文章编号:1000-582X(2012)12-120-13

剔除支持向量回归中异常数据算法

曾绍华[1,2],魏 延[2],唐远炎[1]

(1.重庆大学 计算机学院,重庆 400044;2.重庆师范大学 模式分析与信息处理研究所,重庆 401331)

摘 要:定义了回归问题中异常数据及其不满足回归映射关系差异程度的度量,分析了回归问题中理论映射模式与回归估计模式关系,提出并证明了回归问题中逐个剔除异常数据,建立回归估计模式逐步逼近理论模式的逐步逼近定理,并构建了以逐步逼近定理为理论依据的剔除支持向量回归中异常数据算法,理论分析了算法的收敛性和有效性。然后,引入逐步搜索算法改进剔除异常数据算法以解决大规模样本的支持向量回归中异常数据剔除问题,理论分析显示改进算法也是收敛的和有效的。最后,应用给定已知函数生成样本和 UCI 机器学习数据库样本数据仿真实验,结果显示算法是有效的和鲁棒的。

关键词:支持向量回归;异常数据;剔除异常数据算法;仿真

中图分类号:TP301.6;TP389.1 **文献标志码**:A

Algorithm of removing outliers in SVR

ZENG Shaohua[1,2], *WEI Yan*[2], *TANG Yuanyan*[1]

(1. College of Computer Science, Chongqing University, Chongqing 400044, China; 2. Institute of Pattern Analysis & Information Processing, Chongqing Normal University, Chongqing 401331, China)

Abstract: The outlier and the measurement that an outlier does not fit the theoretical model in the regression problems are defined. The relationship between the theoretical model and the regression model in the regression problem is analyzed. An approximate theorem is proposed and verified by deleting outlier one by one to construct SVR to approximate the theoretical model. An algorithm of detecting outliers in the SVR problems is constructed based on the approximate theorem. The theoretical analysis of the convergence and effectiveness of the proposed algorithm is given. Then, the step-by-step search algorithm is introduced to improve the outlier removing algorithm to remove outliers in SVR with large-scale samples. The theoretical analysis shows that the improved algorithm is convergent and effective. Finally, the samples produced by two test functions and the samples in UCI data set are used for simulation, and the results show that the proposed algorithm is effective and robust.

Key words: SVR(support vector regression); algorithm; algorithm of detecting outliers; simulation

在数据分析和机器学习中,回归(系统)问题所蕴含的 $x \mapsto y$ 的映射关系通常叫做模式。它里面蕴含的真实的 $x \mapsto y$ 映射是理论模式,用函数模拟或学习机学习所获得的是回归模式(或叫回归估计)。

由于异常数据影响数据分析和机器学习的精度,甚至导致错误的分析结果或者机器学习获得的是伪模式。从传统的数据分析到机器学习,剔除异常数据一直是广大学者研究的重要课题之一[1-2]。

收稿日期:2012-06-06

基金项目:重庆市教委科学技术研究项目(KJ110632);重庆市自然科学基金资助项目(CSTC2011JJA4008)

作者简介:曾绍华(1969-),重庆大学博士,主要从事智能计算与知识发现方向研究,(Tel)18908398198;(E-mail)zsh_cqu@126.com。

研究发现,回归问题的异常数据大致分为 2 类[3]:一类是超出定义域范围的异常数据,通常可以用定义域去检测、剔除;另一类是在定义域内,但不符 $x\mapsto y$ 的映射关系的异常数据,也是检测、剔除异常数据领域研究的重点。

许多学者在剔除异常数据的研究上做出了重大的贡献,发现了许多有效的算法。一些学者根据特定的统计分布提出了相应分布下的异常数据检测算法。李云飞[4]、李文东[5]提出了指数分布下的异常数据检测算法;乐立利[6]提出了双参数对数正态分布下的异常数据检测算法。Knorr E. M.[7-8]提出了基于距离异常数据检测算法。许多学者对它进行了深入研究,丰富和发展了基于距离异常数据检测算法。李星毅[9]提出了加权快速聚类的异常数据检测算法;Zhang Jun[10]和肖瑛[11]应用 3σ 准则检测异常数据;翁小清[12]和宋美娇[13]提出了子序列相似性和滑动时间窗的异常时序数据检测算法;周福娜[14]设计了投影能量显著性检验的异常数据检测算法。还有一些学者将蚁群算法[15]、小波分析[11]、神经网络[16]、遗传算法[17-18]、免疫算法[19]、流形理论[20]及其它智能算法技术[21-30]引入异常数据检测,提出了异常数据检测智能算法。

支持向量回归问题一般是高维样本。为了保证学习所获得模式的有效性,样本需覆盖整个特征空间。因此,正常样本间距离也相对较大,异常样本到正常样本间距离较大的特征不能充分展现,应用投影或聚类检测异常数据算法可能失效。应用文献[10-11]中 3σ 准则检测异常数据,其算法系数 3 的确定本身不科学,且未剔除异常样本之前无法精确估计理论模式 $y=F(x)$ 和 $|y-F(x)|$ 的方差 σ。如果用包含异常样本的训练样本集训练获得的模式 $y=f(x)$ 代替理论模式 $y=F(x)$ 去估计 σ,有可能 $y=f(x)$ 本身就是一个伪模式,导致 $|y-f(x)|$ 误差较大,估计的 σ 误差也较大,用 $|y-f(x)|<3\sigma$ 准则无法检测异常样本。其它的智能算法多是基于距离的演变算法,也具有相似的问题。尽管前人在异常数据检测算法研究上取得了丰硕的研究成果,但是应用在支持向量回归问题上效果不佳。支持向量回归问题中的异常数据具有特殊性,新颖的剔除支持向量回归中异常数据智能算法尚待进一步研究。

1　算法思想

1.1　回归问题中异常数据定义

根据 Hawkins[31] 给出的异常数据的定义,这里将回归问题中异常数据定义形式描述为:

定义 1　回归问题是 $x\mapsto y$ 的模式,不满足映射关系 $x\mapsto y$ 模式的样本点为异常数据。

根据定义 1,显然可以定义样本点不满足映射关系 $x\mapsto y$ 的差异程度。

定义 2　$y=F(x)$ 是映射关系 $x\mapsto y$ 的理论模式,定义样本点不满足映射关系 $x\mapsto y$ 的差异程度的一种度量 $d=|y-F(x)|$(其中,y 为观测值,$F(x)$ 为 y 的理论值)。

1.2　相关定理证明

定理 1　$y=F(x)$ 是映射关系 $x\mapsto y$ 的理论模式,$y=f(x)$ 是满足一定精度条件的映射关系 $x\mapsto y$ 的回归估计,样本点不满足映射关系 $x\mapsto y$ 的差异程度估计为 $d=|y-f(x)|$(其中,y 为观测值,$f(x)$ 为 y 的估计值)。

证明:在映射关系 $x\mapsto y$ 实例中抽取训练样本集 $S=\{(x_1,y_1),\cdots,(x_l,y_l),\}$,$x\in\mathbf{R}^n$,$y\in\mathbf{R}$。根据大数定理,当 $l\to\infty$,可以获得一个映射关系 $x\mapsto y$ 的回归估计 $y=f(x)$,得 $\lim\limits_{l\to\infty}f(x)=F(x)$。

根据定义 2,$d=|y-\lim\limits_{l\to\infty}f(x)|$。

在 $x\mapsto y$ 映射关系理论模式 $y=F(x)$ 未知的条件下,通常用有限样本集的满足一定精度条件的映射关系 $x\mapsto y$ 的回归估计 $y=f(x)$ 去估计 $\lim\limits_{l\to\infty}f(x)$,则样本点不满足映射关系 $x\mapsto y$ 的差异程度估计 $d=|y-f(x)|$。

命题得证。

推论 1　$y=F(x)$ 是映射关系 $x\mapsto y$ 的理论模式,$y=f(x)$ 是满足一定精度条件的映射关系 $x\mapsto y$ 的回归估计,样本点 i 满足理论模式 $y=F(x)$ 与 i 满足估计模式 $y=f(x)$ 等价。

证明:根据定义 2,样本点 i 不满足理论模式 $y=F(x)$ 的度量 $d_i=|y_i-F(x_i)|$,$y=f(x)$ 是满足一定精度条件的映射关系 $x\mapsto y$ 的回归估计,$F(x_i)$ 未知,用 $f(x_i)$ 估计 $F(x_i)$,即 $F(x_i)=f(x_i)$,则 $d_i=|y_i-f(x_i)|$。

同理,根据定义 2,样本点 i 不满足估计模式 $y=f(x)$ 的度量 $d'_i=|y_i-f(x_i)|$,

$$d_i=|y_i-F(x_i)|=|y_i-f(x_i)|=d'_i,$$

命题得证。

定理 2 $y=F(x)$是映射关系 $x\mapsto y$ 的理论模式，$y=f(x)$是满足一定精度条件的映射关系 $x\mapsto y$ 的回归估计。如果样本点 i 与 $y=f(x)$差异程度的度量为 $d_i=|y_i-f(x_i)|$较大，说明样本点 i 对估计模式 $f(x)$逼近理论模式 $F(x)$的破坏力较大，为异常数据的可能性(概率)也较大。

