DOI: 10.19650/j.cnki.cjsi.J2107724

考虑样本异常值的改进最小二乘支持向量机算法*

付乐天[1,2],李 鹏[1,2],高 莲[1,2]
(1. 云南大学信息学院 昆明 650091; 2. 云南省高校物联网技术及应用重点实验室 昆明 650091)

摘 要:针对最小二乘支持向量机对异常值敏感、缺乏鲁棒性的情况,提出一种考虑样本异常值的改进最小二乘支持向量机算法。该算法首先通过采用局部异常因子检测算法为每个数据样本计算一个 LOF 因子,根据其因子值能够有效地将样本分成正常样本和异常样本,然后针对不同样本进行单独设置样本权重。其有效地保证了在降低异常样本权重的同时而不使正常样本权重受到影响,使最小二乘支持向量机在达到目标函数最优化的同时能够保证正常数据信息不丢失,以提高模型的鲁棒性。最后,通过引入"信息熵"和"平均粒距"来改进粒子群算法,将其应用于模型的参数优化。经过实验仿真表明,该算法能够有效地提高模型的鲁棒性,随着异常样本的增多,其模型精度提高大约 67%。

关键词: 改进最小二乘支持向量机;局部异常因子检测算法;改进粒子群优化算法
中图分类号: TP18 TH165.3 **文献标识码**: A **国家标准学科分类代码**: 510.80

Improved LSSVM algorithm considering sample outliers

Fu Letian[1,2], Li Peng[1,2], Gao Lian[1,2]
(*1. School of Information, Yunnan University, Kunming 650091, China*;
2. Internet of Things Technology and Application Key Laboratory of Universities in Yunnan, Kunming 650091, China)

Abstract: Aiming at the situation that least squares support vector machine is sensitive to outliers and lacks robustness, an improved least squares support vector machine algorithm considering sample outliers is proposed. The algorithm first calculates a LOF for each data sample using the local outlier factor detection algorithm, and can effectively divide the samples into normal and abnormal samples according to their factor values, and then separately set sample weights for different samples. The algorithm effectively ensures that the weight of abnormal samples is reduced while the weight of normal samples is not affected, so that the least squares support vector machine can achieve the optimization of the objective function while ensuring that the normal data information is not lost, so as to improve the robustness of the model. Finally, "information entropy" and "average particle distance" are introduced to improve the particle swarm algorithm, which is applied to the parameter optimization of the model. Experiment simulation shows that the algorithm can effectively improve the robustness of the model. With the increase of abnormal samples, the accuracy of the model is improved by about 67%.

Keywords: improved least square support vector machines; local outlier factor detection algorithm; improved PSO algorithm

0 引 言

支持向量机(support vector machines, SVM)是 Cortes[1]于 1995 提出的,是建立在统计学习的 VC 维理论和结构风险最小原理基础上,其根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷,以期望获得最好的泛化能力。而最小二乘支持向量机[2](least square support vector machines, LSSVM)将支持向量机的不等式约束转化为等式约束,简化了计算的复杂性,加快了运算速度。它已成功应用于实际的许多模式识别问题,如故障预测[3]、故障检测[4]、图像分类[5]、偏微分方程求解[6]、视觉跟踪[7]。它通过求解一个线性方程组来得到解,而不是采用传统支持向量机方法来直接求解一个

收稿日期:2021-04-08 Received Date: 2021-04-08
*基金项目:国家自然科学基金(61763049)、云南省应用基础研究重点课题(2018FA032)、中青年学术和技术带头人后备人才项目(20205AC160115)资助

二次规划问题。因此,LSSVM 的训练比其他支持向量机的训练更简单。

然而,最小二乘支持向量机对异常值很敏感,这些异常值拥有很大的拉格朗日乘数值,这就导致异常值对决策函数构造的影响比其他样本要大,存在的异常值将会使模型的建模精度受到影响。为了克服异常值的敏感性问题,人们采取的方法一般是将样本中的异常值寻找出来,然后根据该异常值所在区域的局部线性关系来推导该异常值的估计值,然后使用估计值来代替该异常值。黄贤源等[8]通过将趋势面滤波法与 LSSVM 算法相结合来构造海底的趋势面,然后基于优化训练样本的海底趋势面针对异常值进行检测,最后通过异常点的平面坐标推导的估计值来代替异常值,对异常值进行剔除。郭战坤等[9]通过插值或者是应用 LSSVM 中的预测值来修正原始序列中的异常值并将其应用到港口的集装箱吞吐量预测中。但是,此类方法却存在数据信息丢失的风险。样本中的异常值也包含有样本的数据信息,由局部线性关系所推导的估计值来代替此异常值将会导致样本的数据信息丢失,并且还会存在估计值也不一定反映此异常样本所代表的实际信息的问题。因此,不少学者采用通过为样本数据设立权重的方法来代替用估计值替代异常值的方法。通过对异常样本数据设立权重,能够减小异常样本的影响程度,同时能够保留样本的数据信息,避免数据信息丢失。

Suykens 等[10]提出了一种加权最小二乘支持向量机(weighted least square support vector machine, WLSSVM)模型,将较小的权值分配给不太重要的样本和异常值,以减少它们对模型的影响。此外,还提出了其他几种设定权重的策略[11-12]。Zhang 等[13]在 Suykens 等提出的模型基础上考虑到实际应用中噪声和异常值可能不都是服从高斯分布,引入改进的正态分布规则来重构加权权重,提出了一种基于改进的正态分布的加权 LSSVM,以削弱异常数据和冗余数据的影响。Chen 等[14]为了消除训练中异常值对模型预测性能的影响,提高模型的鲁棒性,并考虑到 Suykens 等提出的模型都需要预先求解原 LSSVM 来设置权重并且假设的是误差变量服从高斯分布,因此提出了自适应加权技术,并将其引入到标准 LSSVM 模型中。然而,这些方法无法确定哪些样本数据为异常值,并针对性的进行样本加权权重的构造,其是根据模型误差的统计特性,通过数据的误差或者均值来为每一个样本计算一个权重,无法针对异常数据单独进行权重的设立。然而,根据模型误差的统计特性设立权重可能会使正常样本的权重受到影响。因此,该方面仍有改进的空间。

同时,LSSVM 模型的精度不仅受异常值的影响,还受到其参数选取的影响。针对模型的参数优化,常采用的方法有蝙蝠算法[15]、粒子群算法[16]、遗传算法[17]等。粒子群算法作为一种群体智能优化算法,常用于模型的参数优化过程中,其是通过模拟生物的觅食活动来寻找模型参数的最优值,通过设立惯性权重和学习因子来改变生物的觅食状态,以求能够通过不断的迭代过程寻找到参数的最优值。但是,粒子群算法存在易陷入局部极小值的问题,容易导致寻到参数值并不是全局最优值,使用此参数值建模将会导致模型精度受到影响。因此,如何避免在寻优过程中陷入局部极小值,让粒子在更大的空间中寻优,将会是一个值得思考的问题。

所以,为了提高模型的鲁棒性,减小异常数据的影响,同时优化参数的选取,提高模型的精度,本文提出一种基于局部异常因子检测算法的改进 LSSVM 算法。该算法首先通过采用局部异常因子检测方法对样本数据进行异常值检测,为每个样本数据计算一个 LOF 因子。因为当 LOF 因子越大时,则越能表明该数据为异常值。然后,依据每个样本的 LOF 因子值和所设定的阈值判断样本是否是异常样本,将大于阈值的样本认定为异常样本,小于阈值的样本认定为正常样本。随后,由加权权重公式计算异常样本和正常样本的权重值,并应用这些权重来建立加权最小二乘支持向量机模型。最后,通过引入“熵”和“平均粒距”来改进粒子群算法,以求扩大粒子寻优的范围,避免其陷入局部极小值,并将改进后的粒子群算法用于对建立好的加权最小二乘支持向量机模型的参数寻优当中,以求提高模型面对异常值时的建模精度,增强模型的鲁棒性。

