文章编号:1004-9037(2014)05-0790-05

基于改进 LS-SVM 的随钻测量数据传输误码率预测

刘新平 薛希文

(中国石油大学计算机与通信工程学院,青岛,266555)

摘要:针对泥浆连续波随钻测量数据传输误码率预测精度低、数据传输过程中易受干扰信号影响等缺点,提出利用改进的最小二乘向量积(LS-SVM)对连续波数据传输误码率建立预测模型,并引用遗传算法对参数寻优,在建立模型过程中利用狄克逊准则对数据进行筛选,从而提高误码率预测的精度。在小样本数据的情况下,采用 Matlab 建立基于改进的最小二乘支持向量机泥浆连续波数据传输模型。仿真结果表明该模型能够有效地避免陷入局部最优问题,具有较强的泛化能力和预测能力。通过与误差反传前馈(Back propagation,BP)和 Elman 神经网络预测模型对比可知,该模型预测精度更高,预测值更接近于实际值,可以用于泥浆连续波数据传输误码率预测。

关键词:随钻测量;最小二乘支持向量机;遗传算法;狄克逊准则

中图分类号:TE927.6 **文献标识码**:A

Prediction of Error Rate in Measurement-While-Drilling Data Transmission Based on Improved LS-SVM Method

Liu Xinping[1], *Xue Xiwen*[2]

(Computer and Communication Engineering College, China University of Petroleum, Qingdao, 266555, China)

Abstract: In the continuous wave measurement-while-drilling (MWD) system, the accuracy of error rate prediction is low and the data transfer process is affected by interference signals. A model for error rate prediction of continuous-wave data transmission is proposed by using the improved least squares support vector machines (LS-SVM), and the genetic algorithm is used to search the optimized parameter to improve the prediction accuracy of the model. During establishing the model, Dixon criteria is used to screen the data and improve the error rate prediction accuracy. With small samples, mud continuous-wave data transmission model is established by using Matlab based on the improved LS-SVM. The simulation results show that the model can avoid falling into local optimization problem effectively, and has strong generalization and prediction ability. Compared with back propagation(BP) and Elman neural network, the model has higher prediction accuracy, so it can be used to predict the error rate of mud continuous-wave data.

Key words: measurement while drilling; least squares support vector machine(LS-SVM); genetic algorithm; Dixon criteria

引 言

无线随钻测量仪器在石油钻井工程中的应用越来越普遍,利用连续波传输方式可以提高数据传输速率,增强抗干扰能力[1-2]。但由于数据传输过程中受到泥浆泵、扭矩、螺杆等产生的干扰信号共同影响,会使数据接收的误码率增加,因此如何针对误差参数变化快速而准确地建立误差预测数学模型具有重要的工程意义。文献[3]通过小波神经网络提出强干扰信号来检测弱泥浆信号,这种方法的缺点是小波神经网络的训练时间太久,结果易陷

基金项目:山东省自然科学基金(ZR2010FM028)资助项目;中央高校基本科研业务费专项资金(12CX04075A)资助项目。

收稿日期:2012-02-17;**修订日期**:2012-03-18

入局部极小值。文献[4]采用线性滤波方法还原脉冲信号，又利用一种非线性“平顶消除”的方法对现场采集的信号进行了处理，但文献中只是分析信号去噪，并没有进行误差率预测。支持向量机器(Support vector machine，SVM)是神经网络领域中的一种新的重要方法，被广泛应用于模式识别[5]、多类型分类[6]、金融时间序列预测[7]等方面。但 SVM 的求解涉及二次规划问题，计算复杂、效率低，而最小二乘支持向量机(Least square SVM，LS-SVM)用等式约束代替标准 SVM 的不等式约束，将二次规划问题转化为线性方程组求解，降低了计算复杂性，加快了求解速度和抗干扰能力[8]，最小二乘向量机被广泛应用于非线性系统预测[9]、算法优化[10]等问题中。本文利用 LS-SVM 建立泥浆连续波误码率预测模型，并将遗传算法(Genetic algorithm，GA)嵌入到 LS-SVM 中进行参数寻优，同时对异常数据利用狄克逊准则筛选剔除，提高了模型在误码率预测中的适用性与合理性。

