规划城市供水提供科学依据。

1 支持向量机(SVM)

支持向量机(support vector machines, SVM)是建立在统计学理论的 VC 维(vapnik chervonenkis)理论和结构风险最小原理理论基础之上的算法。支持向量机能尽量提高学习机的泛化能力,即使是由有限训练样本,在求解问题时仍能得到较为满意的解,它解决了在神经网络方法中无法避免的局部极值问题,有效避免“过拟合”,泛化能力强,在解决小样本、非线性及高维模式识别问题中表现出许多特有的优

势,使它成为一种优秀的学习算法。

支持向量机分类计算的原理是寻找一个满足分类要求的最优分类超平面,使得超平面在保证分类精度的同时,能够使超平面两侧的空白区域最大化^[1-3]。

支持向量机是从线性可分情况下的最优分类面发展而来的,基本思想可用图 1 所示的两维情况说明,具体的数学描述如下:

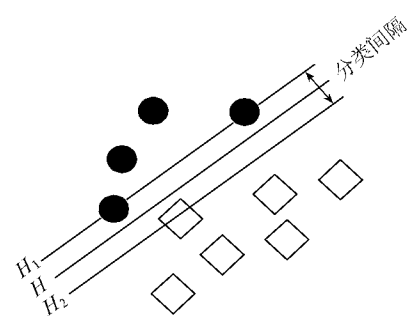


图 1 最优分类线示意图

作者简介 王凯军(1975—),男,吉林安图人,工程师,博士研究生,研究方向为地下水系统及水资源系统工程。E-mail :kaijunwang@126.com

如图 1 中所示,圆和正方形代表两类样本, H 为分类线; H_1, H_2 分别为过各类中离分离线最近的样本且平行于分类线的直线,它们之间的距离叫分类间隔。设线性可分的样本集有 n 个样本 (x_i, y_i) ,其中 $i = 1, 2, \dots, n$; $x \in R^d$; $y, x \in \{-1, 1\}$ 是类别符号。在高维空间中,将两类样本无错分开的分类超平面 H 满足 $g(x) = \omega \cdot x - b = 0$,进行归一化,可以使所有样本满足 $|g(x)| \geq 1$,所有样本被无错误分开时应满足

$$y(\omega \cdot x_i - b) - 1 \geq 0 \tag{1}$$

式中 ω 为可调权向量; b 为阈值。

此时, H_1 和 H_2 之上的点距分类线 H 的距离为 $\frac{1}{\|\omega\|}$,分类间隔为 $\frac{2}{\|\omega\|}$,要使分类间隔最大相当于使 $\|\omega\|^2$ 最小,这样求解最优超平面的问题就可以表示成如下的约束优化问题

$$\varphi(x) = \frac{1}{2} \|\omega\|^2 = \frac{1}{2} (\omega \cdot \omega) \tag{2}$$

满足上述条件的分类面就叫最优分类面, H_1 和 H_2 之上的点共同支持了最优分类面,就称为支持向量,这也是支持向量机名字的由来。

采用 Lagrange 乘子方法解决约束最优问题,可以将上述最优问题转化成为其对偶问题,解上述问题后得到的最优分类函数

$$f(x) = \text{sgn} \{ \omega \cdot x + b \} = \text{sgn} \left\{ \sum_{i=1}^n \alpha_i^* y_i (x_i \cdot x) + b^* \right\} \tag{3}$$

式中 $f(x) = \text{sgn}(x)$ 为符号函数; α_i^* 为与每个样本对应的 Lagrange 乘子, b^* 为分类阈值。

在线性不可分的模式情况下,不可能建立一个不具有分类误差的分离超平面,然而,希望找到一个最优超平面,使对整个训练集平均的分类误差的概率达到最小。因此,引入松弛变量 $\xi_i \geq 0, i = 1, \dots, n$,则式(1)变为

$$y(\omega \cdot x_i - b) - 1 + \xi_i \geq 0 \quad (i = 1, \dots, n) \tag{4}$$

可以将目标改为求

$$\varphi(\omega, \xi) = \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n \xi_i \tag{5}$$

即折中考虑最少错分样本和最大分类间隔,就得到了广义最优分类面。其中惩罚系数 C 是一个常数,它控制对错样本惩罚的程度, $C > 0, C$ 越大表明对错误的惩罚越重。