1 基本方法

1.1 局部异常因子检测算法

局部异常因子(local outlier factor, LOF)检测方法是一种无监督的异常检测算法,通过对每个点计算一个异常因子 LOF 来表征样本数据是否为异常样本数据。当 LOF 值越高时,其所代表的样本数据为异常点的可能性越大,其主要步骤如下:

1)计算点 p 的第 k 距离 $d_k(p)$,其定义为 $d_k(p)=d(p,o)$,满足:

(1)在集合中至少有不包括 p 在内的 k 个点 $o'\in C\{x\neq p\}$,有 $d(p,o')\leqslant d(p,o)$;

(2)在集合中最多有不包括 p 在内的 $k-1$ 个点 $o'\in C\{x\neq p\}$,有 $d(p,o')<d(p,o)$;

2)计算点 p 的第 k 距离邻域 $N_k(p)$,满足 p 的第 k 邻域点的个数 $|N_k(p)|\geqslant k$;

3)计算点 o 到点 p 的第 k 可达距离为:

$$d_k(p,o)=\max\{d_k(o),d(p,o)\} \tag{1}$$

4)计算点 p 的局部可达密度:

$$lrd_k(p)=1\Big/\left(\frac{\sum_{o\in N_k(p)}d_k(p,o)}{|N_k(p)|}\right) \tag{2}$$

5)计算点 p 的局部异常因子:

$$lof_k(p)=\frac{\sum_{o\in N_k(p)}lrd_k(o)}{|N_k(p)|}\Big/lrd_k(p) \tag{3}$$

1.2　最小二乘支持向量机

最小二乘支持向量机模型通过采用最小二乘原理和等式约束将标准支持向量机模型中求解二次规划的问题转化为求解线性方程组的问题,大大方便了模型的建立。

定义一个非线性变换 $\boldsymbol{\Phi}(x)$,将 n 维输入、一维输出样本向量 $\{(\boldsymbol{x}_k,\boldsymbol{y}_k)\}_{k=1}^{N},\boldsymbol{x}_k\in\boldsymbol{R}^n,\boldsymbol{y}_k\in\boldsymbol{R}$ 由原来低维空间映射到高维空间,并构建最优线性回归函数如式(4)所示:

$$\boldsymbol{y}(x)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{x})+\boldsymbol{b} \tag{4}$$

式中:$\boldsymbol{w}$ 为权向量;b 为阈值。

LSSVM 将 SVM 的不等式约束问题转化为等式约束,如式(5)所示:

$$\begin{aligned}&\min J=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{C}{2}\sum_{i=1}^{l}e_i^2>0;\\&\text{s.t. }y_i=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}(x_i)+b_i+e_i;i=1,2,\cdots,l\end{aligned} \tag{5}$$

引入拉格朗日乘子 a_i,式(5)的优化问题可以对偶表示:

$$L=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{C}{2}\sum_{i=1}^{l}e_i^2-\sum_{i=1}^{l}a_i\{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}(x_i)+b_i+e_i-y_i\} \tag{6}$$

根据 KKT 条件,求 L 对 $\boldsymbol{w}$、b、e_i、a_i 的偏导数等于 0,可以得到式(7)所示的线性方程组:

$$\begin{aligned}&\frac{\partial L}{\partial w}=0\rightarrow\boldsymbol{w}=\sum_{i=1}^{l}a_i\boldsymbol{\Phi}(x_i)\\&\frac{\partial L}{\partial b_i}=0\rightarrow\sum_{i=1}^{l}a_i=0\\&\frac{\partial L}{\partial e_i}=0\rightarrow a_i=Ce_i;i=1,2,\cdots,l\\&\frac{\partial L}{\partial a_i}=0\rightarrow\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}(x_i)+b_i+e_i-y_i=0\end{aligned} \tag{7}$$

消去式(7)中的 $\boldsymbol{w}$、e_i,转化为如式(8)所示的线性系统:

$$\begin{bmatrix}0&\boldsymbol{s}^{\mathrm{T}}\\\boldsymbol{s}&\boldsymbol{K}+C^{-1}\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{b}\\\boldsymbol{a}\end{bmatrix}=\begin{bmatrix}0\\\boldsymbol{y}\end{bmatrix} \tag{8}$$

式中:$\boldsymbol{y}=[y_1,y_2,\cdots,y_l]^{\mathrm{T}}$;$\boldsymbol{I}$ 为单位矩阵;$\boldsymbol{a}=[a_1,a_2,\cdots,a_l]^{\mathrm{T}}$;$\boldsymbol{b}=[b_1,b_2,\cdots,b_l]^{\mathrm{T}}$;$\boldsymbol{s}=[1,1,\cdots,1]^{\mathrm{T}}$;$K(\cdot)$ 为核函数,$K(x_i,x_j)=\boldsymbol{\Phi}(x_i)^{\mathrm{T}}\boldsymbol{\Phi}(x_j)$。

用最小二乘法求出 $\boldsymbol{a}$ 和 $\boldsymbol{b}$,可以得到 LSSVM 的函数估计为:

$$y(x)=\sum_{i=1}^{l}a_i\boldsymbol{K}(x,x_i)+\boldsymbol{b} \tag{9}$$

2　改进的 LSSVM 算法

本文中,旨在解决的问题是在加权最小二乘支持向量机模型建立过程中因为存在异常样本而导致的模型精度不够的问题。基于主要问题,提出了一种改进的最小二乘支持向量机算法来解决问题。因此,本部分主要介绍所提出的改进的最小二乘支持向量机算法及其理论分析。

首先,针对含有异常值的样本数据集 X 采用局部异常因子检测算法为每个样本数据计算一个 LOF 因子。然后,依据计算所得的 LOF 因子和所设定的阈值判断该样本是否为异常样本,将超过设定阈值的 LOF 因子所代表的样本归为异常样本,小于阈值的为正常样本。如此,将含有异常值的样本数据 X 有效地分为异常样本 X_O 和正常样本 X_N。其次,由加权公式计算正常样本 X_N 和异常样本 X_O 的样本权重并建立加权 LSSVM 模型。最后,运用改进的粒子群算法对于所建立的模型进行参数的优化。

以下为该算法主要创新点“基于局部异常因子的样本权重设定”以及“基于信息熵和平均粒距的粒子群优化”的详细描述及理论分析,包括与相关算法的原理比较。

2.1　基于局部异常因子的样本权重设定

1)样本权重设定算法

将含有异常样本的数据集划分为正常样本和异常样本,然后分别进行样本权重的设立,能够保证在降低异常样本权重的同时不影响正常样本权重,将能有效地解决问题。

该过程首先是采用局部异常因子检测算法计算每个样本的 LOF 因子,然后依据阈值和 LOF 因子判断该样本是否为异常样本,最后由权重公式对样本的权重进行设定。文中,样本权重设定过程如下:

LSSVM 的等式约束问题如式(5)所示,通过引入样本权重后将式(5)变为式(10)所示。

$$\begin{aligned}&\min J=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{C}{2}\sum_{i=1}^{l}\lambda_ie_i^2>0;\\&\text{s.t. }y_i=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}(x_i)+b_i+e_i;i=1,2,\cdots,l\end{aligned} \tag{10}$$

式中:λ 为该样本数据的样本的权重,表示该样本的影响程度。

由拉格朗日乘子法得到:

$$L=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{C}{2}\sum_{i=1}^{l}\lambda_ie_i^2-\sum_{i=1}^{l}a_i\{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}(x_i)+b_i+e_i-y_i\} \tag{11}$$

