

# 基于支持向量机的水质评估模型研究

邓伟萍, 桂超, 汪波, 石黎, 关培超

(湖北经济学院信息与通信工程学院, 湖北 武汉 430205)

✉dwppove@163.com; 1392979674@qq.com; 30857394@qq.com; 30941665@qq.com; 149293834@qq.com



**摘要:** 针对水质评估因子的模糊性和非线性特征, 且水质样本小类(如高污染水质类)因样本量少而容易导致误分的问题, 深入研究了支持向量机(SVM)这一善于解决非线性问题的智能模型, 设计了一种多宽度复合高斯核的支持向量机模型。该模型通过多个复合高斯核扩大和控制核函数宽度, 以此扩大样本间欧氏距离与差异, 以解决小类的误分问题。运用MATLAB平台对2017年全国98个重点断面水质周报数据进行算法对比实验, 结果证实多宽度核评估模型较好地提升了SVM的分类精度, 对水质分类问题是可行有效的, 对其他小样本分类问题也有一定的借鉴作用。

**关键词:** 水质评估; 多宽度高斯核; 支持向量机; 参数寻优

**中图分类号:** TP391.4 **文献标识码:** A

## Research on Water Quality Assessment Model based on Support Vector Machine

DENG Weiping, GUI Chao, WANG Bo, SHI Li, GUAN Peichao

(School of Information and Communication Engineering, Hubei University of Economics, Wuhan 430205, China)

✉dwppove@163.com; 1392979674@qq.com; 30857394@qq.com; 30941665@qq.com; 149293834@qq.com

**Abstract:** In view of the ambiguity and non-linear characteristics of water quality assessment factors, and the problem of small water quality samples (such as high-polluted water quality) that are easy to cause misclassification due to the small sample size, support vector machine (SVM), a smart model which is good at for solving nonlinear problems, is deeply studied. This paper proposes to design a support vector machine model with a multi-width compound Gaussian kernel. The proposed model expands and controls the width of the kernel function through multiple compound Gaussian kernels to expand the Euclidean distance and difference between samples, so that the problem of misclassification of small classes can be solved. The MATLAB platform is used to conduct algorithm comparison experiments on the water quality weekly report data of 98 key sections across China in 2017. The results prove that the multi-width kernel assessment model improves the classification accuracy of SVM, which is feasible and effective for water quality classification problems. It also provides a reference for problems of other small sample classification.

**Keywords:** water quality assessment; multi-width Gaussian kernel; support vector machine; parameter optimization

## 1 引言(Introduction)

随着经济的高速发展, 我国的水资源污染状况也愈发严峻, 局部地区的水质恶化事件时有发生。因此, 实时监测和评估地表水的质量与变化, 将为环境决策、工农业生产服务提供依据, 是防止污染与合理利用水资源的基础<sup>[1]</sup>。水质评

估单因子评估与多因子评估, 在没有突发水污染事件的情况下, 对断面进行长期监测与评估时一般采用多因子评估方法。

鉴于水质评估中各评估因子的模糊不确定性和非线性特征, 有学者采用智能计算方法进行水质评估。陈海洋等人<sup>[2]</sup>通过构建多个子分类器的决策树支持向量机模型评估水质的多

分类问题；马创等人<sup>[3]</sup>使用遗传算法与支持向量机建立自适应权重水质预测模型；方国华等人<sup>[4]</sup>采用粒子群算法，计算水量与水质联合配置模型；石晴宜等人<sup>[5]</sup>采用模糊神经网络计算洪泽湖入湖水质等问题；XIA等人<sup>[6]</sup>采用粒子群、蜂群、支持向量机等多种混合优化算法评定水质富营养化分级。上述方法解决了水质评估的多分类问题，也对参数寻优的过程进行了优化，但对分类过程中的小样本类别的误分问题没有侧重考虑，本文决定对支持向量机(Support Vector Machine, SVM)中的核函数进行改进，重点解决多因子分类中小样本训练不够易误分的问题。