1　基于改进 LS-SVM 泥浆连续波误码率预测模型

1.1　LS-SVM 回归模型原理

设所有训练样本$(x_i, y_i, i=1,2,\cdots,n, \boldsymbol{x}\in\boldsymbol{R}^l, \boldsymbol{y}\in\boldsymbol{R})$能在精度 ε 下无误差地用线性函数 $\boldsymbol{y}=\boldsymbol{\omega}^{\mathrm{T}}\cdot\boldsymbol{x}+b$ 拟合，则根据结构风险最小化准则，这一优化目标在最小化 $\|\boldsymbol{\omega}\|^2/2$ 时可获得较好的推广能力[11]。

$$\min J(\boldsymbol{\omega},\xi)=\frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\cdot\boldsymbol{\omega}+C\sum_{i=1}^{n}(\xi_i+\xi_i^*)$$
$$\text{s.t.}\begin{cases}y_i-\boldsymbol{\omega}^{\mathrm{T}}\cdot x_i-b\leqslant\varepsilon+\xi_i\\ \boldsymbol{\omega}^{\mathrm{T}}\cdot x_i+b-y_i\leqslant\varepsilon+\xi_i^*,\ \xi_i\geqslant 0,\ \xi_i^*\geqslant 0\end{cases}\tag{1}$$

式中：ε 为拟合精度，C 为平衡因子，ξ_i 和 ξ_i^* 为惩罚因子，表示引入训练集的误差，它们可表示样本点超出拟合精度 ε 的程度。LS-SVM 定义了与标准 SVM 不同的损失函数，并将其不等式约束改为等式约束。引入 Lagrange 函数

$$L(\boldsymbol{\omega},b,\xi;\alpha)=J(\boldsymbol{\omega},\xi)-\sum_{i=1}^{n}\boldsymbol{\alpha}_i((\boldsymbol{\omega},\psi(x_i))+b+\xi_i-y_i)\tag{2}$$

式中：$\boldsymbol{\alpha}_i\in\boldsymbol{R}$ 为拉格朗日乘子。

根据最优性条件有

$$\begin{cases}\dfrac{\partial L}{\partial\boldsymbol{\omega}}=0\rightarrow\boldsymbol{\omega}=\sum_{i=1}^{n}\boldsymbol{\alpha}_i\psi(x_i)\\ \dfrac{\partial L}{\partial b}=0\rightarrow\sum_{i=1}^{n}\boldsymbol{\alpha}_i=0\\ \dfrac{\partial L}{\partial\xi_i}=0\rightarrow\boldsymbol{\alpha}_i=\gamma\xi_i\quad i=1,\cdots,N\\ \dfrac{\partial L}{\partial\alpha_i}=0\rightarrow(\boldsymbol{\omega},\psi(x_i))+b+e_i-y_i=0\end{cases}\tag{3}$$

消去 ω,ξ，优化问题转化为求解以下线性方程组

$$\begin{bmatrix}0 & 1_v^{\mathrm{T}}\\ 1_v & Z+I/\gamma\end{bmatrix}\begin{bmatrix}b\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\ \boldsymbol{y}\end{bmatrix}\tag{4}$$

式中：$\boldsymbol{y}=[y_1,\cdots,y_N]^{\mathrm{T}}$，$\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_n]^{\mathrm{T}}$，$1_v=[1,\cdots,1]^{\mathrm{T}}$，$Z_{i,l}=(\psi(x_i),\ \psi(x_l))$。根据 Mercer 条件，内积可由一个核函数 $K(\cdot,\cdot)$ 表示，则有 $Z_{i,l}=K(x_i,x_l)$，其中 $i,l=1,\cdots,N$。