为了得到 a_i 和 b_i，根据 KKT 条件可以得到：

$$\begin{bmatrix} 1 & K(x_1,x_1)+\dfrac{1}{C\lambda_1} & \cdots & K(x_1,x_l) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & K(x_l,x_1) & \cdots & K(x_l,x_l)+\dfrac{1}{C\lambda_l} \end{bmatrix} \begin{bmatrix} b \\ a_1 \\ \vdots \\ a_l \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ \vdots \\ y_l \end{bmatrix} \tag{12}$$

其中，λ_i 的公式如式(13) 所示。

$$\lambda_i = \begin{cases} 1, lof < \nu \\ \exp\left(\dfrac{\mu-\varepsilon}{\mu}\right), lof > \nu \end{cases} \tag{13}$$

式中：μ 为该异常样本周围的 5 个相邻样本的均值；ε 为异常样本与 μ 的差的绝对值；lof 为每个样本的 LOF 因子值；ν 为设定的阈值。

2)算法分析

面对含有异常值的样本建模，通过对样本设立权重，可以降低异常样本的影响并保留样本的数据信息。其中，Suykens 等提出的 WLSSVM 算法的加权权重构造如下所示：

$$v_i = \begin{cases} 1, & |\varepsilon_i/\hat{s}| \leqslant c_1 \\ \dfrac{c_2-|\varepsilon_i/\hat{s}|}{c_2-c_1}, & c_1 \leqslant |\varepsilon_i/\hat{s}| \leqslant c_2 \\ 10^{-4}, & \text{其他} \end{cases} \tag{14}$$

式中：$\hat{s}$ 是对误差标准差的鲁棒性估计；ε_i 为误差；$c_1 = 2.5$，$c_2 = 3$；v_i 为样本权重。

WLSSVM 的权重设立是基于误差服从高斯分布的，但是实际中误差不一定服从高斯分布，因为考虑到此问题，Zhang 等[13]和 Chen 等[14]在 WLSSVM 的基础上提出了基于改进正态分布的加权 LSSVM(weighted least square support vector machine based on improved normal distribution, INDWLSSVM)和自适应加权 LSSVM(adaptive weighted least square support vector machine, AWLSSVM)。其中 INDWLSSVM 加权系数公式如下所示：

$$v_i = \begin{cases} \exp\left(\dfrac{(|\varepsilon_i|-\mu)^2}{u_1 s^2}\right) & |\varepsilon_i| < \mu \\ \exp\left(\dfrac{(|\varepsilon_i|-\mu)^2}{u_2 s^2}\right) & |\varepsilon_i| \geqslant \mu \end{cases} \tag{15}$$

式中：μ 为 $|\varepsilon_i|$ 的均值；s 为 $|\varepsilon_i|$ 的标准差；$u_1 = 9.7$，$u_2 = 7.6$；ε_i 为误差。

AWLSSVM 的加权权重如下所示：

$$v_i = \exp\left(\frac{-(|\xi_i|-\mu)^2}{us}\right) \tag{16}$$

式中：ξ_i 为误差；μ 为 $|\xi_i|$ 的均值；s 为 $|\xi_i|$ 的标准差；u 为调整系数。

由式(15)可知，INDWLSSVM 的权重公式的指数部分大于 0：

$$\frac{(|\varepsilon_i|-\mu)^2}{u_1 s^2} > 0 \tag{17}$$

$$\frac{(|\varepsilon_i|-\mu)^2}{u_2 s^2} > 0 \tag{18}$$

因此，由式(15)计算 INDWLSSVM 的权重的时候，计算得到的样本权重大于 1，如式(19)：

$$v_i > 1 \tag{19}$$

又因为，LSSVM 在原始权值空间中的优化目标如式(20)所示：

$$\min J = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{l} e_i^2 > 0 \tag{20}$$

而加权 LSSVM 在原始权值空间中的优化问题如式(21)所示：

$$\min J = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{l} v_i e_i^2 > 0 \tag{21}$$

当由式(15)计算样本权重时，会使所得的正常样本权重大于 1，将其带入式(21)中时，会使正常样本的惩罚项的值偏大，如式(22)所示：

$$\frac{C}{2}\sum_{i=1}^{l} v_i e_i^2 > \frac{C}{2}\sum_{i=1}^{l} e_i^2 \tag{22}$$

而对于同一个确定的样本来说，其几何间隔是保持不变的。因此，由式(15)计算样本权重时，将导致优化目标的值偏大，无法得到最优。

由式(16)可知，AWLSSVM 权重公式的指数部分小于 0：

$$\frac{-(|\xi_i|-\mu)^2}{us} < 0 \tag{23}$$

因此，由式(16)计算 AWLSSVM 的权重的时候，计算得到的样本权重小于 1，如式(24)所示：

$$v_i < 1 \tag{24}$$

而采用 k-means 的聚类最小二乘支持向量机[18]其样本权重的表达式如式(25)所示：

$$\beta_j = e^{-\frac{D}{\sigma^2}} \tag{25}$$

式中：D 为第 j 个样本到其 m 个最近邻点的欧式距离之和；参数 σ 和 m 是由经验或优化确定的。

其是通过相对密度来计算得到样本权重的，由此计算样本权重时，也会存在指数小于 0 的情况：

$$-\frac{D}{\sigma^2} < 0 \tag{26}$$

因此，由式(25)计算得到的样本权重小于 1：

$$\beta_j < 1 \tag{27}$$

而加权 LSSVM 通过特定的方法制定样本的权重,赋予样本不同的权重,其代表了样本在模型中的贡献度。而通过式(16)和(25)计算样本的权重将会导致在降低异常样本权重的同时降低了正常样本权重,不利于模型的构建,有可能会造成模型预测精度偏低。

因此,本文基于局部异常因子检测算法来计算样本权重能够有效避免样本权重偏大或者偏小的情况,如式(13)所示,当样本为正常样本时,其样本权重由式(28)计算;当样本为异常样本时,其样本权重由式(29)计算:

$$\lambda_{正} = 1, lof < v \tag{28}$$

$$\lambda_{异} = \exp\left(\frac{\mu - \varepsilon}{\mu}\right), lof > \nu \tag{29}$$

式中:ν 为所设定的阈值,表明当样本的 LOF 因子大于阈值时,为异常样本;小于阈值时,为正常样本。

因此,由基于局部异常因子检测算法得到的样本权重其异常样本的权重都小于 1,即;

$$\lambda_{异} = \exp\left(\frac{\mu - \varepsilon}{\mu}\right) < 1, lof > \nu \tag{30}$$

而正常样本的权重都等于 1,如式(28)所示。由此,保证了在降低异常样本权重的同时不会降低正常样本的权重,避免出现如式(16)和(25)的情况。同时,因为正常样本权重都等于 1,在对样本数据进行建模时,其正常样本将不会出现如式(22)所示的情况。

2.2　基于信息熵和平均粒距的粒子群优化算法及分析

最小二乘支持向量机是在支持向量机的基础上发展而来的,因此最小二乘支持向量机模型参数的作用可以通过观察支持向量机的模型参数的作用得到。在文献[1]中,为了解决样本数据存在异常值时对构建超平面的影响,从而引入了松弛变量,即如:

在考虑到样本异常值问题时,支持向量机的约束条件变为了如式(31)所示:

$$y_i(\boldsymbol{\omega}^{\mathrm{T}} x_i + \boldsymbol{b}) \geqslant 1 - \xi_i \tag{31}$$

其中,$\xi_i \geqslant 0$ 称为松弛变量,即为对应的样本数据点允许偏离超平面的量。

如果当存在 ξ_i 任意大的时候,那么任意的超平面都是符合条件的了。因此,为了在面对异常值时找到适合的超平面,所以在原来的目标函数的后面加上一项,使得这些 ξ_i 的总和也要最小,如式(32)所示:

$$\min \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega} + Y\sum \xi_i \tag{32}$$