证明：$y=f(x)$是满足一定精度条件的映射关系 $x\mapsto y$ 的回归估计，$f(x_i)$是关于 x_i 对 y_i 估计。由定理1可知，样本点 i 不满足映射关系 $x\mapsto y$ 的差异程度的估计为 $d_i=|y_i-f(x_i)|$。

$y=F(x)$是映射关系 $x\mapsto y$ 的理论模式，则 $y_i=F(x_i)$，$d_i=|y_i-f(x_i)|=|F(x_i)-f(x_i)|$。

根据文献[31]可知，$E[|f(x)-F(x)|]$为 $f(x)$逼近 $F(x)$的一种度量。其它条件不变的情况下，由于$|y_i-f(x_i)|$较大，对 $E[|f(x)-F(x)|]$影响也较大，则样本点 i 使 $f(x)$逼近 $F(x)$降低，因而其对 $f(x)$逼近 $F(x)$起破坏作用也较大。

同时，$d_i=|y_i-f(x_i)|$越大，$d_i=|F(x_i)-f(x_i)|$也越大，根据定理2和定义1样本点 i 为异常数据的可能性(概率)也越大。

命题得证。

定理 3 $y=F(x)$是映射关系 $x\mapsto y$ 的理论模式，$y=f(x)$是满足一定精度条件的映射关系 $x\mapsto y$ 的回归估计，严重影响 $f(x)$逼近 $F(x)$的样本点是异常数据；或者说严重影响 $f(x)$拟合精度的样本点是异常数据。

证明：$y=F(x)$是映射关系 $x\mapsto y$ 的理论模式，$y=f(x)$是满足一定精度条件的映射关系 $x\mapsto y$ 的回归估计，则 $f(x)$逼近 $F(x)$的一种精度度量

$$\begin{aligned}\mathrm{MSE}=&\frac{1}{l}\sum_{j=1}^{l}[F(x_j)-f(x_j)]^2=\\&\frac{1}{l}\sum_{j=1,j\neq i}^{l}[F(x_j)-f(x_j)]^2+\\&\frac{1}{l}[F(x_i)-f(x_i)]^2。\end{aligned}\tag{1}$$

当训练样本点 i 严重影响 $f(x)$逼近 $F(x)$时，$[F(x_i)-f(x_i)]^2$ 显著大，$|F(x_i)-f(x_i)|$显著大；反之亦然。根据定义1，训练样本点 i 是异常数据。

同样，$f(x)$拟合映射关系 $x\mapsto y$ 的一种精度度量

$$\begin{aligned}\mathrm{MSE}'=&\frac{1}{l}\sum_{j=1}^{l}[y_j-f(x_j)]^2=\\&\frac{1}{l}\sum_{j=1,j\neq i}^{l}[y_j-f(x_j)]^2+\frac{1}{l}[y_i-f(x_i)]^2,\end{aligned}\tag{2}$$

当训练样本点 i 严重影响 $f(x)$拟合精度时，$[y_i-f(x_i)]^2$ 显著大，$|y_i-f(x_i)|$显著大；反之亦然。根据推论1和定义1，训练样本点 i 是异常数据。

命题得证。

定理 4(逐步逼近定理)

训练样本集 $S_0=\{(x_1,y_1),\cdots,(x_l,y_l),x\in\mathbf{R}^n,y\in\mathbf{R}\}$($l$ 足够大，满足随机抽样定理规定的精度要求；包含异常样本点个数 k 已知)，建立映射关系 $x\mapsto y$ 的回归估计。每次建立回归估计模型之后，从前一次训练样本集中剔除 $d_i=|y_i-f(x_i)|=\max\limits_{j=1\cdots l}|y_j-f(x_j)|$的训练样本点 i，这样直到剔除 k 个样本点，获得一个训练样本集序列$\{S_0,S_1,\cdots,S_k\}$及映射关系 $x\mapsto y$ 对应的回归估计序列$\{f_0(x),f_1(x),\cdots,f_k(x)\}$，则$\{f_0(x),f_1(x),\cdots,f_k(x)\}$逐步逼近映射关系 $x\mapsto y$ 的理论模式 $y=F(x)$。

证明：训练样本集 $S_0=\{(x_1,y_1),\cdots,(x_l,y_l),x\in\mathbf{R}^n,y\in\mathbf{R}\}$，$y=f_0(x)$是它关于映射关系 $x\mapsto y$ 的回归估计，则 $y=f_0(x)$拟合精度的一种度量[32]

$$\begin{aligned}\mathrm{MSE}_0=&\frac{1}{l}\sum_{j=1}^{l}[y_j-f_0(x_j)]^2=\\&\frac{1}{l}\Big\{\sum_{j=1,j\neq i}^{l}[y_j-f_0(x_j)]^2+[y_i-f_0(x_i)]^2\Big\},\end{aligned}\tag{3}$$

其中，训练样本点 i 的 $d_i=|y_i-f(x_i)|=\max\limits_{j=1\cdots l}|y_j-f(x_j)|$。

训练样本集 $S_1=\{S_0-$训练样本点 $i\}$，$y=f_1(x)$是它关于映射关系 $x\mapsto y$ 的回归估计，则

$$\mathrm{MSE}_1=\frac{1}{l-1}\sum_{j=1,j\neq i}^{l}[y_j-f_1(x_j)]^2。\tag{4}$$

假定映射关系 $x\mapsto y$ 的回归估计均最优化技术获得，且等效实现 $\min\|y-f(x)\|^2$。如果

$$\begin{aligned}&\sum_{j=1,j\neq i}^{l}[y_j-f_0(x_j)]^2\leqslant\\&\sum_{j=1,j\neq i}^{l}[y_j-f_1(x_j)]^2。\end{aligned}\tag{5}$$

因为建立训练样本集 S_1 关于映射关系 $x\mapsto y$ 的回归估计 $y=f_1(x)$时，要实现 $\min\sum_{j=1,j\neq i}^{l}[y_j-f_1(x_j)]^2$，训练样本集 S_1 关于映射关系 $x\mapsto y$ 的回归估计 $y=f_1(x)$与假设条件相矛盾，必然 $\sum_{j=1,j\neq i}^{l}[y_j-f_0(x_j)]^2\geqslant\sum_{j=1,j\neq i}^{l}[y_j-f_1(x_j)]^2$。

实现 $\min\|y-f(x)\|^2$ 最坏的情形是

$$\sum_{j=1,j\neq i}^{l}[y_j-f_0(x_j)]^2=\sum_{j=1,j\neq i}^{l}[y_j-f_1(x_j)]^2。\tag{6}$$

由于 $\mathrm{MSE}_1=\frac{1}{l-1}\sum_{j=1,j\neq i}^{l}[y_j-f_1(x_j)]^2$，按最坏情形，

$$\begin{aligned}\mathrm{MSE}_0 &= \frac{1}{l}\sum_{j=1}^{l}[y_j - f_0(x_j)]^2 = \\ &\frac{1}{l}\{\sum_{j=1,j\neq i}^{l}[y_j - f_0(x_j)]^2 + [y_i - \\ &f_0(x_i)]^2\} = \frac{1}{l}\left\{\frac{1}{l-1}\sum_{j=1,j\neq i}^{l} \right. \\ &\left.[y_j - f_1(x_j)]^2 \cdot (l-1) + \mathrm{MSE}_1\right\} + \\ &\frac{1}{l}\{[y_i - f_0(x_i)]^2 - \mathrm{MSE}_1\} = \\ &\mathrm{MSE}_1 + \frac{1}{l}\{[y_i - f_0(x_i)]^2 - \mathrm{MSE}_1\}。\end{aligned} \tag{7}$$

又由于 $d_i = |y_i - f(x_i)| = \max\limits_{j=1\cdots l}|y_j - f(x_j)|$,$[y_i - f_0(x_i)]^2 > \mathrm{MSE}_1$,因此,$\mathrm{MSE}_0 > \mathrm{MSE}_1$,$y = f_1(x)$的精度大于 $y = f_0(x)$的精度。

这样剔除 k 个样本点。同理,$\mathrm{MSE}_0 > \mathrm{MSE}_1 > \cdots > \mathrm{MSE}_k$,$\{f_0(x), f_1(x), \cdots, f_k(x)\}$的精度逐渐增大。

又有 $y = F(x)$是映射关系 $x \mapsto y$ 的理论模式,$y_j - f(x_j)$与 $F(x_j) - f(x_j)$等价,MSE 也代表 $y = f(x)$逼近 $y = F(x)$的精度,则$\{f_0(x), f_1(x), \cdots, f_k(x)\}$也逐步逼近映射关系 $x \mapsto y$ 的理论模式 $y = F(x)$。