## 2 支持向量机(Support Vector Machine)

支持向量机分类模型的原理是通过核函数将样本点映射到多维特征空间，通过构造最优分类超平面，使得超平面与不同类样本集之间的距离最大，从而达到最大的泛化能力。

SVM标准算法中，设有训练样本集 $\{(x_1, y_1), \dots, (x_i, y_i)\}$ ,  $x \in R, y \in \{-1, 1\}$ ,  $y = -1$ 与 $y = +1$ 分别表示两类不同的样本；样本集可被一超平面 $H$ , 即 $(w \cdot x) + b = 0$ 没有错误地分开，对任意一个训练样本 $(x_i, y_i)$ 都有：

$$y_i[(w \cdot x) + b] \geq 1, i = 1, 2, \dots, n \quad (1)$$

使分类间隔最大的分类面称为最优分类超平面，寻找过程转化为求如下一个二次规划问题： $\min \|w\|^2/2$ ，满足约束条件(1)。采用LaGrange转换，将二次规划问题转为如下一个对偶问题：

$$\max Q(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m a_i a_j y_i y_j x_i' x_j \quad (2)$$

满足约束条件： $\sum_{i=1}^n a_i y_i = 0$ 且 $a_i \geq 0, i = 1, 2, \dots, n$ 。求解该问题的最优解，则求得 $a_i^*$ 是最优的拉格朗日乘子，对应的 $x_i$ 是支持向量，得到最优分类面的判别函数：

$$f(x) = \text{sgn} \left\{ \sum a_i^* y_i K(x_i, x) + b \right\} \quad (3)$$

$K(x_i, x)$ 称为核函数， $K(x_i, x)$ 将高维特征空间中内积运算转化为低维模式空间上一个简单的函数计算。核函数中以高斯核函数(Radial Basis Function, RBF)运用最广泛：

$$K(x_i, x_j) = \exp \left( -\frac{\|x_i - x_j\|^2}{2\sigma^2} \right) \quad (4)$$

其中， $\sigma$ 为高斯分布的宽度。

## 3 多宽度高斯核支持向量机(Multi-width Gaussian kernel Support Vector Machine)

高斯核支持向量机中可调的参数仅有核宽度 $\sigma^2$ ，样本映射到特征空间后其分布是不均匀的，容易导致支持向量机模型在样本集中的区域产生过学习现象，而在样本稀疏区又学习不足，从而产生错分问题或陷入局部最优解现象。针对此

类问题，多宽度高斯核被提出<sup>[7]</sup>，其核函数是一种复合函数，形式如下：

$$K(x_i, x_j) = \left[ \exp \left( -\frac{\|x_i - x_j\|^2}{2\sigma^2} \right) + \theta \right]^d \quad (5)$$

$\forall \theta \geq 0, \forall d \geq 0, d \in N$

通过二项式定理展开后：

$$\begin{aligned} K(x_i, x_j) &= \sum_{s=0}^d \binom{d}{s} \theta^d - s \exp \left( -\frac{s \|x_i - x_j\|^2}{2\sigma^2} \right) \\ &= \theta^d + \sum_{s=1}^d \binom{d}{s} \theta^d - s \exp \left( -\frac{s \|x_i - x_j\|^2}{2\sigma^2} \right) \end{aligned} \quad (6)$$

形成一种复合核函数，是由一系列不同宽度的高斯核构成，每个核的宽度为 $\sigma^2$ ， $\left[ \frac{\theta}{s} \right] \theta^{d-s}$ 影响单个高斯核的空间收放程度。文献[8]阐述了当 $\theta > 1$ 时，二项式展开式的前面若干个高斯核将被放大，影响程度高，权重加大；反之，当 $\theta < 1$ 时，二项式展开式的后面若干个高斯核将被放大，权重加大。