其中,Y 为一个参数,代表寻找到最大间隔超平面和保证数据点偏差量最小之间的权重,也即为模型所需优化的参数值,其与样本数据是否存在异常值息息相关。当样本数据不存在异常偏离点或异常值较少时,则着重于寻找到最大间隔的超平面,其参数值就较小;当样本数据存在较多异常偏离点时,则着重于保证数据点偏差量较小,其参数值就较大。

由此可见,模型参数的选取与样本异常值的影响息息相关,在面对有异常值的样本时选择准确的模型参数,有助于寻找到准确的超平面,确保目标函数最小化。

针对传统的粒子群算法因为容易陷入局部极小值的问题,本文提出了一种改进的粒子群算法来用于模型参数的优化。本文引入“信息熵”和“平均粒距”的概念来描述粒子的某时刻所处的状态,针对不同状态的粒子采取不同的更新模式。

信息熵是信息论中用于度量信息的一个概念。1948 年,香农将统计物理中的熵概念引入到信道通信过程中,定义了信息熵的概念。根据信息熵的定义可知,信息熵代表了某种特定信息出现的概率,即当一种信息出现概率更高时,表明它被传播的更广泛。因此,当一个系统越是有序时,其出现变量的不确定性就越小,把其弄清楚所需要的信息量也就越小,其信息熵就越低;反之,一个系统越是混乱,其出现变量的不确定性就越大,把其弄清楚所需要的信息量也就越大,其信息熵就越高。本文依据“信息熵”的概念来描述粒子在搜索空间中的混乱程度。当熵值越高的时候,代表粒子在搜索空间中越混乱,其相应的分布也就越发的广泛,其粒子越不易陷入局部极小值。本文定义的熵值公式如式(33)所示。

$$S = -K\sum_{i=1}^{20}(\Theta_i(x) * \log_2(\Theta_i(x))) \tag{33}$$

式中:$\Theta_i(x)$ 表征每一个粒子在搜索空间中出现的相对位置;K 为常数。

平均粒距[18]表达了种群各个个体相互之间的分布离散程度。因此,本文采用平均粒距来描述种群的多样性,避免陷入局部极小值,其如式(34)所示。

$$D(t) = \frac{1}{oL}\sum_{i=1}^{o}\sqrt{\sum_{d=1}^{\Psi}(a_{id} - \overline{a_d})^2} \tag{34}$$

式中:o 为种群数;Ψ 为种群维数;L 是搜索空间对角线最大长度;a_{id} 为粒子 i 的 d 维坐标;$\overline{a_d}$ 为所有粒子第 d 维的均值。由上式可知,D 越小,种群分布越不均匀。

在每次迭代中,根据式(33)、(34)计算粒子的熵和平均粒距,判断在当前迭代时的粒子的状态。其值大于设定最大值时,说明粒子在搜索空间中较为分散,不易陷入局部极小值,则粒子为“活跃状态”,粒子正常更新速度和位置;当其值小于设定最小值时,粒子在搜索空间中团聚在一起,极易陷入局部极小值,则粒子为“惰性状态”,则粒子更新时采用较大的惯性权重来更新粒子的速度和位置。其主要步骤如下:

1)初始化粒子的速度和位置,设定粒子的种群数量、最大迭代次数、学习因子 $C1$ 和 $C2$、熵的最大值与最小值、平均粒距的最大值与最小值。

2)计算粒子的适应度值,寻找粒子适应度的个体极值和群体极值。

3)在每一次的迭代过程中,根据式(33)、(34)计算粒子的熵和平均粒距。

4)由计算得到的熵和平均粒距来判断粒子的状态,将粒子的状态分为"活跃状态"和"惰性状态"。

5)如果粒子状态为"活跃状态",则采用较小的惯性权重更新粒子的速度和位置;如果粒子的状态为"惰性状态",则采用较大的惯性权重更新粒子速度和位置。

6)由更新后的粒子的速度和位置计算粒子的适应度值,并寻找粒子适应度的个体极值和群体极值。

7)判断是否满足迭代的条件。如果满足,则结束迭代,返回寻找到的个体极值和群体极值;如果不满足,则返回步骤3)继续迭代寻优。

8)迭代结束,返回寻找到粒子的群体极值和个体极值。

通过在每一次迭代过程中由式(33)和(34)来判断粒子的状态,并针对不同的状态来更新粒子,将有利于使粒子在整个搜索空间中分布的更加广泛,而不会陷入局部极小值中。

本文提出的改进的粒子群算法的参数设置如表1所示。种群规模为20,最大迭代次数为100,学习因子 $C1$ 和 $C2$ 通常设为2。惯性权重因子 w 代表粒子的寻优能力,其值越大,表明粒子全局寻优能力强,局部寻优能力弱;反之亦然。因此,对于处于"活跃状态"的粒子需要加强其局部寻优能力,以便于搜索全局最优值,需将惯性因子设立的较小;对于处于"惰性状态"的粒子需要避免陷入局部极小值,应该加强其全局寻优能力,应将惯性因子设立的较大。所以,对于"活跃状态"的惯性因子设为0.9,对于"惰性状态"的惯性因子设为1.5。其次,经过多次实验,选取熵值的上、下限分别为2和0.25,平均粒距的上、下限为0.25和0.01。

表1 参数设置

Table 1 Parameter setting

参数	设置
种群数量	20
最大迭代次数	100
$C1$	2
$C2$	2
S_{max}	2
S_{min}	0.25
D_{max}	0.25
D_{min}	0.01
活跃状态 w	0.9
惰性状态 w	1.5

3 实验及结果分析

为了有效地评估模型性能的好坏,本文采用均方根误差(RMSE)作为评估模型性能的指标并通过数值仿真实验和轴承仿真实验来进行验证,其如式(35)所示。

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y'-y)^2} \tag{35}$$

式中:n 为样本数量;y' 为预测值;y 为真实值。

3.1 数值仿真

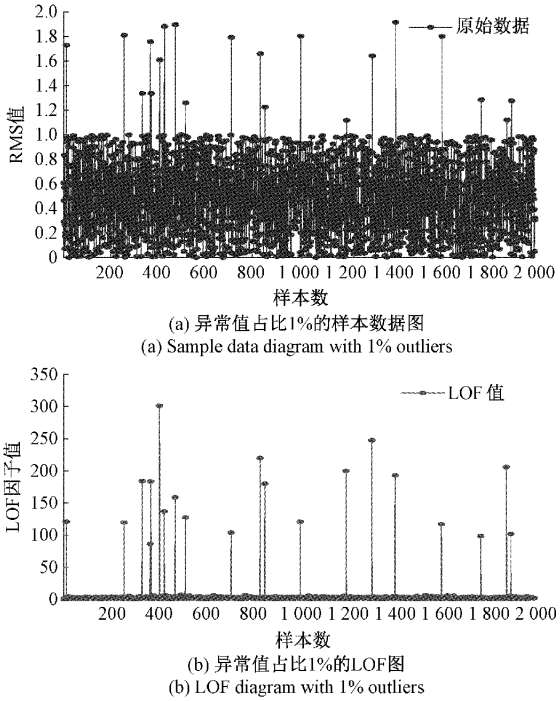
本节的数据集是采用以下模型生成的:

$$y=50(x^3-5)^2+2\varepsilon+5 \tag{36}$$

其中,y 为输出;$\boldsymbol{x}$ 为输入,且 $\boldsymbol{x}$ 的输入范围在[0,1]之间;$\boldsymbol{\varepsilon}$ 为服从参数 $\sigma=0.7$ 的瑞利分布。