命题得证。

推论 2 训练样本集 $S_0 = \{(x_1, y_1), \cdots, (x_l, y_l), x \in \mathbf{R}^n, y \in \mathbf{R}\}$($l$ 足够大,满足随机抽样定理规定的精度要求;包含异常样本点个数 k 未知),建立映射关系 $x \mapsto y$ 的回归估计。每次建立回归估计模型之后,从前一次训练样本集中剔除 $d_i = |y_i - f(x_i)| = \max\limits_{j=1\cdots l}|y_j - f(x_j)|$的训练样本点 i 构成新的训练样本集,再建立新的回归估计。直到剔除 k 个异常样本点后获得的训练样本集 S_k 建立的关于映射关系 $x \mapsto y$ 的回归估计 $f_k(x)$满足一定的精度要求的终止,即 $\Delta\mathrm{MSE}_m = \mathrm{MSE}_m - \mathrm{MSE}_{m+1}$ 小于阀值 $\Omega(>0)$。获得一个训练样本集序列$\{S_0, S_1, \cdots, S_k\}$及映射关系 $x \mapsto y$ 对应的回归估计序列$\{f_0(x), f_1(x), \cdots, f_k(x)\}$,则$\{f_0(x), f_1(x), \cdots, f_k(x)\}$逐步逼近映射关系 $x \mapsto y$ 的理论模式 $y = F(x)$。

证明:由定理 4 证明可知,每次建立回归估计模型之后,从前一次训练样本集中剔除 $d_i = |y_i - f(x_i)| = \max\limits_{j=1\cdots l}|y_j - f(x_j)|$的训练样本点 i 构成新的训练样本集,再建立新的回归估计获得的序列$\{f_0(x), f_1(x), \cdots, f_k(x), f_{k+1}(x)\}$和$\{S_0, S_1, \cdots, S_k, S_{k+1}\}$对应的 $\mathrm{MSE}_0 > \mathrm{MSE}_1 > \cdots > \mathrm{MSE}_k > \mathrm{MSE}_{k+1}$。$\{f_0(x), f_1(x), \cdots, f_k(x), f_{k+1}(x)\}$的精度逐步提高,并逐步逼近映射关系 $x \mapsto y$ 的理论模式 $y = F(x)$。

由于每次剔除训练样本点的 $d = \max\limits_{j=1,\cdots l}|y_j - f(x_j)|$,形成的递减序列 $d = \{d_0^*, d_1^*, \cdots, d_k^*, d_{k+1}^*\}$,所以 $\Delta\mathrm{MSE}_m$($\mathrm{MSE}_m - \mathrm{MSE}_{m+1}$)也为递减序列。当取合适的 $\Omega(>0)$,$\Delta\mathrm{MSE} > \Omega$,使剔除 S_0 的异常样本点的过程在终止 S_k 和 $f_k(x)$处终止。截取$\{f_0(x), f_1(x), \cdots, f_k(x), f_{k+1}(x)\}$获得的$\{f_0(x), f_1(x), \cdots, f_k(x)\}$同样逐步逼近映射关系 $x \mapsto y$ 的理论模式 $y = F(x)$。

命题得证。

推论 3 训练样本集 $S_0 = \{(x_1, y_1), \cdots, (x_l, y_l), x \in \mathbf{R}^n, y \in \mathbf{R}\}$($l$ 足够大,满足随机抽样定理规定的精度要求;包含异常样本点个数 k 未知),建立映射关系 $x \mapsto y$ 的回归模型。每次建立回归模型之后,从前一次训练样本集中剔除 $d_i = |y_i - f(x_i)| = \max\limits_{j=1\cdots l}|y_j - f(x_j)|$的训练样本点 i 构成新的训练样本集,建立新的回归模型。直到相邻 2 次回归的离差平方和的变化 $\Delta\mathrm{SES}_m$($= \mathrm{SES}_m - \mathrm{SES}_{m+1}$)小于合适阀值 $\theta(>0)$时终止。获得一个训练样本集序列$\{S_0, S_1, \cdots, S_k\}$及映射关系 $x \mapsto y$ 对应的回归估计序列$\{f_0(x), f_1(x), \cdots, f_k(x)\}$,则$\{f_0(x), f_1(x), \cdots, f_k(x)\}$逐步逼近映射关系 $x \mapsto y$ 的理论模式 $y = F(x)$,且剔除的都是异常样本点。

证明:训练样本集 $S_0 = \{(x_1, y_1), \cdots, (x_l, y_l), x \in \mathbf{R}^n, y \in \mathbf{R}\}$,$y = f_0(x)$是它关于映射关系 $x \mapsto y$ 的回归估计,则 $y = f_0(x)$拟合精度的一种度量

$$\begin{aligned}\mathrm{MSE}_0 &= \frac{1}{l}\sum_{j=1}^{l}[y_j - f_0(x_j)]^2 = \\ &\frac{1}{l}\{\sum_{j=1,j\neq i}^{l}[y_j - f_0(x_j)]^2 + [y_i - f_0(x_i)]^2\}。\end{aligned} \tag{8}$$

由于 l 为训练样本集 S_0 的样本个数,是常量,上述度量可等效地表示为

$$\begin{aligned}\mathrm{SES}_0 &= \sum_{j=1}^{l}[y_j - f_0(x_j)]^2 = \\ &\{\sum_{j=1,j\neq i}^{l}[y_j - f_0(x_j)]^2 + [y_i - f_0(x_i)]^2\},\end{aligned} \tag{9}$$

其中,训练样本点 i 的 $d_i = |y_i - f(x_i)| = \max\limits_{j=1\cdots l}|y_j - f(x_j)|$。

训练样本集 $S_1 = \{S_0 - \text{训练样本点 } i\}$,$y = f_1(x)$是它关于映射关系 $x \mapsto y$ 的回归估计,则

$$\mathrm{SES}_1 = \sum_{j=1,j\neq i}^{l}[y_j - f_1(x_j)]^2。 \tag{10}$$

假定映射关系 $x \mapsto y$ 的回归估计通过实现最优化技术获得,且等效实现 $\min \| y - f(x) \|^2$。如果

$\sum_{j=1,j\neq i}^{l}[y_j-f_0(x_j)]^2 \leqslant \sum_{j=1,j\neq i}^{l}[y_j-f_1(x_j)]^2$。 (11)

因为建立训练样本集 S_1 关于映射关系 $x\mapsto y$ 的回归估计 $y=f_1(x)$时，要实现 $\min\sum_{j=1,j\neq i}^{l}[y_j-f_1(x_j)]^2$，训练样本集 S_1 关于映射关系 $x\mapsto y$ 的回归估计 $y=f_1(x)$ 与假设条件相矛盾，必然 $\sum_{j=1,j\neq i}^{l}[y_j-f_0(x_j)]^2 \geqslant \sum_{j=1,j\neq i}^{l}[y_j-f_1(x_j)]^2$。所以

$$\begin{aligned}\Delta SES_0 &= SES_0 - SES_1 = \\ &[y_i-f_0(x_i)]^2+\sum_{j=1,j\neq i}^{l}[y_j-f_0(x_j)]^2 - \\ &\sum_{j=1,j\neq i}^{l}[y_j-f_1(x_j)]^2 \geqslant \\ &[y_i-f_0(x_i)]^2。\end{aligned} \quad (12)$$

同理，获得$\{\Delta SES_1,\cdots,\Delta SES_k\}$及相应的$\{f_2(x),\cdots,f_{k+1}(x)\}$。由于每次剔除训练样本点的 $d=\max\limits_{j=1\cdots l}|y_j-f(x_j)|$，形成的递减序列 $d=\{d_0^*,d_1^*,\cdots,d_k^*,d_{k+1}^*\}$，则$\{\Delta SES_0,\Delta SES_1,\cdots,\Delta SES_k\}$是严格的递减序列。所以能够取合适的 $\theta(>0)$，$\Delta SES>\theta$，使剔除 S_0 的异常样本点的过程在终止 S_k 和 $f_k(x)$处终止。获得$\{f_0(x),f_1(x),\cdots,f_k(x)\}$。根据定理4证明，$\{f_0(x),f_1(x),\cdots,f_k(x)\}$逐步逼近映射关系 $x\mapsto y$ 的理论模式 $y=F(x)$。

由于第 k 次剔除样本点 k^* 是异常样本点，则

$$d_k^* = |y_k^*-f_k(x_k^*)| = \max|y-f_k(x)|, \quad (13)$$

在第 k 次之前的第 m 次剔除的样本点 m^*，则

$$d_m^* = |y_m^*-f_m(x_m^*)| = \max|y-f_m(x)| > d_k^*, \quad (14)$$

所有，前面 $k-1$ 次剔除样本点也是是异常样本点。

命题得证。

1.3 算法思想

在推论3中，用SVR映射关系 $x\mapsto y$ 的回归估计，获得剔除支持向量回归中异常数据的算法。

从包含 k 个异常样本点(k 未知)的训练样本集 $S_0=\{(x_1,y_1),\cdots,(x_l,y_l),x\in\mathbf{R}^n,y\in\mathbf{R}\}$ 出发，建立映射关系 $x\mapsto y$ 的SVR。每次建立SVR之后，剔除样本点 $i(d_i=|y_i-f(x_i)|=\max\limits_{j=1\cdots l}|y_j-f(x_j)|)$构成新的训练样本集，建立新的SVR。直到 $\Delta SES=SES_{before}-SES_{after}<\theta$ 终止。被剔除的训练样本点为异常数据($S_{outliers}=\{S_0-S_k\}$)，$f_k(x)$为最后的SVR。异常数据阀值为 $D=\min|y_i-f_k(x_i)|$，$(x_i,y_i)\subset S_{outliers}$。