多个不同宽度的高斯核累加在一起，通过多参数调节，能解决其可调参数单一容易导致的过学习问题，提高单一高斯核函数的抗干扰性和泛化能力；常量因子 $\theta^d$ 将点集之间的矢量距离放大了，这也将扩大样本点映射在特征空间的差异，从而提高分类器的效率<sup>[9]</sup>。

多宽度核的参数增加，给参数寻优增加了困难<sup>[10]</sup>。现讨论参数对核函数的影响，对式(5)做简化处理：

$$K(x) = \left[ \exp \left( -\frac{\|x\|^2}{2\sigma^2} \right) + \theta \right]^d \quad (7)$$

令 $\lambda = 1/(2\sigma^2)$ ，则：

$$K(x) = \left[ \exp(-\lambda \|x\|^2) + \theta \right]^d \quad (8)$$

为对比不同参数下多宽度核函数的作用域与收敛情况，现设几组参数进行对比，令：

$$y1 = \exp(-2 \times x^2)$$

$$y2 = \exp(-8 \times x^2)$$

$$y3 = (\exp(-2 \times x^2))^8$$

$$y4 = (\exp(-8 \times x^2) + 1.5)^3$$

$$y5 = (\exp(-2 \times x^2) + 0.5)^3$$

当 $\theta = 0, d = 1$ 时，是普通的高斯核； $\lambda$ 增大，核宽度明显缩小，有利于函数收敛；当 $\theta = 0, d$ 增大时，等同于 $\lambda$ 增大，核宽度缩小；当 $\theta > 0, d > 1$ 时，核函数径向作用范围被显著拉伸，加大样本在特征空间的距离差异，有利于分类，三个参

数需要结合样本进行协调<sup>[11]</sup>。

4 实例验证(Example verification)

本文以全国主要流域重点断面水质自动检测周报<sup>[12]</sup>作为数据源，选取了2017年第1周、第8周、第31周、第44周不同时期的水质监测数据。以PH值、溶解氧(DO)、高锰酸盐指数(CODMn)与氨氮(NH<sub>3</sub>-N)四项因子作为水质评估指标。2017年有效断面监测点为98个，包括长江、黄河、淮河、松花江、太湖、辽河、海河等主要江河流域和重要湖泊，取58个样本作为训练集进行SVM模型的训练，另外取40个样本作为测试集进行预测。表1列出了中华人民共和国国家地表水环境质量标准(GB3838—2002)提供的指标限值标准。

表1 地表水环境质量标准基本指标标准限值

| Tab.1 Standard limits of basic indicators of surface water environmental quality standards |      |      |     |      |     |     |
|--------------------------------------------------------------------------------------------|------|------|-----|------|-----|-----|
| 评估因子/mg/L                                                                                  | 限值关系 | I类   | II类 | III类 | IV类 | V类  |
| PH值(无量纲)                                                                                   |      | 6—9  |     |      |     |     |
| 溶解氧(DO)                                                                                    | ≥    | 7.5  | 6   | 5    | 3   | 2   |
| 高锰酸盐指数(CODMn)                                                                              | ≤    | 2    | 4   | 6    | 10  | 15  |
| 氨氮(NH <sub>3</sub> -N)                                                                     | ≤    | 0.15 | 0.5 | 1.0  | 1.5 | 2.0 |

本文以LIBSVM软件包为开发工具，在MATLAB 2016平台上分别对水质样本做了比较试验，对比研究了不同样本集、不同参数下多宽度高斯核与标准高斯核的分类结果。以2017年第31周水质样本为例，阐述实验结果：参数组如表2所示，参数对应式(7)中的 $\sigma$ 、 $\theta$ 、 $d$ 和惩罚因子 $c$ ，令 $\lambda=1/(2\sigma^2)$ ，为突出比较其他主要参数，表2中统一设 $c=2$ 。多宽度高斯核分类准确率对比如表3所示。多宽度高斯核与标准高斯核(RBF)的最优预测结果如图1、图2所示，反映一类至五类及劣五类共六个水质等级，符号○表示Labels，代表实际水质评估结果；符号\*与▼分别表示多宽度高斯核与标准高斯核的预测值。