本例中,从式(34)中随机提取2 000个样本作为数据集,并向2 000个样本数据集中随机添加异常样本数据构建含有异常值的新样本数据集,其添加的异常值数量占总数的1%、2%、3%、4%。随后,从新的样本数据集中随机提取1 500个样本作为训练数据集,剩下的500个样本作为测试数据集,从而应用于模型的训练与测试。

通过采用局部异常因子检测算法对数据集进行异常值检测,其相应的结果如图1~5所示。

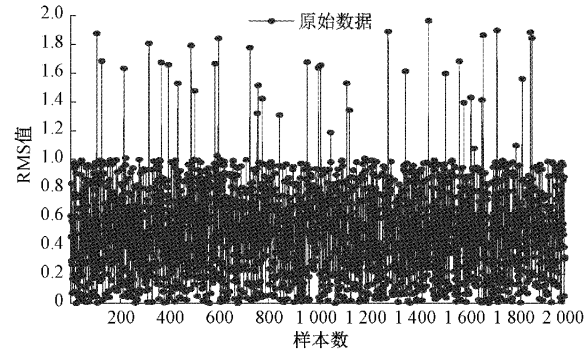


(a) 异常值占比1%的样本数据图
(a) Sample data diagram with 1% outliers

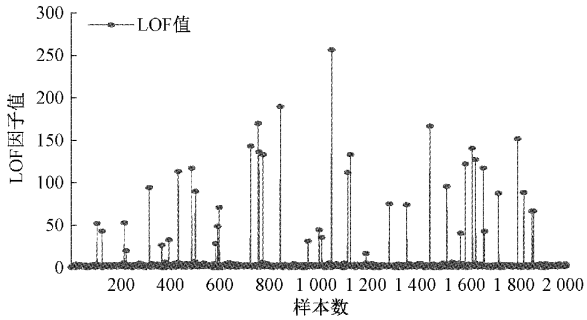
(b) 异常值占比1%的LOF图
(b) LOF diagram with 1% outliers

图1 异常值占比1%的检测图

Fig. 1 Detection diagrams with 1% outliers



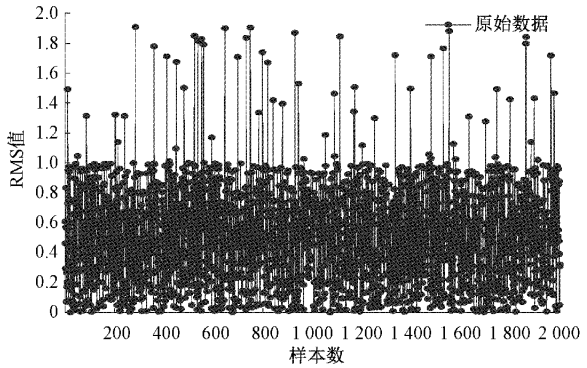
(a) 异常值占比2%的样本数据图
(a) Sample data diagram with 2% outliers



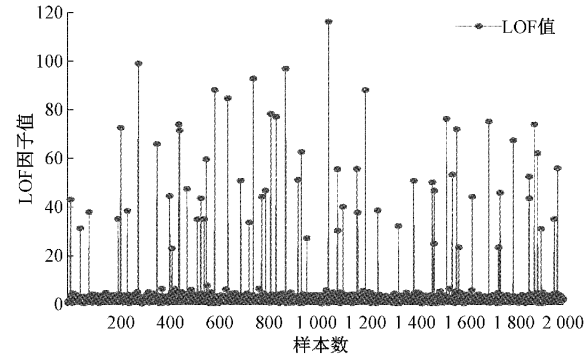
(b) 异常值占比2%的LOF图
(b) LOF diagram with 2% outliers

图 2　异常值占比 2% 的检测图

Fig. 2　Detection diagrams with 2% outliers



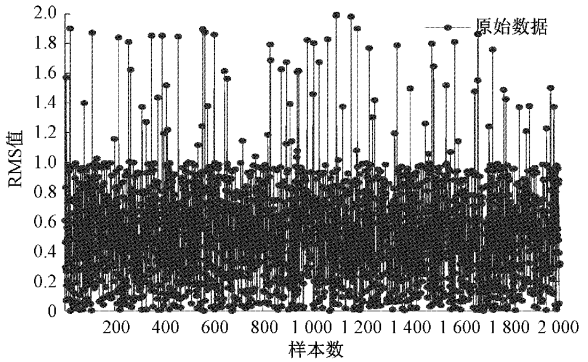
(a) 异常值占比3%的样本数据图
(a) Sample data diagram with 3% outliers



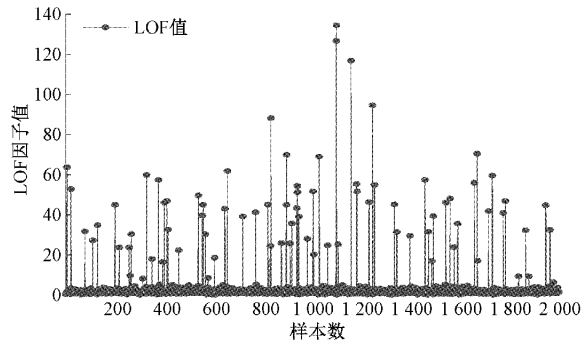
(b) 异常值占比3%的LOF图
(b) LOF diagram with 3% outliers

图 3　异常值占比 3% 的检测图

Fig. 3　Detection diagrams with 3% outliers



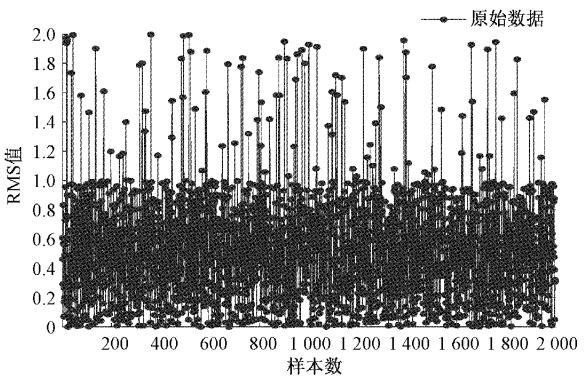
(a) 异常值占比4%的样本数据图
(a) Sample data diagram with 4% outliers



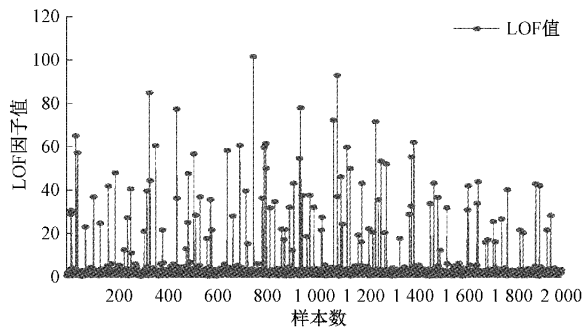
(b) 异常值占比4%的LOF图
(b) LOF diagram with 4% outliers

图 4　异常值占比 4% 的检测图

Fig. 4　Detection diagrams with 4% outliers



(a) 异常值占比5%的样本数据图
(a) Sample data diagram with 5% outliers



(b) 异常值占比5%的LOF图
(b) LOF diagram with 5% outliers

图 5　异常值占比 5% 的检测图

Fig. 5　Detection diagrams with 5% outliers

由图 1~5 可以看出,异常样本数据的 LOF 因子的值明显有异于正常样本数据的 LOF 因子的值,采用局部异常因子检测算法能够很好地将样本数据集中的异常样本和正常样本分开,并分别对其进行权重的设定。