用新样本距 $f_k(x)$超平面的距离是否大于等于 D 检测新样本是否是异常数据。

2 剔除支持向量回归中异常数据算法及分析

2.1 剔除异常数据算法及分析

2.1.1 剔除支持向量回归中异常数据算法

依据推论3和1.3的算法思想，设计剔除支持向量回归中异常数据算法如下

算法1 剔除支持向量回归中异常数据算法

输入	1. 训练样本集 $S=\{(x_1,y_1),\cdots,(x_l,y_l),x\in\mathbf{R}^n,y\in\mathbf{R}\}$；
初始化	1. 选择核类型和设置核参数； 2. 设置终止条件 θ。

过程	1. 置 $S_{normal}=S,S_{work}=S$；
	2. 训练 S_{work} 的 SVR——$f(x_i)=\alpha K(x,x_i)+b$；
	3. 计算 S_{work} 关于 SVR 的离差平方和，即：
	$SES\_before=\sum_{j=1}^{l}(y_j-f(x_j))^2$；
	4. 用 S_{work} 和 SVR 计算 $D=\max\limits_{j=1\cdots l}\|y_j-f(x_j)\|$；
	5. $S_{work}=\{S-\{\|y_i-f(x_i)\|\geqslant D$ 的样本点$\}\}$；
	6. 训练 S_{work} 的 SVR——$f(x_i)=\alpha K(x,x_i)+b$；
	7. 计算 S_{work} 关于 SVR 的离差平方和，即：
	$SES\_after=\sum_{j=1}^{l}(y_j-f(x_j))^2$；
	8. delta_SES=SES_before−SES_after；
	9. while(delta_SES$\geqslant\theta$){
	9.1 $S_{normal}=S_{work}$；
	9.2 用 S_{work} 和 SVR 计算 $D=\max\limits_{j=1\cdots l}\|y_j-f(x_j)\|$；
	9.3 $S_{work}=\{S-\{\|y_i-f(x_i)\|\geqslant D$ 的样本点$\}\}$；
	9.4 SES_before=SES_after；
	9.5 训练 S_{work} 的 SVR——$f(x_i)=\alpha K(x,x_i)+b$；
	9.6 计算 S_{work} 关于 SVR 的离差平方和，即：
	$SES\_after=\sum_{j=1}^{l}(y_j-f(x_j))^2$；
	9.7 delta_SES=SES_before−SES_after;}
	10. $S_{outier}=S-S_{normal}$；
	11. 训练 S_{normal} 的 SVR——$f(x_i)=\alpha K(x,x_i)+b$；
	12. if($S_{outier}!=\varphi$)
	计算 S_{outier} 中样本距 SVR 超平面的最小值
	$min\_D\_outlier=\min\limits_{j=1\cdots k}\|y_j-f(x_j)\|$
	else
	min_D_outlier=inf
	13. 计算 S_{normal} 中样本距 SVR 超平面的最大值
	$max\_D\_normal=\max\limits_{j=1\cdots l}\|y_j-f(x_j)\|$；

输出	1. Del_Outlier.S=S； 2. Del_Outlier.$S_{normal}=S_{normal}$； 3. Del_Outlier.$S_{outier}=S_{outier}$； 4. Del_Outlier.SVR=SVR； 5. Del_Outlier.min_D_outlier=min_D_outlier； 6. Del_Outlier.max_D_normal=max_D_normal。

2.1.2　收敛性证明

剔除支持向量回归中异常数据算法的初始工作集为训练样本集 $S=\{(x_1,y_1),\cdots,(x_l,y_l),x\in\mathbf{R}^n,y\in\mathbf{R}\}$;训练工作集的 SVR,剔除工作集中 $d_i=|y_i-f(x_i)|=\max\limits_{j=1\cdots l}|y_j-f(x_j)|$ 的样本点 i 重构工作集;循环直到 $\Delta SES<\theta$ 终止。

假定在上述循环中,剔除 k 个样本点后,终止条件 $\Delta SES<\theta$ 被满足,形成工作集序列 $S_{work}=\{S_0,S_1,\cdots,S_k\}$;工作集序列对应的 SVR 为 $\{f_0(x),f_1(x),\cdots,f_k(x)\}$ 和 SVR 的误差平方和是 $\{SES_0,SES_1,\cdots,SES_k\}$。

由推论 3 的证明可知,$SES_0>SES_1>\cdots>SES_k$。它说明 $\{f_0(x),f_1(x),\cdots,f_k(x)\}$ 的精度逐步提高,并逐步收敛于 $f_k(x)$。也就是说工作集 $S_{work}=\{S_0,S_1,\cdots,S_k\}$ 序列对应的 SVR 将收敛于 S_k 的 SVR。

2.1.3　算法的有效性分析

在算法 1 中,最耗时步骤是训练 S_{work} 的 SVR——$(f(x_i)=\alpha K(x,x_i)+b)$,其时间复杂度为 $O(l^3)$。假定剔除 k 个样本点后,终止条件 $\Delta SES<\theta$ 被满足,则算法 1 的时间复杂度为 $O(k\cdot l^3)$。

训练 S_{work} 的 SVR 也使算法 1 占用最大空间复杂度。其中,计算核矩阵和核矩阵的逆是计算过程中占用最大存储空间,需要新增 $3\times l\times l$ 的存储空间。算法 1 空间复杂度的最大块是 $3\times l\times l$。

上述时空复杂度分析显示,在剔除中小刻度样本集异常数据时算法 1 是高效的。但,用算法 1 剔除大刻度样本集异常数据时,由于空间复杂度的最大块是 $3\times l\times l$,可能造成计算核矩阵和核矩阵的逆的内存申请量超过机器的物理存储量,使算法 1 执行困难;同时,由于算法 1 的时间复杂度为 $O(k\cdot l^3)$,随样本数 l 的增大,算法 1 耗时将以 l^3 级数增加,使算法 1 不能满足生产实践中剔除大刻度样本集异常数据的时间要求。

2.2　剔除大刻度样本集异常数据算法及分析

2.2.1　剔除大刻度样本集异常数据的算法思想

前面的定理 4 和推论 3 给剔除支持向量回归异常数据算法提供了理论依据。但逐个搜索异常数据循环中要构建剔除异常数据后剩余样本的支持向量机。由于支持向量机的时间复杂度与样本数有关,随样本数的增加,训练时间呈几何级数增加[33],导致算法 1 在剔除大刻度样本集异常数据中失效。

曾绍华和唐远炎在文献[34]中提出了基于分块算法的训练大刻度样本集 ε-SVR 的逐步搜索算法思想。引入该算法思想和 SVR 稀疏算法可大大降低搜索大刻度样本集稀疏 SVR 的时间、空间复杂度。

稀疏 SVR 的支持向量是一个小样本集,可用算法 1 剔除其异常数据。然后,通过稀疏 SVR、样本点距超平面距离逐步搜索异常数据,解决剔除大刻度样本集异常数据的异常问题。

2.2.2　大刻度样本集的稀疏 SVR 算法

基于曾绍华和唐远炎在文献[34]中提出的算法思想的大刻度样本集稀疏 SVR 算法如下

算法 2　大刻度样本集的稀疏 SVR 算法

输入	1. 训练样本集 $S=\{(x_1,y_1),\cdots,(x_l,y_l),x\in\mathbf{R}^n,y\in\mathbf{R}\}$;
初始化	1. 设置支持向量个数 num_SVs; 2. 设置搜索新样本个数 num_search_samples; 3. 选择核类型 ker_type; 4. 设置核参数 ker_parameters; 5. 设置终止条件 Ter。

过程	1. $T=S$;对 T 随机排序;在 T 中抽取 num_SVs 个样本构成 S_{work},并在 T 中删除这 num_SVs 个样本;$S_{sv}=S_{work}$;
	2. 清除 S_{work} 中相同样本点;训练 S_{work} 的 SVR $f(x_i)=\alpha K(x,x_i)+b$;
	3. 计算 S 关于 SVR 的平均离差平方和,即:$\text{MSE\_before}=\frac{1}{l}\sum_{j=1}^{l}(y_j-f(x_j))^2$;
	4. do{
	4.1 MSE=MSE_before;
	4.2 while($T!=\phi$){
	4.2.1 $S_{work}=S_{sv}$;训练 S_{work} 的 SVR $f(x_i)=\alpha K(x,x_i)+b$;
	4.2.2 稀疏掉 S_{work} 中 $\lvert\alpha\rvert$ 较小的 num_search_samples 个样本点,在 T 中抽取 num_search_samples 个新样本点添加到 S_{work} 中;并在 T 中删除这 num_search_samples 个样本点;
	4.2.3 清除 S_{work} 中相同样本点;重新训练 S_{work} 的 SVR——$f(x_i)=\alpha K(x,x_i)+b$;
	4.2.4 计算 S 关于 SVR 的平均离差平方和,即:$\text{MSE\_after}=\frac{1}{l}\sum_{j=1}^{l}(y_j-f(x_j))^2$;
	4.2.5 if(MSE_before>MSE_after){
	4.2.5.1 MSE_before=MSE_after;
	4.2.5.2 $S_{sv}=S_{work}$;
	}
	4.3 $T=S$;并随机排列 T;
	}While(MSE−MSE_before<Ter)
	5. 训练 S_{sv} 的 SVR——$f(x_i)=\alpha K(x,x_i)+b$;