表2 多宽度高斯核的参数组合

| Tab.2 Parameter combination of multi-width Gaussian kernel |     |     |     |     |     |     |
|------------------------------------------------------------|-----|-----|-----|-----|-----|-----|
| 参数                                                         | 组号  |     |     |     |     |     |
|                                                            | I   | II  | III | IV  | V   | VI  |
| $c$                                                        | 2   | 2   | 2   | 2   | 2   | 2   |
| $\lambda$                                                  | 2   | 4   | 8   | 16  | 4   | 8   |
| $d$                                                        | 5   | 20  | 8   | 3   | 5   | 5   |
| $\theta$                                                   | 2.5 | 0.8 | 1.5 | 0.5 | 0.5 | 2.5 |

表3 不同参数组下的多宽度高斯核的分类准确率

Tab.3 Classification accuracy of multi-width Gaussian kernel under different parameter groups

| 分类准确率/% |      |      |      |      |      | 平均准确率/% |
|---------|------|------|------|------|------|---------|
| I       | II   | III  | IV   | V    | VI   |         |
| 87.5    | 90.0 | 97.5 | 92.5 | 87.5 | 92.5 | 91.3    |

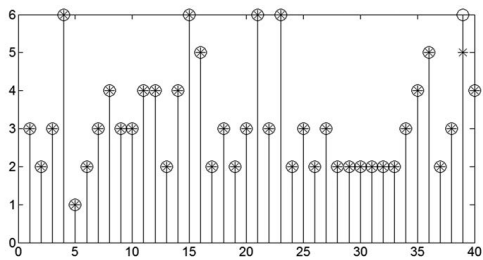


图1 多宽度高斯核的最优预测结果

Fig.1 Optimal prediction results of multi-width Gaussian

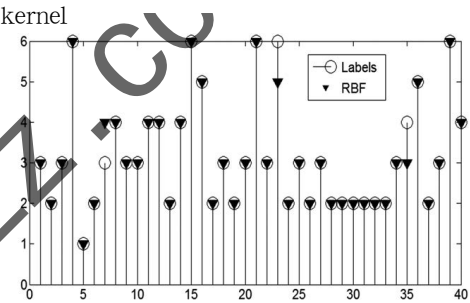


图2 标准高斯核的最优预测结果

Fig.2 Optimal prediction results of standard Gaussian kernel

从实验结果可以看出，参数 $d$ 放大高斯核径向作用的范围，拉大了样本距离，对高斯影响最大，直接影响精度。对比参数组II与V、III与VI，在 $\lambda$ 相同的情况下， $d$ 增大，准确率均有提高。 $\lambda$ 决定高斯核的宽度，对比参数组I、V、VI，在 $d$ 相同的情况下， $\lambda$ 增大，核宽度缩小，有利于加快收敛，快速找到最优解。但 $\lambda$ 和 $d$ 增长到一定程度，精度将不再提高，如参数组II、III、IV，提高将导致过学习状态产生。

现讨论多宽度核的特例情况，令 $d=1$ ， $\theta=0$ ，多宽度核转变为标准高斯核，其可调节参数为 $c$ 和 $\lambda$ ，调节参数后经实验演算，得到如图2所示的最优解。

对比图1、图2中相同40个预测样本在不同方法下的最优预测结果可以发现，多宽度高斯核评估结果更吻合实际评估值，其分类精度要优于标准高斯核(RBF)，特别是在易错分的四、五、六等小类上分类精度也有所提高。

5 结论(Conclusion)

分析水质评估问题的特点后，引入多宽度高斯核SVM方法进行水质评估。多宽度高斯核是一种复合核函数，较之标准高斯核，其通过拉大样本在特征空间的欧式距离来降低误

(下转第28页)