求解式(4)得到 b 和 α 并结合最优条件式(3)，由 LS-SVM 确定的回归函数可表示为

$$f(x)=\sum_{i=1}^{n}\alpha_k K(x,x_k)+b\tag{5}$$

核函数 $K(\cdot,\cdot)$的选择有多种可能[12]，常用的包括线性核函数、多项式核函数、多层感知器核函数和高斯径向基核函数。本文采用应用最为普遍的高斯径向基核函数，其表达式为

$$K(x,x_k)=\exp(-\|x-x_k\|^2/(2\delta)^2)\tag{6}$$

式中：δ 为核宽度。

1.2　基于遗传算法的参数优化

由以上分析可知，LS-SVM 的性能依赖于学习机的参数，通过确定调整参数 γ 和径向基函数的宽度 σ 来确定最优 LS-SVM 模型。因此本文将遗传算法[13]嵌入到 LS-SVM 建模中进行参数寻优，利用遗传算法的全局搜索能力选择最佳的 γ 和 σ。对算法的改进及其实现主要包括以下几个方面：

(1) 适应度函数

由于适应度函数的作用是对个体作评价，因此适应度函数的设计至关重要。针对 LS-SVM 需要优化的参数以及模型的泛化能力，建立以训练样本平均相对误差 (Mean relative error，MRE) $\varepsilon_{\mathrm{MRE1}}$ 的 15% 与测试样本平均相对误差 $\varepsilon_{\mathrm{MRE2}}$ 的 85% 的加权和作为适应度函数，即

$$\mathrm{Fit}(\gamma,\delta)=15\%\varepsilon_{\mathrm{MRE1}}+85\%\varepsilon_{\mathrm{MRE2}}\tag{7}$$

(2) 选择算子

本文选择运算使用比例选择算子，种群个体数为 N，个体 i 的适应度为 f_i，则个体被选取的概率为

$$P_i = f_i / \sum_{k=1}^{N} f_k \tag{8}$$

个体 i 被复制的个数 $Rp(i) = N \times P_i$。从初始化种群中经过选择与复制形成一个子群 $P_1(t)$。

(3) 交叉与变异

参数基因的交叉操作采用线性组合方式，将两个基因串对应交叉位的值相组合生成新的基因串，这样可以保证交叉后产生新的参数值，并开辟出新的搜索空间。变异在遗传操作中属于辅助性的搜索操作，主要目的是维持群体的多样性。较低的变异概率可以防止群体中重要的单一基因丢失，但降低了遗传算法开辟新搜索空间的能力；较高的变异概率将使遗传操作趋于纯粹的随机搜索，降低了算法的收敛速度和稳定性。一般根据具体问题，变异概率取 0.001～0.01 之间的值。

对于采用径向基核函数的 LS-SVM，利用遗传算法进行调整参数 γ 和径向基函数的宽度 σ 的确定。参数优化的具体过程如图 1 所示。

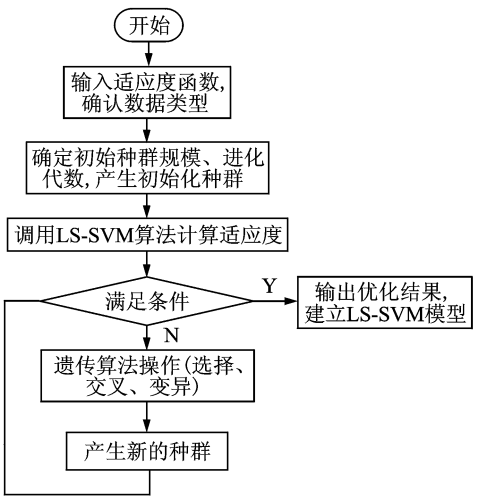

图 1 基于遗传算法的 LS-SVM 的参数优化流程图

Fig. 1 LS-SVM parameters optimization flowchart based on GA

当训练样本、核函数和相关参数确定后，就可以将训练样本集输入 LS-SVM 学习机进行训练，就可求出式(5)中的回归参数 γ，σ，从而得到 LS-SVM 连续波误码率预测模型。

1.3 预测误差函数模型

由信息传输速率 R_b、码元传输速率 R_B 和进制数 M 可得连续波误差数据相互之间的映射关系为

$$R_B = \frac{R_b}{\log_2 M}(B) \tag{9}$$

各种二进制数字调制系统的误码率 P_e 与输入信噪比 r 的数学关系如下

$$Pe_{2ASK} = \frac{1}{2}e^{-\frac{r}{4}};Pe_{2FSK} = \frac{1}{2}e^{-\frac{r}{2}};Pe_{2PSK} = \frac{1}{2}e^{-r} \tag{10}$$