输出	1. SVR.MSE=MSE_before; 2. SVR.$SV=S_{sv}$; 3. SVR.$\alpha=\alpha$; 4. SVR.$b=b$。

2.2.3　剔除大刻度样本集异常数据算法

根据 2.2.1 的算法思想,建立剔除大刻度样本集异常数据的搜索算法如下:

算法3 剔除大刻度样本集异常数据算法

输入	1. 训练样本集 $S=\{(x_1,y_1),\cdots,(x_l,y_l),x\in\mathbf{R}^n,y\in\mathbf{R}\}$；
初始化	1. 支持向量个数 p； 2. 每次搜索的新样本的个数 q； 3. 选择核类型 ker_type； 4. 设置核参数 ker_parameters； 5. 设置终止条件 Ter,θ。

过程	1. $S_{normal}=S$；并对 S_{normal} 随机排序；Mark=true；DD=inf；
	2. while(Mark){
	2.1Mark=false；
	2.2 用算法2搜索关于 S_{normal} 的稀疏SVR；
	2.3S_{work}=SVR.SV；用算法1剔除工作集 S_{work} 的异常数据，获得 Del_Outlier；
	2.4 if($DD>$Del_Outlier.min_D_outlier)
	DD=Del_Outlier.min_D_outlier；
	SVR 和 DD 剔除 S_{normal} 中异常样本，重构 S_{normal}；
	2.5.2S_{work}=Del_Outlier.S_{normal}；Mark=true；
	}
	2.6label=true；
	2.7While(label){
	2.7.1label=false；
	2.7.2 在 S_{normal} 中抽取距SVR超平面距离最大的样本点加入 S_{work}；并将该样本点从 S_{normal} 中删除；
	2.7.3 用算法1剔除工作集 S_{work} 的异常数据，获得 Del_Outlier；
	2.7.4 if(Del_Outlier.S_{outier} 不是空集){
	2.7.4.1label=true；Mark=true；
	2.7.4.2if($DD>$Del_Outlier.min_D_outlier)
	DD=Del_Outlier.min_D_outlier；
	}
	}
	2.8$S_{normal}=S$；用SVR和 DD 剔除 S_{normal} 中异常样本，重构 S_{normal}，并随机排列；
	}
	3. $S_{normal}=S$；
	4. 用SVR和 DD 剔除 S_{normal} 中异常样本，即
	$S_{normal}=\{S$ 中$|y_i-f(x_i)|<DD$ 的样本点$\}$；
	5. 用SVR计算关于 S_{normal} 的平均离差平方和MSE；

输出	1. Del_Lar_Outlier.$S=S$； 2. Del_Lar_Outlier.SV=SVR.SV； 3. Del_Lar_Outlier.SVR=SVR； 4. Del_Lar_Outlier.MSE=MSE； 5. Del_Lar_Outlier.D=DD。

2.2.4 剔除大刻度样本集异常数据算法的复杂度分析

在剔除大刻度样本集异常数据算法中，最耗时的步骤是用算法2搜索关于 S_{normal} 最多包含 p 个支持向量的稀疏SVR。为了简化分析，取 S_{normal} 的样本个数为样本集 S 的样本个数为 l，工作集 S_{work} 的样本个数为 p，每次循环搜索新样本的个数为 q，样本数据维数为 n。算法2完成一轮循环搜索的时间复杂度为 $O((l/q)\cdot p^3)$。如果执行 η 轮循环达到终止条件，算法2的时间复杂度为 $O(\eta\cdot(l/q)\cdot p^3)$。空间复杂度为 $3\times p\times p+3\times p\times n+2\times l\times n$。

其次是剔除了工作集 S_{work} 的 τ 个异常数据，需训练 S_{work} 的SVR执行 $\tau+2$ 次，其时间复杂度为 $O((\tau+2)\cdot p^3)$。可借用搜索稀疏SVR申请的存储空间，没有新的存储空间花销。

再就是算法3中的步骤2.7修改异常样本点距SVR超平面地最小距离。根据 S_{normal} 中样本点距SVR超平面距离，从大到小逐点与 S_{work} 的正常样本一起重构 S_{work}，再剔除工作集 S_{work} 异常数据，直到加入点不被剔除时终止。在上述循环终止前，每次剔除工作集 S_{work} 异常数据中，有且仅有一个样本点被剔除。假定有 m 样本点被剔除后步骤2.7执行结束，从算法1可以得知，需训练 S_{work} 的SVR执行 $3m+2$ 次，其时间复杂度为 $O((3m+2)\cdot p^3)$。也没有新的存储空间花销

由于 τ、$m\ll p$，完成一轮步骤2的循环，时间复杂度仍为算法2的时间复杂度。最好的情形是步骤2执行2轮循环算法3终止。通常情况下，执行几轮循环算法3也就终止了。算法3的时间复杂度就是几倍算法2的时间复杂度，且保持了算法2的空间复杂度。

与直接用算法1剔除大刻度样本集异常数据比较，算法3的时间复杂度为 $O(p^3)$ 远小于算法1的时间复杂度为 $O(l^3)$；在空间复杂度上，剔除异常数据过程在存储样本集的基础上，算法1需新增存储空间的最大块是 $3\times l\times l$ 远大于算法3需新增存储空间的最大块是 $3\times p\times p$。显然，在剔除大刻度样本集异常数据上，算法3在算法1的基础上极大地降低了时、空复杂度。

2.2.5　剔除大刻度样本集异常数据算法的收敛性分析

算法 3 的关键步骤是算法 1 和算法 2。2.1.2 证明了算法 1 是收敛的。曾绍华和唐远炎在文献[34]中证明基于逐步搜索大刻度样本 SVR 的收敛性,因而,基于该算法思想的算法 2 也是收敛的。在剔除异常数据搜索过程中,算法 3 保留 SVR 的 SES 逐步减小,形成一个 SES 的递减序列,且 SES>0 的。由于 SES 有界且极限存在,因此,算法 3 也是收敛的。

3　仿真实验

3.1　仿真实验样本的获取

3.1.1　标准函数 1 生成模拟样本方法

1)正常样本:在[−3,3],等间距生成仿真实验样本的 x,$y=\sin\pi x/\pi x+\text{rand}()$,其中 rand()∈[−0.05,+0.05]。

2)普通异常数据:在[−3,3],随机生成仿真实验样本的 x,$y=\sin\pi x/\pi x+(-1)^j\text{rand}()$,其中:$j=\text{int}(10\text{rand}()+1)$,rand()∈[+0.10,+0.15]。

3)鲁棒性实验异常数据:在[−3,3],随机生成仿真实验样本的 x,$y=\sin\pi x/\pi x+\text{rand}()$ 或 $y=\sin\pi x/\pi x-\text{rand}()$二者取一,rand()∈[+0.10,+0.15]。

3.1.2　标准函数 2 生成模拟样本方法

1)正常样本:在[−3,3],等间距生成仿真实验样本的 x,$y=x_1\cdot e^{-x_1^2-x_2^2}+\text{rand}()$,其中 rand()∈[−0.05,+0.05]。

2)普通异常数据:在[−3,3],随机生成仿真实验样本的 x,$y=x_1\cdot e^{-x_1^2-x_2^2}+(-1)^j\text{rand}()$,其中:$j=\text{int}(10\text{rand}()+1)$,rand()∈[+0.1,+0.15]。

3)鲁棒性实验异常数据:在[−3,3],随机生成仿真实验样本的 x,$y=x_1\cdot e^{-x_1^2-x_2^2}+\text{rand}()$或 $y=x_1\cdot e^{-x_1^2-x_2^2}-\text{rand}()$二者取一,rand()∈[+0.1,+0.15]。

3.1.3　普通实验样本集

Data1:标准函数 1 生成正常样本 100 个,普通异常数据 10 个,对样本随机排列。

Data2:标准函数 2 生成正常样本 400 个,普通异常数据 20 个,对样本随机排列。

Data3:标准函数 1 生成正常样本 1000 个,普通异常数据 50 个,对样本随机排列。

Data4:标准函数 2 生成正常样本 10 000 个,普通异常数据 1 000 个,对样本随机排列。

Data5:UCI 数据集[35] wine-quality Data Set (12 维),样本 4 898 个。

Data6:UCI 数据集[35] Concrete Compressive Strength Data Set (9 维),样本 1 030 个。

3.1.4　鲁棒性实验样本集

Data7:标准函数 1 生成正常样本 100 个,鲁棒性实验异常数据 5 个,对样本随机排列。

Data8:标准函数 2 生成正常样本 400 个,鲁棒性实验异常数据 20 个,对样本随机排列。

Data9:标准函数 1 生成正常样本 1 000 个,鲁棒性实验异常数据 100 个,对样本随机排列。

Data10:标准函数 2 生成正常样本 10 000 个,鲁棒性实验异常数据 500 个,对样本随机排列。

Data11:UCI 数据集[35] wine-quality Data Set (12 维),样本 4 898 个。用算法 3 获得异常样本,删除 $y_i-f(x_i)<0$ 的异常样本后的样本集。

