文章编号:1671-4598(2020)06-0140-04 DOI:10.16526/j.cnki.11—4762/tp.2020.06.029 中图分类号:X832 文献标识码:A

基于支持向量机的水质浊度补偿研究

李 文,王兴浩,何云霄,罗学科
(北方工业大学 机械与材料工程学院,北京 100144)

摘要:针对传统浊度传感器的非线性误差,无法满足直接对水中浊度进行精确测量的需求,提出了一种支持向量机的方法补偿其性能;而支持向量机中惩罚系数 C 和核参数 γ 决定了其补偿的性能,传统支持向量机寻参方法速度慢、运算量大,具有一定的局限性;针对其参数的选择优化提出了改进的网格搜索法优化支持向量机,即采用改进的网格搜索法来针对水质浊度监测传感器补偿系统的特性来优化选择 C 和 γ;实验结果表明,基于网格搜索法的支持向量机测量精度达到 93.0%,其各项测量误差满足实际标准要求。

关键词:支持向量机;水质监测;浊度;网格搜索法

Turbidity Sensor Compensation Method Based on SVM

Li Wen, Wang Xinghao, He Yunxiao, Luo Xueke
(North China University of Technology, Beijing 100144, China)

Abstract: Aiming at the non-linear error of the traditional turbidity sensor, which can't meet the demand of directly measuring the turbidity in water, a support vector machine method is proposed to compensate its performance. The performance of compensation is determined by the penalty coefficient C and kernel parameter γ in SVM. The traditional method of finding parameters in SVM is slow and requires a lot of computation, which has some limitations. An improved grid search method is proposed to optimize support vector machine (SVM) for the selection and optimization of its parameters. That is to say, the improved grid search method is used to optimize the selection and compensation of water quality turbidity monitoring sensor compensation system. The experimental results show that the measurement accuracy of SVM based on grid search method is 93.0%, and the measurement errors meet the actual standard requirements.

Keywords: support vector machine; water quality monitoring; turbidity; grid search method

0 引言

随着水环境污染越来越严重,传统的手动实验检测水环境质量已不能满足水质监测的实时性和准确性的标准要求。近些年来,水质在线监测发展迅速,由于其能够实时检测到水域污染的变化和较高的准确率,被推广并应用于水质监测的各个领域[1]。在这些水质监测仪中关键器件是传感器。水质监测传感器在实际应用中由于环境等诸多因素导致测量精度低、稳定性差,而对水质监测传感器的输入输出非线性关系的补偿是提高系统测量精度的必要方法[2]。近年来智能算法在补偿建模中发展迅速,其中有人工神经网络、支持向量机等各种算法。神经网络最严重的问题是没能力来解释自己推理的过程和依据,而且数据不充分时就无法工作,同时神经网络的理论和学习算法还有待进一步提高。支持向量机(SVM)是一种新颖的小样本学习方法,它有着坚实的理论基础,在实际应用中,支持向量机能够有效避免从归纳到演绎的传统过程,能够高效地从训练样本中推导出预测样本,在分类和回归等问题上,能够有效地简化步骤,提高了效率和准确率。大量实验和研究表明,基于支持向量机建立的回归模型,无论是在逼近能力,还是在泛化性上,都要优于神经网络以及其他智能算法。

支持向量机以统计学为理论基础,从 1995 年提出后,在小样本、非线性和模式识别等各个领域中迅速发展,并且具有很多优势,并能够推广到函数拟合等其他实际问题中。支持向量机在与神经网络相比较,支持向量机的原理是结构风险最小化,弥补了神经网络的缺点,在数据量较少的情况下依然具有很好的推广能力。但是在实际应用中,支持向量机有两个重要的参数,即惩罚系数 C 和核参数 γ,如果参数的选择不当,则会直接影响整体的性能。到目前为止支持向量机的参数优化并没有标准化的方法,所以目前应用在支持向量机的参数选择的优化方法各种各样。针对研究问题,以水质浊度参数检测作为实验背景,提出一种改进的网格搜索法优化支持向量机参数,来提高其准确率和优化速度,与其他优化方法相比较,并在实际测量环境中取得了很好的结果。

1 水质监测传感器补偿原理

传统的传感器输入输出特性为 $y=f(x)$,$x\in(\zeta a,\zeta b)$,

收稿日期:2019-10-21; **修回日期**:2019-11-07。
基金项目:国家自然科学基金(51205005);北京市科技创新服务能力建设-PXM2017-014212-000013。
作者简介:李 文(1975-),男,山东泰安人,博士,副教授,硕士生导师,主要从事机器人技术和光学智能传感器方向的研究。