本文针对含有异常样本数据的建模,并基于局部异常因子检测算法提出了一种改进的 LSSVM 算法。采用局部异常因子检测算法检测出异常值,然后根据检测出的异常值确定样本的权重,并用改进粒子群算法对模型的参数进行寻优。为了验证本文所提的改进算法能有效地解决所提出的问题,因此将本文算法与相关算法作对比实验,即将本文方法与最小二乘支持向量机算法(LSSVM)、自适应加权最小二乘支持向量机算法(AWLSSVM)、基于改进正态分布的加权最小二乘支持向量机算法(INDWLSSVM)、聚类最小二乘支持向量机算法[19](C-LSSVM)、不进行参数优化的本文算法(LOF-LSSVM)进行实验对比,发现本文方法能够提高 LSSVM 在面对异常样本时的建模精度,其效果对比如表 2 所示。

表 2 6 种方法的 RMSE 效果对比

Table 2 Comparison of RMSE effects of 6 methods

异常值占样本总数/%	LSSVM (文献[2],1999)	AWLSSVM (文献[14],2020)	INDWLSSVM (文献[13],2018)	C-LSSVM (文献[19],2020)	LOF-LSSVM	本文算法
1	4.636 9	2.863 5	2.595 2	1.766 5	1.759 8	1.624 3
2	4.869 3	2.945 9	2.863 5	1.603 0	1.599 0	1.486 3
3	5.689 1	4.289 2	4.168 5	1.585 7	1.585 7	1.467 1
4	6.325 4	4.456 6	4.368 9	1.856 2	1.729 2	1.695 4
5	8.967 3	7.787 1	6.594 2	2.967 1	2.578 4	2.232 1

由表 2 可知,本方法的模型效果良好。对经过局部异常因子检测后的样本数据集分别设定权重,能够避免出现权重过大或者过小的情况,对于模型的建模来说,既能够有效降低异常样本的影响,又能够完整保留正常样本信息,有利于最小二乘支持向量机对于含有异常样本的建模,完好地解决了问题。其中,在异常值占比不高时,如占比 1%、2%、3% 时,C-LSSVM 算法与本文算法相差不大;当异常值数据增多时,如 4%、5% 时,本文算法性能越来越好,其与其余 5 种算法性能差距拉大。因此,可以发现,随着异常值的增多,本文算法性能相较于其余算法性能将越来越好。

3.2 轴承数据实验

1)数据来源

本节采用来源于 FEMTO-ST 研究所提供的轴承数据进行仿真实验。该数据是在专门开发的轴承退化平台上采集得到,该平台可以在恒定或者可变工的情况下加速轴承的退化,并在线采集轴承的振动和温度信息。实验的振动传感器包括水平轴和垂直轴的两个微型加速度计,径向放置在轴承的外圈上。由于径向力在水平方向上加载,水平加速度计在失效分析中更有用,因此本例选择水平振动信号作为实验数据进行分析。

该数据集使用 17 个轴承在 3 种不同的操作条件下进行退化实验,并给出了 17 个完整的退化数据集,每个数据集包含轴承从正常完好到失效的全寿命周期数据。本文研究的是对含有异常值样本的最小二乘支持向量机进行建模,以增强模型的鲁棒性,而又因为轴承数据集中的失效部分的数据因为波动剧烈,很难体现出存在异常值对该部分数据的影响,并且不利于最小二乘支持向量机建模。所以,本文选取轴承数据集中部分正常工作状态的数据进行实验仿真,以减少其它干扰因素的影响,体现本文算法的优越性。故本文选取轴承 1-1 数据集中前 2 000 组正常工作状态的数据进行实验仿真,其相应的 RMS 图如图 6 所示。

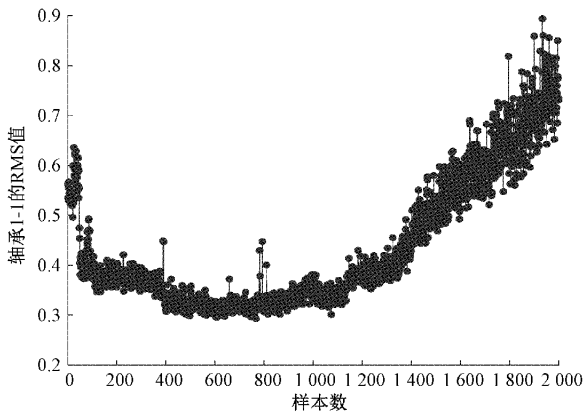


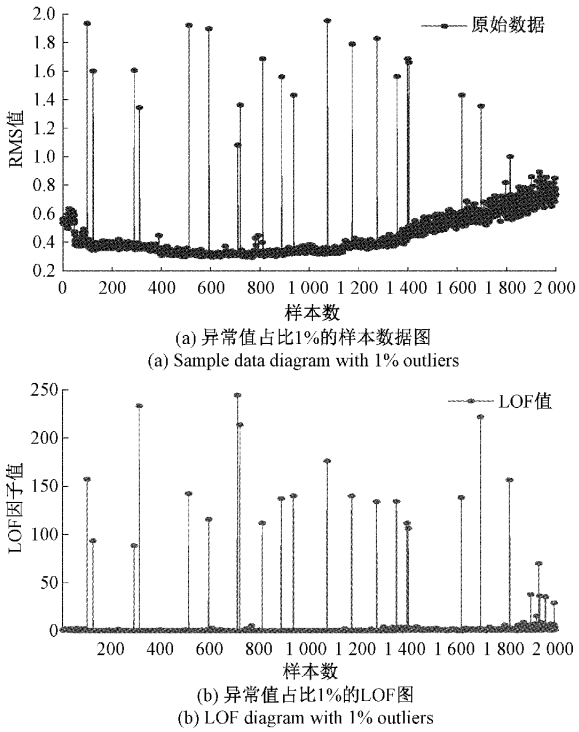
图 6 轴承 1-1 的 RMS 图

Fig. 6 RMS diagram of bearing 1-1

本例向提取后的 RMS 样本数据中,随机选取多个样本数据,并在此基础上将随机选取的样本数据值加 1 形成异常数据,然后将其和正常的样本数据一起组合成新的含有异常值样本的数据集来进行实验。其中,选取新的样本数据集的前 80% 作为训练数据集,后 20% 作为测试数据集进行本文的研究。

2)效果分析

本文随机向样本数据中添加异常样本点构建新的样本数据,其数量为占据样本数量总数的 1%、2%、3%、4%、5%,并运用 LOF 算法为每个样本点计算一个 LOF 因子。其所对应的样本数据和 LOF 值如图 7~11 所示。