Data12:UCI 数据集[35] Concrete Compressive Strength Data Set (9 维),样本 1 030 个。用算法 3 剔除的异常样本,删除 $y_i-f(x_i)>0$ 的异常样本后的样本集。

3.2　仿真实验与分析

仿真实验环境:Windows 7 (32bit),AMD Turion 64×2 TL50,4G 内存,Matlab R2010a。

3.2.1　确定终止条件 θ

从回归问题异常数据的定义可知,异常数据距 SVR 超平面的距离较大。相应地,剔除一个异常数据后,训练样本集关于 SVR 的离差平方和 *SES* 显著减小。而剔除一个正常样本后,训练样本集关于 SVR 的平均离差平方和变化较小。对一个特定的样本集,这种变化的程度无法预知。它体现为算法 1 的终止条件 θ。如果没有一个合适的终止条件 θ,算法 1 可能不能完全剔除异常数据或将部分正常样本视为异常数据剔除。如何获取一个合适的终止条件 θ 成为了一个问题。

以 data1 为例,做算法 1($\theta=0.000\,5$)关于 Data1 实验,得逐个剔除距 SVR 超平面的距离最大点的 SES 曲线(见图 1)和逐个剔除距 SVR 超平面的距离最大点前后 SES 的变化(delta-SES)曲线(见图 2)

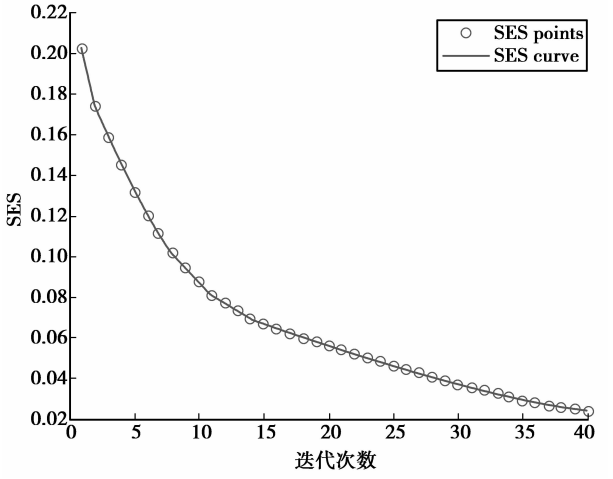


图 1　算法 1 计算 data1($\theta=0.000\,5$)的 SES 曲线

图 1 显示，SES 曲线前端 SES 减小较快，然后 SES 减小的幅度显著减小，有明显的剔除的是异常数据点过渡到剔除了正常数据点的特征。图 2 将 SES 的变化直观化。delta-SES 值明显分为两部分，前面部分是剔除异常数据的 SES 变化，delta-SES 值较大；后面部分是剔除了正常样本的 SES 变化，delta-SES 值相对较小。终止条件 θ，大于 delta-SES 值较小部分的最大 delta-SES 值，且小于 delta-SES 值较大部分的最小 delta-SES 值。如图 2 所示，直观地确定终止阀值 $0.004<\theta<0.006$，算法 1 都能正确剔除 Data1 的异常数据。

3.2.2 普通实验样本集仿真实验结果

用 3.2.1 确定终止条件 θ 的方法后，将 θ 带入相应算法，获得实验结果。

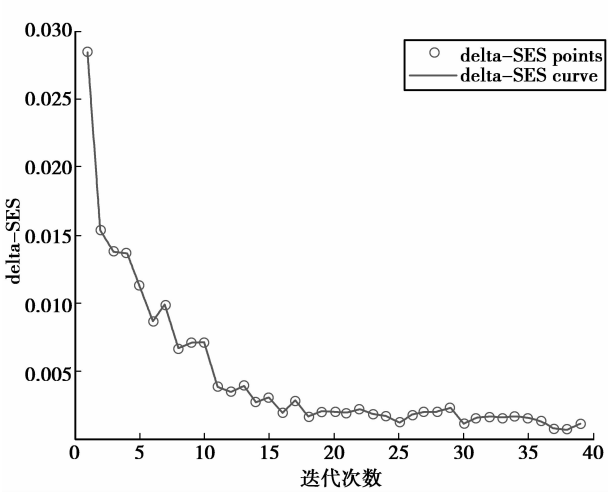
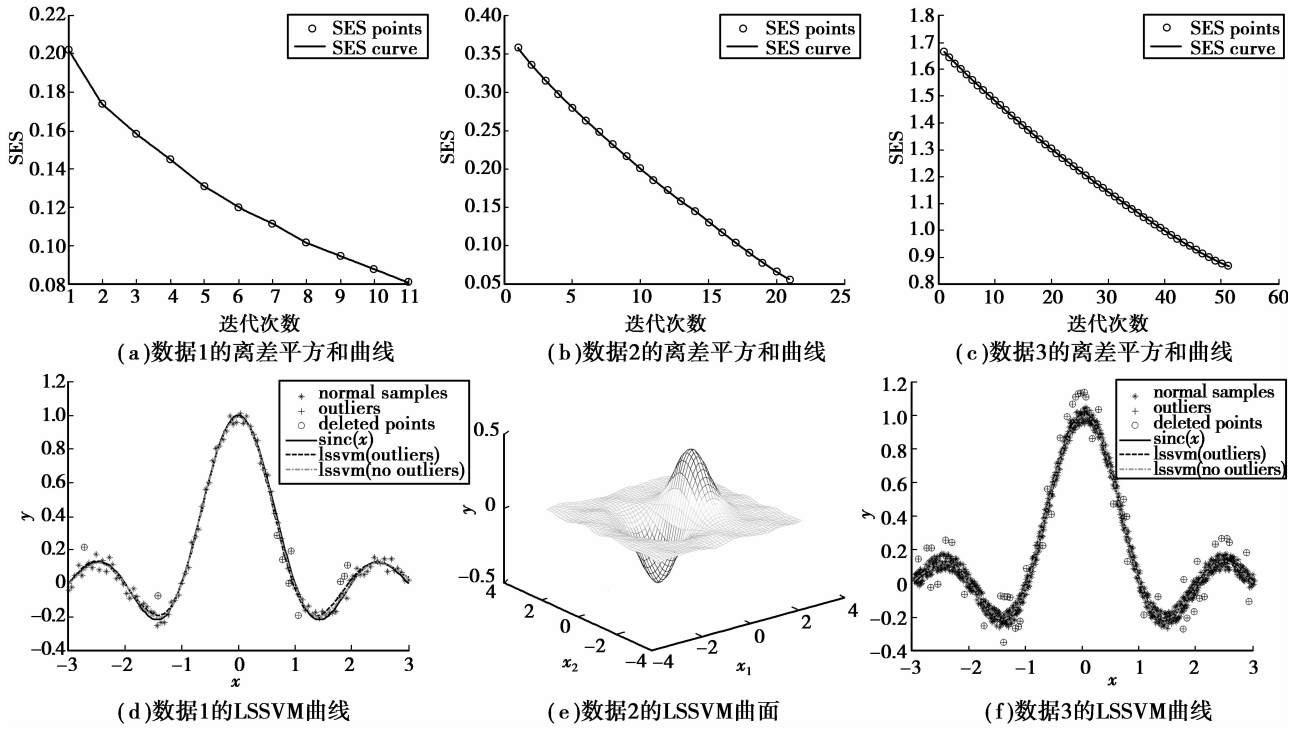


图 2 算法 1 计算 data1($\theta=0.000\ 5$)的 delta-SES 曲线

1)部分仿真实验的图形结果



(a)数据1的离差平方和曲线 (b)数据2的离差平方和曲线 (c)数据3的离差平方和曲线

(d)数据1的LSSVM曲线 (e)数据2的LSSVM曲面 (f)数据3的LSSVM曲线

图 3 算法 1 计算部分普通实验样本集($\theta=0.005$)的 SES 变化曲线和 LSSVM 曲(线)面

2)时间花销及有效性仿真实验结果

表 1 相应算法关于普通实验样本集的仿真结果

算法	数据集	Ter	θ	p	q	计算时间/s	MSE	总样本个数	异常数据个数	实际剔除样本个数	剔除异常数据个数	命中率/%
算法 1	Data1	—	5.0E−3	—	—	0.782±0.017	7.18E−4	110	10	10	10	100.0
算法 1	Data2	—	5.0E−3	—	—	39.16±0.137	1.39E−4	420	20	20	20	100.0
算法 1	Data3	—	5.0E−3	—	—	237.9±0.403	8.69E−4	1050	50	50	50	100.0
算法 1	Data6	—	100	—	—	530.8±1.279	1.056	1030	—	73	—	—