式中 $f(x)$ 为非线性函数，y 表示测量参数后输出的电压值信号，x 表示测量参数输入的溶液值，ζa，ζb 为溶液真实值的范围。在已知溶液值 x 的情况下，y 电压值信号可使用浊度传感器测量溶液得到，其目的是根据输出的电压值 y 求得未知的输入变量 x，既表示为 $x = y^{-1}(y)$。

而在实际应用中，由于环境或者传感器自身硬件会导致的测量的值存在非线性误差。为了校正这种非线性误差，使其输出的电压值信号 y 通过一个校正环节[3]，如图 1 所示。校正模型的函数为 $u = g(y)$，式中 u 为校正系统非线性后的输出，它与输入的溶液值 x 呈线性关系，使得补偿后的传感器具有理想特性[4]。在实际中，$g(*)$ 的表达式难以确认，那么建立支持向量机回归补偿模型就成了解决此模型表达式的重要因素[5-6]。

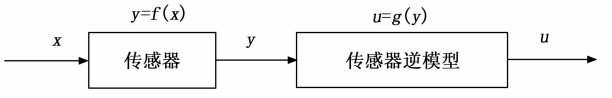

图 1　非线性误差校正模型

2　支持向量回归机理论

20 世纪 90 年代，Vapnik 等人基于小样本统计学提出了支持向量机理论，其基本原理是以训练误差作为要解决问题的约束条件，以最小置信区间作为优化的最终目标。其本质就是解决一种凸规划或者二次规划问题[7]。支持向量机首先通过内积核函数将非线性的变换问题映射到一个高维空间，变成一个线性问题来求广义分类面或回归问题。

对于给定的一组数据 $T = \{(x_1, y_1), \cdots, (x_i, y_i)\} \subset R^d \times R, i = 1, \cdots, n$，我们要解决的回归问题简单来说就是找到 x_i 与 y_i 之间的映射关系：

$$y = f(x) = [\omega, \varphi(x)] + b, \quad x \in R^d; y, b \in R \tag{1}$$

式中，$[\omega, \varphi(x)]$ 对应的是 R^d 空间的内积。$\varphi(x)$ 为核函数，把训练样本数据映射到高维空间 F 上，因此它的思想就是把原空间的非线性问题映射到高维空间中转变为高维空间的线性问题，解决其对应的线性回归问题[8]。

支持向量机回归理论对这一类问题的表述为在一组函数 $\{f(x, \omega)\}$ 种，寻找最优的一个函数 $\{f(x, \omega^*)\}$，使预期的期望风险 $R(\omega)$ 达到最小化[9]。

$$R(\omega) \leqslant R_{emp}(\omega) + \sqrt{\frac{h(\ln(2n/h)+1) - \ln(\eta/4)}{n}} \tag{2}$$

式中，n 为样本容量，h 为 VC 维。支持向量机把上式转化为寻求下式的最优解：

$$W_{\max} = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i - a_i^*)(a_j - a_j^*)k(x_i, x_j) - \varepsilon\sum_{i=1}^{n}(a_i + a_i^*) + \sum_{i=1}^{n} y_i(a_i - a_i^*)$$

$$s.t.\begin{cases}\sum_{i=1}^{n}(a_i - a_i^*) = 0 \\ 0 \leqslant a_i, a_i^* \leqslant C, i = 1, 2, \ldots, n\end{cases} \tag{3}$$

其中：ε 根据不敏感损失函数 $L(y, f(x, a))$ 来决定回归曲线的平坦度，给定 $0 < \varepsilon < 1$。当 x 点处的实际结果值 y 与预测值 $f(x)$ 之间的误差值不超过预先给定的 ε 时，那么就认为该点的预测值 $f(x)$ 是无损失的[10]。

$$L(y, f(x, a)) = L(|y - f(x, a)|_\varepsilon) \tag{4}$$

其中：

$$|y - f(x, a)|_\varepsilon = \begin{cases} 0, y - f(x, a) | \leqslant \varepsilon \\ |y - f(x, a)| - \varepsilon, \quad \text{其他} \end{cases} \tag{5}$$