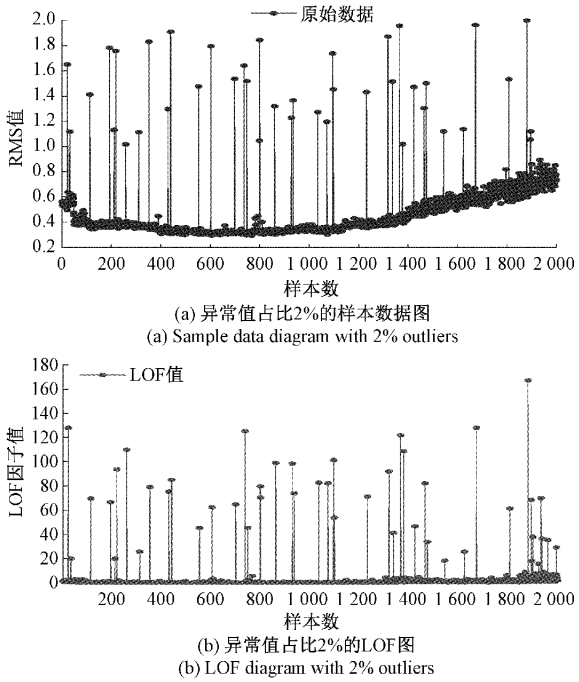


(a) 异常值占比1%的样本数据图
(a) Sample data diagram with 1% outliers

(b) 异常值占比1%的LOF图
(b) LOF diagram with 1% outliers

图 7　异常值占比 1% 的检测图

Fig. 7　Detection diagrams with 1% outliers

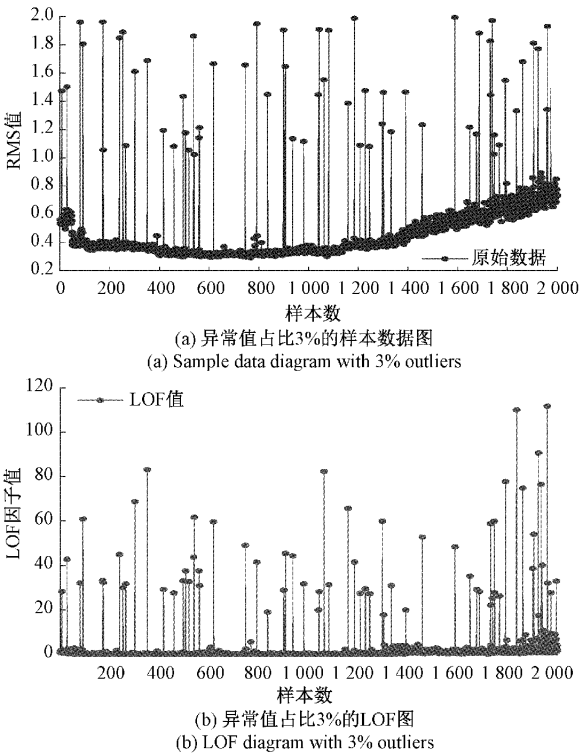


(a) 异常值占比2%的样本数据图
(a) Sample data diagram with 2% outliers

(b) 异常值占比2%的LOF图
(b) LOF diagram with 2% outliers

图 8　异常值占比 2% 的检测图

Fig. 8　Detection diagrams with 2% outliers

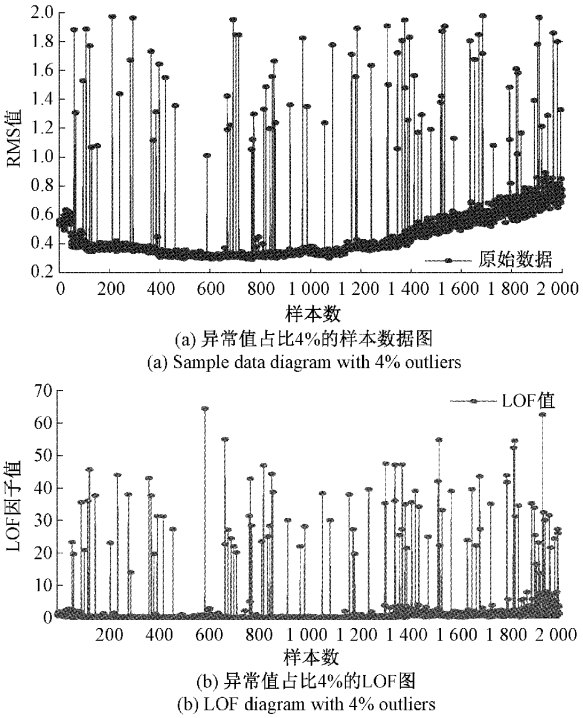


(a) 异常值占比3%的样本数据图
(a) Sample data diagram with 3% outliers

(b) 异常值占比3%的LOF图
(b) LOF diagram with 3% outliers

图 9　异常值占比 3% 的检测图

Fig. 9　Detection diagrams with 3% outliers

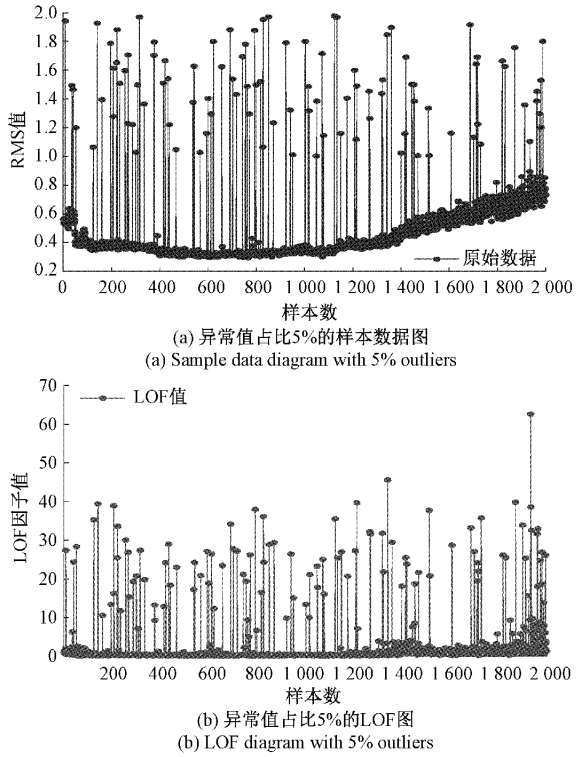


(a) 异常值占比4%的样本数据图
(a) Sample data diagram with 4% outliers

(b) 异常值占比4%的LOF图
(b) LOF diagram with 4% outliers

图 10　异常值占比 4% 的检测图

Fig. 10　Detection diagrams with 4% outliers



(a) 异常值占比5%的样本数据图
(a) Sample data diagram with 5% outliers

(b) 异常值占比5%的LOF图
(b) LOF diagram with 5% outliers

图 11 异常值占比 5% 的检测图

Fig. 11 Detection diagrams with 5% outliers

本文针对含有异常样本数据的建模,并基于局部异常因子检测算法提出了一种改进的 LSSVM 算法。采用局部异常因子检测算法检测出异常值,然后根据检测出的异常值确定样本的权重,并用改进粒子群算法对模型的参数进行寻优。为了验证本文所提的改进算法能有效地解决所提出的问题,因此将本文算法与相关算法作对比实验,即将本文方法与最小二乘支持向量机算法(LSSVM)、自适应加权最小二乘支持向量机算法(AWLSSVM)、基于改进正态分布的加权最小二乘支持向量机算法(INDWLSSVM)、聚类最小二乘支持向量机算法[19](C-LSSVM)、不进行参数优化的本文算法(LOF-LSSVM)进行实验对比,发现本文方法能够提高 LSSVM 在面对异常样本时的建模精度,其效果对比如表 3 所示。

在图 7~11 中,对应的 5 幅图依次为 1%、2%、3%、4%、5%异常样本数据的 LOF 图,观察 5 幅图可以发现,所对应的异常样本数据的 LOF 值大于周围正常数据的 LOF 值,其能够有效辨别样本数据中异常值,并将样本数据中的异常样本检测出来,然后针对异常值构建样本数据的权重,而不需要从数据的统计特性入手,可以避免对正常样本的权重的影响,提高 LSSVM 建模的精度。

表 3 6 种方法的 RMSE 效果对比

Table 3 Comparison of RMSE effects of 6 methods

异常值占样本总数/%	LSSVM (文献[2],1999)	AWLSSVM (文献[14],2020)	INDWLSSVM (文献[13],2018)	C-LSSVM (文献[19],2020)	LOF-LSSVM	本文算法
1	0.103 8	0.049 08	0.046 2	0.045 4	0.044 08	0.043 18
2	0.118 1	0.049 85	0.057 2	0.048 8	0.045 5	0.044 5
3	0.114 2	0.050 60	0.062 1	0.047 9	0.044 7	0.040 8
4	0.127 6	0.051 10	0.064 1	0.049 6	0.045 3	0.043 2
5	0.159 8	0.869 10	0.832 4	0.678 2	0.062 8	0.051 6