当前样本的预测误差的数据可根据上式已知的数据获得，此时的预测函数关系可表示为

$$Pe = \varphi(A, r, f, M) \tag{11}$$

1.4 数据集选择及异常数据处理

狄克逊准则是通过极差比判定和剔除异常数据。与一般比较简单极差的方法不同，该准则为了提高判断效率，对不同的实验量测定应用不同的极差比进行计算。该准则认为异常数据应该是最大数据和最小数据，因此其基本方法是将数据按大小排队，检验最大数据和最小数据是否异常数据。具体做法如下：

将实验数据 x_i 按值的大小排成顺序统计量 $x_1 \leqslant x_2 \leqslant x_3 \leqslant \cdots \leqslant x_n$，利用给定值计算 f_0 值，然后此给定值将 f_0 与 $f_{(n,a)}$ 进行比较，如果 $f_0 > f_{(n,a)}$，则判定该数据为异常数据，予以剔除。通过对原始数据进行基于狄克逊准则的筛选，可以得到较为理想的样本数据。

利用狄克逊准则对数据进行筛选，将误码率较大的数据视为异常数据剔除。然后利用抽样技术将数据集分为训练集和测试集。通过对式(10)的分析可得到，信噪比对误码率的影响较其他因素大，本文中使用分层抽样[14]的方法，按照信噪比的区间进行抽样得到训练集和测试集。

2 泥浆连续波误码率预测模型对比实验及结果分析

本文实验条件为井深从 1～8 km，钻井液粘度从 10～50 mPa·s，钻管内径从 126 mm(套管尺寸 5 in)～660 mm(套管尺寸 26 in)，最高载波频率从 20～36 Hz，分别计算地面接收端压力波信号幅值，选择接收信号幅值大于 1 kPa 的各种传输方式，用风洞模拟[15]发送并接收数据试验。对接收并存储的数据用 Matlab 处理，并施加不同频率的干扰噪声信号，再用 Matlab 仿真接收并统计其误码率，利用狄克逊准则对数据进行筛选，筛选结果如图 2 所示。

由图 2 可得，通过狄克逊准则对误码率进行数据筛选，可以将误码率较高的部分数据清除，得到

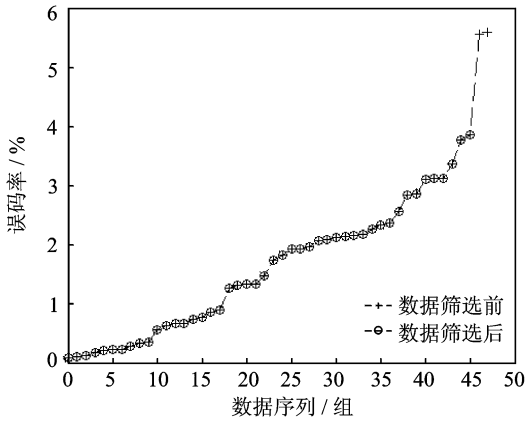

图 2　利用狄克逊准则筛选误码率数据

Fig. 2　Screening data error rate using Dixon criterion

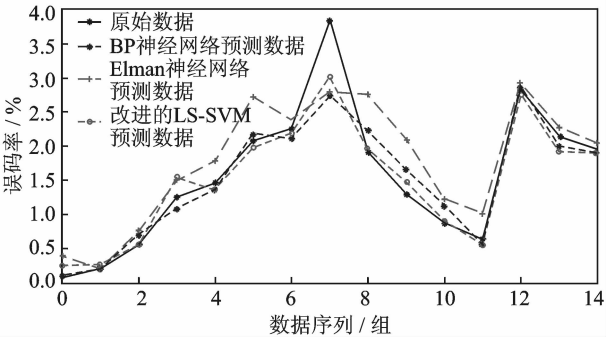

图 3　原始数据、BP 网络、Elman 网络和改进 LS-SVM 数据预测对比图

Fig. 3　Data forecasting contrast figure of original data, BP network, Elman network and improved LS-SVM