式（3）中，C 为惩罚因子，表示对错分样本的惩罚。

由此得到最优解 $a^{(*)} = (a_1, a_1^*, \ldots, a_i, a_i^*)^T$，因此支持向量机对应的回归函数为：

$$f(x) = \sum_{i=1}^{SV}(a_i - a_i^*)k(x_i, x_j) + b \tag{6}$$

在支持向量机中，综合考虑到 RBF 高斯径向基函数所体现出的较好性能，选取式（4）中的 RBF 核函数作为支持向量机的核函数。在实际应用中，传统的参数选择方法大多都是凭借大量经验或者反复试算法，导致选择不准确使得补偿精度达不到目标精度的要求，且效率低。因此正确的方法来选择核函数参数和惩罚系数，对 SVM 的性能以及水质监测的补偿精度至关重要。

$$k(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \gamma > 0 \tag{7}$$

3　支持向量机参数选择方法

3.1　支持向量机模型选择的研究现状

对于支持向量机的性能，最重要的影响因素就是两个参数值惩罚系数 C 和核参数 γ 的选取。惩罚系数 C 体现了对误差的宽容度。C 的取值越高，建立的回归模型越不能容忍出现误差，会造成过拟合现象。如果 C 的取值过小，则会出现欠拟合。如果 C 取值不当，过大或过小，泛化能力都会变差。核参数 γ，是选取的高斯径向基 RBF 核函数自带的参数，γ 取值过大，其支持的向量会越少。γ 取值过小，其支持的向量会越多[11]。

到目前为止，关于 SVM 的参数选择优化并没有标准的结构化方法。相关的优化方法各有优缺点，常用的方法有：实验法、遗传算法、粒子群算法和网格搜索算法等。实验法就是通过大量的实验比较结果精度来确定参数，这种方法虽然能够找到合适的参数，但是效率低。遗传算法思想来源于自然界的生物遗传和进化，是一种应用较为广泛的全局搜索功能的优化算法[12]。遗传算法依据适者生存的进化原理，通过众多的个体不断地经过选择、遗传、变异的过程，筛选出最优的个体，即为最优的参数解，遗传算法对于问题本身可以不用知道，它只是对优化过程中的每个个体进行评估和筛选[13]。粒子群算法基本概念源于对鸟群觅食行为的研究，即自由个体组成的群体与周边环境以及个体之间的互动性为，是一种新颖的优化算法。它的基本思想就是将问题所有可能的解都看作是一个微粒，每个微粒在其解空间中飞行，通过其适应度函数的标准判别粒子

的优劣性，并根据解空间中其他微粒传递的飞行经验进行调整，想着最好的微粒位置飞行，以此来得到最优解[14]。遗传和粒子群算法属于启发式算法，他们不必遍历所有参数集合也可以找到全局最优解，但是这两种算法操作比较复杂，并且容易陷入局部循环，得到的解也只是局部最优解[15]。

3.2 网格搜索法

网格搜索法是一种穷举遍历算法，它将所有可能的参数组合在空间中划分成若干网格，遍历网格中所有交点，对每个参数集合应用交叉验证来计算误差，得到误差最小的为全局最优解。网格搜索法可以从较多参数中获得最优解，但是效率低[16]。

针对上述网格搜索法的缺点，选择改进的网格搜索法，即先在给定的参数范围内进行大步距粗略搜索，确定一个结果较优的参数组合存在的区间，在此区间附近内再进行小步距精确搜索，来改进传统网格搜索法的缺点，提高其优化精度和优化速度。

3.3 改进的网格搜索法

网格搜索法本质是让惩罚系数 C 和核函数参数 γ 的集合在其范围内生成网格，并对网格内所有点进行评价，最终取得整个模型训练集的平均验证均方根误差（MSE）最小的那组为最优参数组合[17]。计算得到的最优参数为图 2 所示，其中 C 的范围在 $[2^{-10},2^{20}]$，γ 设置的范围是 $[2^{-10},2^{10}]$，步距为 0.1。

由图 1 可以看出，参数组合在一定的区间范围内准确率很高，但是在整个范围内准确率相对偏低，如果以 0.1 步距全部遍历整个区间，将使得整个算法效率降低，因此先找到平均验证均方根误差较小的参数区间再进行精确搜索，将能减少大量的计算，节约时间提高效率。

针对上述传统网格搜索法的问题，选择改进的网格搜索法作为参数优化方法。首先在给定的参数组合范围内进行大步距粗搜，选择训练集的平均验证均方根误差最小的一组参数组合。若参数选择过程中搜索出多组达到最小平均验证均方根误差的参数组合，则选择 C 最小的那组，如果对应 C 最小的有多组 γ，那么就选择搜索到的第一组作为最佳参数组合。因为惩罚系数 C 如果过高将导致过学习现象，寻得这组局部最优参数组合之后，在此参数组合点附近选择一个小区间，采用小步距进行第二次精搜，找到的最优参数即为全局最优参数组合。