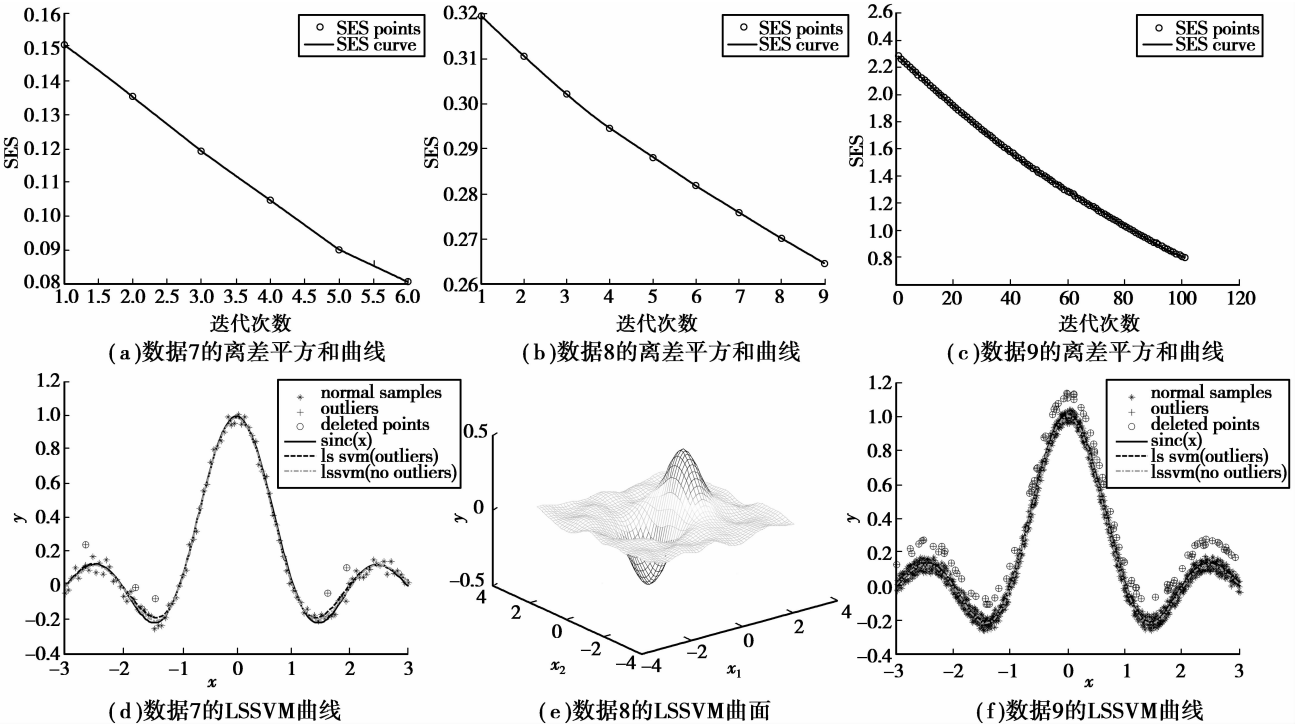
续表 1

算法	数据集	Ter	θ	p	q	计算时间/s	MSE	总样本个数	异常数据个数	实际剔除样本个数	剔除异常数据个数	命中率/%
算法 3	Data3	1.0E−9	5.0E−3	200	20	192.0±54.08	8.78E−4	1050	50	50	50	100.0
算法 3	Data4	1.0E−9	5.0E−3	200	20	5677.3±856	2.39E−04	11000	1000	1000	1000	100.0
算法 3	Data5	1.0E−9	0.25	200	20	907±229.4	0.215	4898	—	1142	—	—
算法 3	Data6	1.0E−9	100	300	20	363.5±74.8	26.297	1030	—	73	—	—
3σ 准则 *	Data1	—	—	—	—	0.236±0.071	1.83E−3	110	10	1	1	10.0
3σ 准则 *	Data2	—	—	—	—	2.125±0.183	8.53E−4	420	20	20	20	100.0
3σ 准则 *	Data3	—	—	—	—	4.937±1. 337	1.58E−3	1050	50	33	33	66.0
3σ 准则 *	Data6	—	—	—	—	7.875±1.428	28.093	1030	—	11	—	—
3σ 准则	Data3	1.0E−9	—	200	20	182.6±39.59	8.65E−4	1050	50	8	8	16.0
3σ 准则	Data4	1.0E−9	—	200	20	4979.3±696	7.63E−04	11000	1000	147	147	14.7
3σ 准则	Data5	1.0E−9	—	200	20	817±189.2	0.286	4898	—	295	—	—
3σ 准则	Data6	1.0E−9	—	300	20	310.5±69.4	29.375	1030	—	76	—	—

*“3σ 准则 *”中用 SVR 标准算法,“3σ 准则”中用 SVR 稀疏算法(算法 2);用 LS-SVM 计算数据集 Data1、Data2、Data3 和 Data4,LS-SVM 标准算法来源于 LS-SVMlab1.5;用 ε-SVR 计算数据集 Data5 和 Data6,ε-SVR 标准算法来源于 Matlab 工具箱的 quadprog();自编 LS-SVM 和 ε-SVR的稀疏算法。

3.2.3　鲁棒性实验样本集仿真实验结果

1)鲁棒性实验的图形结果



(a)数据7的离差平方和曲线　(b)数据8的离差平方和曲线　(c)数据9的离差平方和曲线

(d)数据7的LSSVM曲线　(e)数据8的LSSVM曲面　(f)数据9的LSSVM曲线

图 4　算法 1 计算部分鲁棒实验样本集(θ=0.005)的 SES 变化曲线和 LSSVM 曲(线)面

2)时间花销及有效性仿真实验结果

表 2 相应算法关于鲁棒性实验样本集的仿真结果

算法	数据集	Ter	θ	p	q	计算时间/s	MSE	总样本个数	异常数据个数	实际剔除样本个数	剔除异常数据个数	命中率%
算法 1	Data7	—	5.0E−3	—	—	0.453±0.012	8.06E−4	105	5	5	5	100.0
算法 1	Data8	—	5.0E−3	—	—	40.36±0.375	5.32E−4	420	20	20	20	100.0
算法 1	Data9	—	5.0E−3	—	—	482.4±7.455	7.98E−4	1100	100	100	100	100.0
算法 1	Data12	—	100	—	—	501.1±1.079	1.056	992	35	35	35	100.0
算法 3	Data9	1.0E−9	5.0E−3	200	20	167.7±27.48	8.58E−04	1100	100	100	100	100.0
算法 3	Data10	1.0E−9	5.0E−3	200	20	5563±901	8.53E−04	10500	500	500	500	100.0
算法 3	Data11	1.0E−9	0.25	200	20	853±193.8	0.209	4351	595	595	595	100.0
算法 3	Data12	1.0E−9	100	300	20	330.9±79.6	26.765	992	35	34	34	97.3
3σ 准则 *	Data7	—	—	—	—	0.203±0.063	1.83E−3	105	5	2	2	40.0
3σ 准则 *	Data8	—	—	—	—	2.046±0.137	7.60E−4	420	20	2	2	10.0
3σ 准则 *	Data9	—	—	—	—	5.406±1.431	2.07E−3	1100	100	6	6	6.0
3σ 准则 *	Data12	—	—	—	—	6.875±1.337	2.983	992	35	7	7	20.0
3σ 准则	Data9	1.0E−9	—	200	20	159.6±28.18	8.49E−04	1100	100	16	16	16.0
3σ 准则	Data10	1.0E−9	—	200	20	5467±759	8.42E−04	10500	500	124	124	24.8
3σ 准则	Data11	1.0E−9	—	200	20	855±168.3	0.211	4351	595	167	167	28.1
3σ 准则	Data12	1.0E−9	—	300	20	336.8±74.8	26.969	992	35	7	7	20.S0

* 用 LS−SVM 计算数据集 Data7、Data8、Data9 和 Data10；用 ε−SVR 计算数据集 Data11 和 Data12；其它同表 1。

4 结 论

通过前述分析，可以得出如下结论：

1)回归问题中，$d=|y-f(x)|$ 可定义为样本点不满足回归映射关系 $x\mapsto y$ 的差异程度的一种度量。从 $\max|y_i-f(x_i)|$ 出发，每次剔除 $d=\max|y_i-f(x_i)|$ 的异常样本点，逐步剔除异常数据获得的回归模式序列 $f_0(x)$，$f_1(x)$，…，$f_{k-1}(x)$，$f_k(x)$ 逐步逼近回归映射关系 $x\mapsto y$ 的理论模式 $F(x)$。

2)逐步剔除异常数据获得的回归模式序列 $f(x)$ 是一个渐变序列。定理 4 与推论 2 的证明及算法理论分析显示逐步支持向量回归中剔除异常数据算法是收敛的。同时，算法理论分析还显示，剔除支持向量回归中异常数据算法时间花销没有大的增加，保持了不剔除异常数据直接训练 SVR(相同数量级)的时间复杂度，几乎没有额外的空间花销。

3)仿真结果显示：算法 1、算法 3 是有效的和鲁棒的；算法 1 和算法 3 比 3σ 准则命中率更高；剔除异常数据算法能有效提高最终 SVR 模型精度；在剔除大刻度样本集异常数据中，算法 3 的效率高出算法 1 数十甚至上百倍。

尚待继续研究的问题：理论上分析剔除支持向量回归中异常数据算法收敛速度。

参考文献：

[1] Barnett V, Lewis T. Outliers in statistical data[M]. 3rd ed. New York: John Wiley & Sons, 1994.