由表 3 可知,本文所提算法在 6 种算法模型中精度最好,其主要原因在于本文算法在设立样本权重的同时只降低异常样本的影响,而不影响正常数据的权重。而 AWLSSVM、INDWLSSVM 和 C-LSSVM 是根据模型误差的统计特性计算加权权重,计算得出的样本权重出现了偏大和偏小现象,导致模型无法达到最优化和使正常样本贡献度降低,丢失数据信息使建模不精确。因此,基于局部异常因子检测算法确定异常样本,然后针对异常样本确定其权重,其模型建模的精度要高。

4 结 论

本文提出的改进的 LSSVM 算法有如下特点:

1)针对最小二乘支持向量机模型对异常值敏感的问题,本文提出一种改进的最小二乘支持向量机算法。通过采用局部异常因子检测算法对样本数据进行检测,从而将样本数据划分为正常数据和异常数据,然后根据划分的数据对样本进行权重设置,从而建立模型。由此,将样本数据有效地分隔开来,并分别对其进行权重设定,避免了在减小异常样本数据影

响的同时减小正常样本数据的贡献,使模型的鲁棒性提高,降低了最小二乘支持向量机算法对异常值的敏感度。

2)针对模型参数与样本数据异常值的关系,本文引入信息熵和平均粒距的概念来改进粒子群算法。通过基于信息熵和平均粒距将粒子在迭代过程中划分为“活跃状态”和“惰性状态”两种状态,然后分别对不同状态的粒子进行更新。由此,针对粒子在迭代过程中不同状态进行粒子更新,有效地防止粒子群算法陷入局部极小值,使其能在全局范围内寻找到最优解,提高了模型在面对异常值时的鲁棒性。

本文通过数值仿真实验和轴承仿真实验验证了模型的性能,采用本文算法能有效地提高模型的鲁棒性和精度。

参考文献

[1] CORTES C, VAPNIK V. Support-vector networks[J]. Mach Learn,1995, 20: 273-297.

[2] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Process. Lett, 1999,9(3): 293-300.

[3] 陈法法,杨勇,马婧华,等. 信息熵与优化 LS-SVM 的轴承性能退化模糊粒化预测[J]. 仪器仪表学报,2016,37(4):779-787.
CHEN F F, YANG Y, MA J H, et al. Fuzzy granulation prediction for bearing performance degradation based on information entropy and optimized LSSVM[J]. Chinese Journal of Scientific Instrument,2016,37(4):779-787.

[4] 王佩丽,彭敏放,杨易旻,等. 应用模糊最优小波包和 LS-SVM 的模拟电路诊断[J]. 仪器仪表学报,2010,31(6):1282-1288.
WANG P L, PENG M F, YANG Y M, et al. Analog circuit diagnosis using fuzzy-rule based optimal waveletpacket and LSSVM[J]. Chinese Journal of Scientific Instrument,2010,31(6):1282-1288.

[5] YANG L, YANG S, LI S, et al. Coupled compressed sensing inspired sparse spatial-spectral LSSVM for hyperspectral image classification[J]. Knowl. Based Syst, 2015,79(5): 80-89.

[6] MEHRKANOON S, SUYKENS J A. Learning solutions to partial differential equations using LS-SVM[J]. Neurocomputing, 2015,159 (2): 105-116.

[7] GAO Y, SHAN X, HU Z, et al. Extended compressed tracking via random projection based on MSERs and online LS-SVM learning[J]. Pattern Recognit, 2016, 59: 245-254.

[8] 黄贤源,翟国君,隋立芬,等. LS-SVM 算法中优化训练样本对测深异常值剔除的影响[J]. 测绘学报,2011,40(1):22-27.
HUANG X Y, ZHAI G J, SUI L F, et al. The influence of optimized train samples on elimination of sounding outliers in the LS-SVM arithmetic[J]. Acta Geodaetica et Cartographica Sinica,2011,40(1):22-27.

[9] 郭战坤,金永威,梁小珍,等. 基于异常值检测的港口集装箱吞吐量预测模型[J]. 数学的实践与认识,2019,49(17):26-34.
GUO ZH K, JIN Y W, LIAN X ZH, et al. Prediction model of port container throughput based on outlier detection[J]. Mathematics in Practice and Theory,2019,49(17):26-34.

[10] SUYKENS J A K, DE BRABANTER J, LUKAS L, et al. Weighted least squares support vector machines: Robustness and sparse approximation[J]. Neurocomputing,2002, 48(1):85-105.

[11] VALYOU J, HORYTH G. A weighted generalized LS-SVM[J]. Period. Polytech. Ser. Electr. Eng, 2003, 47(3-4): 229-251.

[12] YOU L, JIZHEN L, YAXIN Q. A new robust least squares support vector machine for regression with outliers[J]. Adv. Control Eng. Inf. Sci, 2011, 15: 1355-136 .

[13] ZHANG C, LI C, PENG T, et al. Modeling and synchronous optimization of pump turbine governing system using sparse robust least squares support vector machine and hybrid backtracking search algorithm[J]. Energies, 2018, 11, 3108.

[14] CHEN Y J, GU CH SH, SHAO CH F, et al. An approach using adaptive weighted least squares support vector machines coupled with modified ant lion optimizer for dam deformation prediction[J]. Mathematical Problems in Engineering, 2020, DOI: 10.1155/2020/9434065.

[15] 吴博,赵法锁,贺子光,等. 基于 BA-LSSVM 模型的黄土滑坡致灾范围预测[J]. 中国地质灾害与防治学报,2020,31(5):1-6.
WU B, ZHAO F S, HE Z G, et al. Prediction of the disaster area of loess landslide based on least square support vector machine optimized by bat algorithm[J]. The Chinese Journal of Geological Hazard and Control,

2020,31(5):1-6.

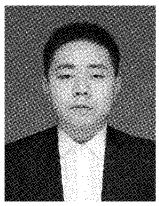
[16] 朱光轩,张庆松,刘人太,等. 基于 PSO-LSSVM 的砂层可注性预测模型及其敏感性分析[J]. 哈尔滨工业大学学报,2020,52(11):175-182.
ZHU G X, ZHANG Q S, LIU R T, et al. Groutability prediction in sand stratum using PSO-LSSVM and its sensitivity analysis [J]. Journal of Harbin Institute of Technology,2020,52(11):175-182.

[17] ZHU X, MA SH, XU Q, et al. WD-GA-LSSVM model for rainfall-triggered landslide displacement prediction[J]. Journal of Mountain Science, 2018, 15(1):156-166.

[18] KRINK T, VESTERSTROEM J S, RIGET J. Particle swarm optimisation with spatial particle extension [C]. Proceedings of the 2002 Congress on Evolutionary Computation, 2002: 1474-1479.

[19] LU X, MING L, HU T, et al. Collaborative learning-based clustered support vector machine for modeling of nonlinear processes subject to noise [J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2018:1-10.

作者简介



付乐天,2018 年于武汉工程大学获得学士学位,现为云南大学信息学院硕士研究生,主要研究方向为工业过程故障预测。
E-mail: 2022445415@qq.com

Fu Letian received his B. Sc. degree in 2018 from Wuhan Institute of Technology. Now, he is pursuing his M. Sc. degree in School of Information, Yunnan University. His main research interest is fault prediction of industrial process.

李鹏(通信作者),2007 年于华东理工大学获得博士学位,现为云南大学信息学院副教授,主要研究方向为工业过程故障诊断、可靠性分析与维护决策。
E-mail: lipeng@ynu.edu.cn

Li Peng (Corresponding author) received his Ph. D. degree in 2007 from East China University of Science and Technology. Now, he is an associate professor at School of Information, Yunnan University. His main research interest is fault diagnosis, reliability analysis and maintenance decision-making of industrial process.