4 实验结果及分析

4.1 实验器材准备

浊度是水体中一种重要的特征参数，体现了水环境的清洁度和卫生状况，它是衡量水环境质量的重要依据，并且也作为影响其他参数的干扰因素，不管是民用还是环境监测都是必须要测量的参数[18]。所以选取了浊度参数作为研究对象。

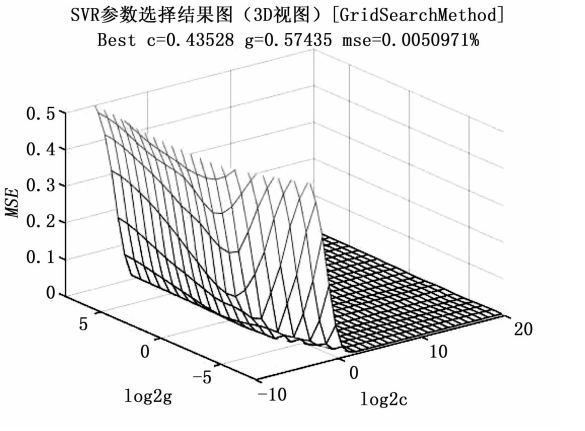

图 2 网格搜索法参数选择结果

实验器材包括配置浊度溶液所需要的烧杯、玻璃棒、计量筒等准备工具，以及浊度传感器。所用到的标准试剂是中国计量院化学所购置的标准溶液。采用超纯水作为零点校正液，主要用于稀释溶液。

通过上述材料工具来对溶液进行测量。选取数据时，由于温度对于浊度测量结果有着影响，分别选取 5°、10°、15°、20°和 25°的输入电压值。实验数据一共选取 10 组传感器有效数据共 70 个样本，随机打乱顺序选取 60 个样本作为训练样本，剩下 10 个作为测试样本，实验数据如表 1 所示。

表 1 实验数据

	标准溶液值	5°测量电压值	10°测量电压值	15°测量电压值	20°测量电压值	25°测量电压值
1	0	799.07	824.33	850.65	879.81	907.51
2	25	777.67	804.93	832.50	855.48	876.18
3	50	763.22	778.61	811.58	834.95	853.83
4	100	709.93	739.98	758.85	787.30	803.39
5	200	609.82	633.16	658.99	681.06	692.96
...	...	...	...	...	...	...
66	50	764.89	782.63	812.63	834.95	855.20
67	100	712.08	740.12	765.40	787.96	805.14
68	200	608.91	634.74	659.86	682.66	694.21
79	500	358.95	373.47	387.56	401.17	411.60
70	800	175.08	184.43	189.66	197.78	202.29

4.2 实验过程

采用 Matlab 平台结合开源的 LIBSVM 工具包，进行网格搜索法优化支持向量机参数仿真测试。实验过程如下：

1）确定网格搜索法的参数变量 C 和 γ 的取值范围，C 的初始范围在 $[2^{-10},2^{20}]$，γ 设置的初始范围是 $[2^{-10},2^{10}]$。传统的网格搜索法的步距一般为 0.1，改进的方法将步距放大 100 倍，即步距为 10。以 2 的幂次方沿着两个区间范围方向生成网格。将整个网格区间分别分为 M、N 等分，网格中的节点即为给定范围内所有可能的参数组合[19]。

2）针对所有的参数组合 $(C_i,\gamma_j)(i=1,\ldots,M,j=1,\ldots,N)$，对训练样本集进行训练，得到训练样本集的平均验证均方根误差最小的参数组合 (C_i,γ_j)，判断是否达到精度标准要求，如果满足转到 4），否则转到 3）。

3）在参数 (C_i,γ_j) 相邻的两个区间作为新的参数范围 $C\in[C_{i-1},C_{i+1}],\gamma\in[\gamma_{j-1},\gamma_{j+1}]$，并分别减少搜索步距的两倍，因为网格的范围是以 2 的幂次方的。再次进行最优参数组合的搜索，判断是否满足平均验证均方根误差要求，如果满足则跳转到 4），否则继续在 3）循环进行直到找到最优的参数组合。