对于非线性问题,根据泛函的有关理论,采用满足 Mercer 条件的某种核函数 $K(x_i, x_j)$,无须知道变换的形式,就可以实现某一非线性变换后的线性分类,而不增加计算的复杂程度。相应的分类函数也

变为^[4-6]

$$f(x) = \text{sgn} \{ \omega \cdot x + b \} = \text{sgn} \left\{ \sum_{i=1}^n \alpha_i^* y_i K(x_i \cdot x) + b^* \right\} \tag{6}$$

目前 3 种常用的核函数分别为^[7-9]：

多项式核函数：

$$K(x_i, y_i) = [(x \cdot x_i) + 1]^q \tag{7}$$

径向基核函数：

$$K(x_i, y_i) = \exp \left(-\frac{1}{2\sigma^2} |x - x_i|^2 \right) \tag{8}$$

式中 σ^2 为径向基核函数。

Sigmoid 核函数：

$$K(x_i, y_i) = \tanh [v(x - x_i) + C] \tag{9}$$

式中 v 为 Sigmoid 核参数。

本文计算中利用了最小二乘支持向量机分类算法,该方法是支持向量机算法的一种解法,即分类函数可以用最小二乘算法解出,因此称为最小二乘支持向量机分类算法。该方法最主要输入的参数是惩罚因子 C 和核参数 σ^2 。

2 多层次分类器的支持向量机

支持向量机只是一个二类分类算法,即它只能把数据分成两类。而在综合评价的过程中级别一般都超过了两类,例如本文计算的地下水水质标准有 5 级,因此必须要对其算法进行改进,以满足多级别分类的要求。本文研究设计了 1 个多层次的分类器来解决支持向量机的多级别分类问题。

该分类器主要由 5 个支持向量机两类分类器串联而成,每个支持向量机两类分类器用来辨别地下水环境健康属于哪个等级。在对数据进行训练时,将训练样本集中水质等级 I 作为一类,其类别标志为 1,其余样本为另一类,其类别标志为 -1,训练分类器 1 然后,再选取训练样本集中水质等级 II 作为一类,将其类别标志为 1,并将该类以外的其余样本为另一类,标志为 -1,训练分类器 2。按此方法,依次训练分类器 3、4、5。在进行地下水水质评价时,将水质样本输入分类器 1,若输出为 1,则评价测试结束,表明该地区的地下水水质等级为 I 级,否则该指标属性值样本将被自动输入到分类器 2 进行判别。若决策函数输出仍不为 1,则该指标属性值样本将依次自动输入到下一级分类器,继续进行识别分类,直到决策函数输出为 1,分类结束。这样,基于多层次分类器的支持向量机算法就可以进行多级别的分类计算。

3 磐石市地下水水质评价

在研究区选出 10 个采样点中 9 个水质指标:pH

值、总硬度、 Cl^- 、 SO_4^{2-} 、溶解性总固体、 Fe 、 NO_3^- 、 NO_2^- 、 F^- 。依据 GB/T14848—93《地下水质量标准》进行地下水质量分类。

由于支持向量机只接受 $-1 \sim 1$ 之间的数值计算,因此首先将所有的等级标准值换算到 $0 \sim 1$ 之间,即作为地下水环境健康预警诊断子系统中的数据输入部分,将原始数据进行预处理,应用公式(10)进行换算。

$$\bar{x}_{ij} = (x_{ij} - x_{j\min}) / (x_{j\max} - x_{j\min}) \quad (10)$$

最后将评价分级指标作为训练样本,具体测试结果作为分类样本,进行支持向量机的分类计算。

计算采用交叉验证比较方法来选择参数 C 和 σ^2 。首先确定惩罚因子参数集 $S_C = \{C_1, C_2, \dots, C_n\}$ 和核参数集 $S_{\sigma^2} = \{\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2\}$ 。然后从参数集中选择参数进行组合,对最小二乘支持向量机进行训练,选出最佳的参数组合作为支持向量机模型的最终参数。

最终经过多次模型训练,当 $C = 500, \sigma_1^2 = \sigma_2^2 = 0.1$ 时,参数为最佳组合,模型可以用来进行分类计算。

将具体的水质样本数据代入到训练好的模型中可以得到利用支持向量机进行水质分类的结果。同时为了对比,另外还用模糊综合评判法、灰色聚类法和神经网络法进行了评价计算。具体结果如表 1 所示。