误码率较低、较为理想的样本数据，为模型的建立做好准备。

本文中将误差反传前馈（Back propagation，BP）神经网络、Elman 神经网络和基于 GA 的 LS-SVM 建立泥浆连续波误码率预测模型，进行仿真对比。将筛选后的 29 组数据作为训练样本，建立各自的误差预测模型，然后利用其余 14 组数据作为检验样本，通过 Matlab 得出的误码率与实际误码率以进行对比。其中 BP 神经网络使用三层结构网络，隐含层采用 S 型传递函数，输出层采用线性传递函数；Elman 神经网络使用两层结构网络，隐含层函数使用 Tansig 传递函数，输出层使用 Purelin 传递函数；基于 GA 的 LS-SVM 回归算法中，利用遗传算法得到调整参数 γ 和径向基函数的宽度 σ，其中种群大小 $N=20$，进化最大代数为 100。为便于比较各模型对训练样本的回归精度，定义均方误差（Mean squared error，MSE）为检验指标

$$\mathrm{MSE}=\sum_{i=1}^{n}(P_i^*-P_i)^2/n \qquad (12)$$

式中：P_i^* 为模型输出；P_i 为训练样本集输出；n 为训练样本个数。

从图 3 各模型的预测值对比结果可以看出，改进型 LS-SVM 预测模型的预测能力要强于 BP 神经网络和 Elman 神经网络，相对于神经网络 LS-SVM 具有较高的回归精度，这一点也是由于本实验样本相对较少，而 LS-SVM 是利用结构风险最小化准则来代替传统的经验风险最小化准则，因而更适合小样本情况下的学习问题。

预测结果表明，BP 神经网络模型的预测均方误差为 0.482 2，Elman 模型为 0.679 5，LS-SVM 模型为 0.313 3。Elman 神经网络的预测数据趋势与原始数据的差别较大，BP 神经网络虽然趋势基本一致，但预测均方误差值较大，而改进型 LS-SVM 预测模型在以上 3 个模型中对泥浆连续波误码率预测效果最好，预测值也更接近于实际的值。

虽然改进型 LS-SVM 模型相对于神经网络来说能够较好的预测出泥浆连续波数据的误码率，但还是存在一定的误差。分析其原因有：

（1）数据量相对较少，在将数据随机分为训练样本和检测样本的过程中会出现误差。

（2）利用遗传算法对 LS-SVM 模型建立过程中的参数求解存在一定的误差，只是趋近于最佳值。

3　结束语

本文通过对最小二乘支持向量机进行理论分析，利用狄克逊准则对样本数据进行处理，并提出了基于遗传算法改进的 LS-SVM 泥浆连续波误码率预测模型的步骤和方法，并经过与 BP 神经网络和 Elman 神经网络预测方法的对比，通过 Matlab 仿真得出基于遗传算法的 LS-SVM 在小样本情况下，它比 BP 和 Elman 神经网络预测模型具有更强的泛化和预测能力，模型预测精度较高，可以用于泥浆连续波误码率预测。

参考文献：

[1]　苏义脑，窦修荣. 随钻测量、随钻测井与录井工具[J]. 石油钻采工艺，2005，27(1)：74-78.
Su Yinao, Dou Xiurong. Measurement while drill-

ing, logging while drilling and logging instrument [J]. Oil Drilling & Production Technology, 2005,27 (1):74-78.

[2] 房军, 苏义脑. 液压信号发生器基本类型与信号产生的原理[J]. 石油钻探技术,2004,32(2):39-41.
Fang Jun, Su Yinao. The basic types and it's mechanism of the hydraulic signal producer[J]. Oil Drilling & Production Technology, 2004,32(2):39-41.

[3] 张伟,师奕兵,卢涛. 小波神经网络在无线随钻测量系统在泥浆信号检测中的应用研究[J]. 电子测量与仪器学报,2008,22(6):43-46.
Zhang Wei, Shi Yibing, Lu Tao. Research on application of wavelet neural network to mud signal detection in wireless measurement while drilling[J]. Journal of Electronic Measurem Entand Instrument, 2008,22(6):43-46.