4）储存得到的最优参数组合和选择结果，参数优化过程结束。

4.3　实验结果

在采用改进的网格搜索法进行支持向量机的参数选择后，预测结果精度如图 3 所示。

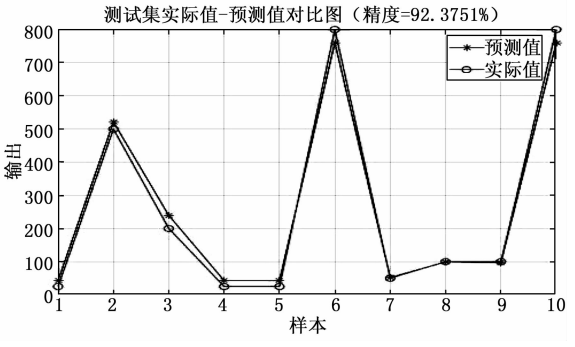

图 3　网格搜索法预测精度

为了便于分析和比较，在本次实验中还分别采用了遗传算法和粒子群算法进行参数寻优，与改进的网格搜索法进行对比。

从表 2 看出，遗传算法虽然也能够得到较高的预测精度，但在实验中容易出现过早收敛，出现局部最优，搜索效果不稳定。粒子群算法搜索性能较稳定，但耗时较长。相比较而言，改进的网格搜索法是精度最高并且时间较短的优化算法。

表 2　不同优化算法性能对比

参数	改进的网格搜索法	遗传算法	粒子群算法
Best C	0.44	12.10	138.54
Best γ	0.57	0.67	0.69
准确率/%	92.38	91.92	91.60
时间/s	18.4	80.2	120.8

5　结束语

应用改进的网格搜索法优化支持向量机方法对浊度传感器进行了预测校正，并与粒子群算法、遗传算法进行了比较。经实验结果表明：改进的网格搜索法优化支持向量机方法更好地实现了对浊度传感器的预测校正，显著改善了传统网格搜索法的性能，提高了准确率，减少了优化时间，相比较其他优化方法具有更好的性能，对传感器的非线性校正提供了一种可行有效的方法。

参考文献：

[1] 张　颖. 浅谈我国地表水水质监测现状 [J]. 科技信息，2011 (26)：59，61.

[2] 曾国栋. 地表水水质监测现状与措施 [J]. 环境与发展，2018，30 (9)：131－133.

[3] 卢智远，周永军，李卫军. 传感器非线性误差校正的 BP 神经网络方法研究 [J]. 传感器技术，2005，24 (2)：11－12.

[4] 周鸣争，汪　军. 基于支持向量机的传感器非线性误差校正 [J]. 电子科技大学学报，2006，35 (2)：242－245.

[5] 李如发，卢文科. 基于 MLP 传感器的非线性校正 [J]. 湖北大学学报 (自然科学版)，2014，36 (2)：181－184.

[6] 刘　涛，王　华. 传感器非线性校正的遗传支持向量机方法 [J]. 电子测量与仪器学报，2011，25 (1)：56－60.

[7] 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)：34－39.

[8] 丁世飞，齐丙娟，谭红艳. 支持向量机理论与算法研究综述 [J]. 电子科技大学学报，2011 (1)：4－12.

[9] 张博洋. 支持向量机理论与应用研究综述 [J]. 无线互联科技，2015，71 (19)：116－117.

[10] Smolaa A J，Scholkopf B，Muller K R. The connection between regularization operators and support vector Kernels [J]. Neural Networks，1998，11 (4)：637－649.

[11] 林升梁，刘　志. 基于 RBF 核函数的支持向量机参数选择 [J]. 浙江工业大学学报，2007，35 (2)：163－167.

[12] 边　霞，米　良. 遗传算法理论及其应用研究进展 [J]. 计算机应用研究，2010，27 (7)：2425－2429.

[13] 李良敏，温广瑞，王生昌. 基于遗传算法的回归型支持向量机参数选择法 [J]. 计算机工程与应用，2008，44 (7)：23－26.

[14] 熊伟丽，徐保国. 粒子群算法在支持向量机参数选择优化中的应用研究 [A]. 中国控制与决策学术年会 [C]. 2007.

[15] 杜树新，吴铁军. 模式识别中的支持向量机方法 [J]. 浙江大学学报 (工学版)，2003，37 (5)：521－527.

[16] Min S，Lee J，Han I. Hybrid genetic algorithms and support vector machines for bankruptcy prediction [J]. Expert Systems with Applications，2006，31 (3)：652－660.

[17] 王　健，张　磊，陈国兴，等. 基于改进的网格搜索法的 SVM 参数优化 [J]. 应用科技，2012 (3)：32－35.

[18] 宋建军. 基于偏最小二乘回归和浊度补偿的化学需氧量监测传感器的算法研究 [D]. 成都：四川师范大学，2017.

[19] 郭　英. 基于改进支持向量机算法的水质监测模型研究 [J]. 水科学与工程技术，2018，210 (4)：25－28.