表 1 各种水质评价方法的地下水环境质量评价结果

采样地点	模糊综合 评判法	灰色 聚类法	神经 网络法	支持 向量机法
常家屯内	II	II	II	II
水利局	V	V	V	V
市政府	III	III	IV	IV
植物油厂	V	V	IV	IV
一中	V	V	V	V
王家屯南	IV	V	IV	IV
武装部仓库南侧	II	II	II	II
七个顶子屯西	IV	IV	IV	IV
仙人洞山下	III	I	II	II
东兴立	III	III	IV	IV

从表 1 的地下水水质评价结果看,4 种综合评价模型的计算结果基本一致,在个别点稍有差别是因为每一种方法在评价中的侧重点不同。支持向量机算法与神经网络法的计算结果虽然相同,但在计算过程中,支持向量机算法的计算步骤、训练速度以及计算效率明显高于神经网络算法。

4 结 语

支持向量机算法是在统计学习理论的基础上提出的一种先进的算法,其本身具有在统计样本量较少的情况下获得良好的统计规律和更好的泛化能力的特点,非常适合综合评价计算。经过实证研究,可

以看出该方法的综合评价结果正确,但计算难度不大。因此,支持向量机算法在地下水综合评价计算中具有很好的发展前景。

参考文献：

[1] HSU Chi-wei, LIN Chih-jen. A comparison of methods for multiclass support vector machine [J]. IEEE Transacatutions on Neural Networks, 2002, 13 (2) :415-425.

[2] DUAN Kai-bo, SATHIYA KEERTHI S, AUN N P. Evaluation of simple performance measures for tuning SVM hyperparameters [J]. Neurocomputing, 2003, 51 :41-59.

[3] MAHMUDUR R, BIPIN C D, PRABIR B. Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion [J]. Computerized Medical Imaging and Graphics, 2008, 32 (2) :95-108.

[4] 赖永标, 乔春生, 刘开云, 等. 支持向量机在围岩稳定性分类中的应用 [J]. 水利学报, 2006, 37 (9) :1092-1096.

[5] 王伟, 赵利飞, 吴耿锋, 等. 地下水观测数据拟合与预测的支持向量机方法 [J]. 地震, 2007, 27 (1) :9-15.

[6] 武晟, 解建仓, 汪志荣, 等. 基于支持向量机的绿地径流系数预测模型的建立 [J]. 沈阳农业大学学报, 2007, 38 (1) :102-105.

[7] 卢敏, 张展羽, 冯宝平, 等. 基于支持向量机的区域水安全预警模型及应用 [J]. 计算机工程, 2006, 32 (5) :44-46.

[8] 徐红敏, 杨天行. 基于支持向量机分类算法的湖泊水质评价研究 [J]. 吉林大学学报, 2006, 36 (4) :570-573.

[9] 徐劲力. 支持向量机在水质评价中的应用 [J]. 中国农村水利水电, 2007 (3) :7-9.

(收稿日期 2008-06-28 编辑 高渭文)

第 5 届世界水资源论坛暨
世界水展 2009 年 3 月在土耳其举办

由世界水理事会和土耳其政府共同举办的第 5 届世界水资源论坛暨世界水展于 2009 年 3 月 16 日在土耳其伊斯坦布尔拉开帷幕, 共有来自 13 个国际机构以及 100 多个国家和地区的 2.5 万各界人士参与本届论坛及相关活动。本届论坛主题是“架起沟通水资源问题的桥梁”, 议题涵盖干旱、全球气候变化以及与水问题相关的健康、能源和农业问题, 并探讨解决水问题的新技术、新方法。

世界水资源论坛每 3 年举办 1 次, 是水资源领域规模最大的国际活动, 此前 4 届论坛分别在摩洛哥、荷兰、日本以及墨西哥举办。

第 5 届世界水论坛的主题是“跨越水的鸿沟”, 强调全球各地区参与水管理或受水影响的各方之间应加强沟通、交流和协调。论坛在中心主题下设立了 6 个子主题, 分别是“全球变化和风险管理”、“推进人类发展和千年发展目标”、“管理和保护水资源及供水设施以满足人类和环境的需求”、“水治理和管理”、“水融资”以及“教育、知识和能力建设”在子主题下又设立 24 个议题。会议期间通过部长级会议、区域日、议题分会、水展和水博览会以及水文化活动等多种形式开展了研讨和交流, 共同为世界水问题的解决做出努力。

(本刊编辑部 供稿)