[4] 赵建辉,王丽艳,盛利民,等. 去除随钻测量信号中噪声及干扰的新方法[J]. 石油学报,2008,29(4):596-600.
Zhao Jianhui, Wang Liyan, Sheng Limin, et al. A nonlinear method for filtering noise and interference of pulse signal in measurement while drilling[J]. Acta Petrolei Sinica, 2008,29(4): 596-600.

[5] 陈后金,袁保宗,Douglas Baxter. 生物神经信号采集与模式识别[J]. 数据采集与处理,2007, 22(1):38-41.
Chen Houjin, Yuan Baozong, Douglas Baxter. Biological neural signal collecting and pattern recognition [J]. Journal of Data Acquisition and Processing, 2007, 22(1):38-41.

[6] 苟博,黄贤武. 支持向量机多类分类方法[J]. 数据采集与处理,2006,21(3):334-339.
Gou Bo, Huang Xianwu. SVM multi-class classification[J]. Journal of Data Acquisition and Processing, 2006,21(3):334-339.

[7] 杨一文,杨朝军. 基于支持向量机的金融时间序列预测[J]. 系统工程理论方法应用,2005,14(2):176-181.
Yang Yiwen, Yang Zhaojun. Financial time series forecasting based on support vector machine[J]. Systems Engineering Theory Method Applacations, 2005,14(2):176-181.

[8] Li Yanhua, He Chunhua, Li Bingjun, et al. An effecient computational model for LS-SVM and its applications in time series prediction [C]// IEEE :ICCSE. Hefei,China:IEEE,2010:24-27,467-470.

[9] 王瑞,罗飞,杨红,等. 基于贝叶斯回归 LS-SVM 的非线性系统观测[J]. 自动化与仪表, 2011, 26(7):5-9.
Wang Rui, Luo Fei, Yang Hong, et al. Observer design for nonlinear systems based on regression LS-SVM with bayesian methods[J]. Automation & Instrumentation, 2011, 26(7):5-9.

[10] 姚全珠,蔡婕. 基于 PSO 的 LS-SVM 特征选择与参数优化算法[J]. 计算机工程与应用, 2010,46(1): 134-136.
Yao Quanzhu, Cai Jie. Feature selection and LS - SVM parameters optimization algorithm based on PSO[J]. Computer Engineering and Applications, 2010,46(1):134-136.

[11] 郑水波, 韩正之, 唐厚君, 等. 最小二乘支持向量机在汽车动态系统辨识中的应用[J]. 上海交通大学学报, 2005, 39(3):392-395.
Zheng Shuibo, Han Zhengzhi, Tang Houjun, et al. Application of LS-SVM in the automobile dynamical system identification[J]. Journal of Shanghai Jiaotong University, 2005, 39(3):392-395.

[12] Colombo G. A genetic algorithm for frequency assignment with problem decomposition [J]. International Journal of Mobile Network Design and Innovation, 2006, 1(2):102-112.

[13] 邵婷婷,张水利,张永波. 两种剔除异常数据的方法比较[J]. 现代电子技术. 2008, 31(24): 148-150.
Shao Tingting, Zhang Shuili, Zhang Yongbo. Comparison of two methods in eliminating the excrescent data[J]. Modern Electronics Technique, 2008, 31 (24): 148-150.

[14] Choi B Y, Park J, Zhang Z L. Adaptive packet sampling for accurate and scalable flow measurement[C] // IEEE Globecom Telecommunications Conference. Dallas, USA:IEEE, 2004:1448-1452.

[15] 艾延廷,黄福幸,李国文. DFD 风洞数据采集与控制系统设计[J]. 仪器仪表学报,2005,26(增刊):821-823.
Ai Yanting, Huang Fuxing, Li Guowen. Design of the data acquisition and control system for DFD wind tunnel[J]. Chinese Journal of Scientific Instrument. 2005,26(S):821-823.

作者简介:刘新平(1966-),男, 副教授,研究方向:数字信号处理、智能控制等,E-mail:liuxinp@upc.edu.cn;薛希文(1987-), 男,硕士研究生,研究方向:智能控制技术。