[2] 张德然. 可靠性统计与数据挖掘[J]. 西华师范大学学报：自然科学版, 2005, 26(3): 334-337.
ZHANG Deran. Reliability statistics & data mining[J]. Journal of China West Normal University: Natural Sciences, 2005, 26(3): 334-337.

[3] 黄洪宇，林甲祥，陈崇成，等. 离群数据挖掘综述[J]. 计算机应用研究, 2006, 23(8): 8-13.
HUANG Hongyu, LIN Jiaxiang, CHEN Chongcheng, et al. Review of outlier detection[J]. Application Research of Computers, 2006, 23(8): 8-13.

[4] 李云飞，黄继伟，朱宏. 双参数指数分布异常数据的检验[J]. 电子科技大学学报, 2005, 34(1): 127-130.
LI Yunfei, HUANG Jiwei, ZHU Hong. Detection of outliers from the two-parameter exponential distribution [J]. Journal of UEST of China, 2005, 34(1): 127-130.

[5] 李文东,张建军,乔昱亚.指数分布场合异常数据的检验[J].长春大学学报,2006,16(4):10-13.
LI Wendong, ZHANG Jianjun, QIAO Yuya. Testing for abnormal data in exponential distribution sample[J]. Journal of Changchun University, 2006, 16(4): 10-13.

[6] 乐立利,曾海群.双参数对数正态分布异常数据的检测方法[J].数学理论与应用,2009,29(1):29-32.
YUE Lili, ZENG Haiqun. Method of detection for outliers in two-parameter lognormal distribution[J]. Mathematical Theory and Applications, 2009, 29(1): 29-32.

[7] Knorr E M, Ng R T. Finding intensional knowledge of distance-based outliers[C]// Proceedings of the 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK. New York: ACM, 1999: 211-222.

[8] Knorr E M, Ng R T. Algorithm for mining distance-based outliers in large datasets[C]//Proceedings of the 24rd International Conference on Very Large Data Bases, August 24-27, 1998. New York, NY, USA. New York: ACM, 1998: 392-403.

[9] 李星毅,包从剑,施化吉,等.基于加权快速聚类的异常数据挖掘算法[J].计算机工程与应用,2007,43(35):153-155.
LI Xingyi, BAO Congjian, SHI Huaji, et al. Outlier data mining algorithms based on weighted fast clustering[J]. Computer Engineering and Applications, 2007, 43(35): 153-155.

[10] Zhang J, Wang H. A new pretreatment approach of eliminating abnormal data in discrete time series[C]// Proceedings of the 2005 IEEE International Geosciences and Remote Sensing Symposium, July 25-29, 2005, Seoul, South Korea. Piscataway: IEEE, 2005,1:665-668.

[11] 肖瑛,董玉华.基于小波网络的遥测视速度异常数据剔除方法[J].大连民族学院学报,2009,11(3):215-218,281.
XIAO Ying, DONG Yuhual. Eliminating abnormal data in apparent velocity of telemetry based on wavelet network [J]. Journal of Dalian Nationalities University, 2009, 11(3): 215-218,281.

[12] 翁小清,沈钧毅.基于滑动窗口的多变量时间序列异常数据的挖掘[J].计算机工程,2007,33(12):102-104.
WENG Xiaoqing, SHEN Junyi. Outlier mining for multivariate time series based on sliding window[J]. Computer Engineering, 2007, 33(12): 102-104.

[13] 宋美娇,唐常杰,乔少杰,等.基于局部 K-距离的靶场异常数据检测算法[J].四川大学学报:自然科学版,2008,45(6):1337-1340.
SONG Meijiao, TANG Changjie, QIAO Shaojie, et al. A new algorithm for outlier detection in rocketdrome based on local K-distance[J]. Journal of Sichuan University: Natural Science Edition, 2008, 45(6): 1337-1340.

[14] 周福娜,文成林,汤天浩,等.基于指定元分析的多故障诊断方法[J].自动化学报,2009,35(7):971-982.
ZHOU Funa, WEN Chenglin, TANG Tianhao, et al. DCA based multiple faults diagnosis method[J]. Acta Automatica Sinica, 2009, 35(7): 971-982.

[15] Ju K Y, Zhou D Q, Zhang Y Q. A novel algorithm for outlier detection in high dimension and its application in mine disaster forewarning[C]// Proceedings of the 4th International Conference on Wireless Communications, Networking and Mobile Computing, October 12-14, 2008, Dalian, China. Piscataway: IEEE, 2008: 1-7.

[16] 张登峰,陆宝春,王执铨.基于动态神经网络的非线性系统鲁棒故障检测[J].数学的实践与认识,2006,36(1):154-159.
ZHANG Dengfeng, LU Baochun, WANG Zhiquan. Robust fault detection based on dynamic neural network for nonlinear systems[J]. Mathematics in Practice and Theory, 2006, 36(1): 154-159.

[17] 施冬冬,贾瑞玉,黄义堂.基于遗传算法的高维离群点检测算法的改进[J].计算机技术与发展,2009,19(3):141-143,147.
SHI Dongdong, JIA Ruiyu, HUANG Yitang. An improved high-dimensional outlier detection algorithm based on genetic algorithm[J]. Computer Technology and Development, 2009, 19(3): 141-143,147.

[18] Chan K Y, Kwong C K, Fogarty T C. Modeling manufacturing processes using a genetic programming-based fuzzy regression with detection of outliers[J]. Information Sciences,2010,180(4): 506-518.

[19] 庞茂,周晓军,孟庆华.基于免疫学的在线故障检测算法的研究及应用[J].中国电机工程学报,2005,25(24):149-153.
PANG Mao, ZHOU Xiaojun, MENG Qinghua. Study and application of on-line fault testing algorithm based on immunology[J]. Proceedings of the CSEE, 2005, 25(24): 149-153.

[20] 徐雪松,宋东明,张谞,等.基于流形学习的离群点检测方法[J].中国工程科学,2009,11(2):82-87.
XU Xuesong, SONG Dongming, ZHANG Xu, et al. The research of detection of outliers based on manifold learning[J]. Engineering Science, 2009, 11(2): 82-87.

[21] 徐雪松,张谞,宋东明,等.基于非线性数据变换的离群点检测算法[J].中国工程科学,2008,10(9):74-78.

XU Xuesong, ZHANG Xu, SONG Dongming, et al. Outliers detection algorithm based on nonlinear data transformation[J]. Engineering Science, 2008, 10(9): 74-78.

[22] 周晓云,孙志挥,张柏礼,等.高维类别属性数据流离群点快速检测算法[J].软件学报,2007,18(4):933-942.
ZHOU Xiaoyun, SUN Zhihui, ZHANG Baili, et al. A fast outlier detection algorithm for high dimensional categorical data streams[J]. Journal of Software, 2007, 18(4): 933-942.

[23] 张宁.离群点检测算法研究[J].桂林电子科技大学学报,2009,29(1):22-25.
ZHANG Ning. Research of outlier detection algorithm[J]. Journal of Guilin University of Electronic Technology, 2009, 29(1): 22-25.

[24] 倪巍伟,陈耿,陆介平,等.基于局部信息熵的加权子空间离群点检测算法[J].计算机研究与发展,2008,45(7):1189-1194.
NI Weiwei, CHEN Geng, LU Jieping, et al. Local entropy based weighted subspace outlier mining algorithm[J]. Journal of Computer Research and Development, 2008, 45(7): 1189-1194.

[25] Isaksson C, Dunham M H. A comparative study of outlier detection algorithms[C]// Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition, July 23-25, 2009, Leipzig, Germany. Berlin, Heidelberg: Springer-Verlag, 2009:440-453.

[26] Mejía-Lavalle M, Obregón R G, Vivar A S. Outlier detection with a hybrid artificial intelligence method[C]// Proceedings of the 8th Mexican International Conference on Artificial Intelligence, November 9-13, 2009, Guanajuato, Mexico. Berlin, Heidelberg: Springer-Verlag, 2009,5845:590-599.

[27] Zhang K, Hutter M, Jin H D. A new local distance-based outlier detection approach for scattered real-world data[C]// Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, April 27-30, 2009, Bangkok, Thailand. Berlin, Heidelberg: Springer-Verlag, 2009, 5476:813-822.

[28] Angiulli F, Fassetti F. DOLPHIN: an efficient algorithm for mining distance-based outliers in very large datasets[J]. ACM Transactions on Knowledge Discovery from Data,2009,3(1):1-57.

[29] Yoon K A, Bae D H. A pattern-based outlier detection method identifying abnormal attributes in software project data[J]. Information and Software Technology, 2010,52(2):137-151.

[30] 黄彬彬,赵久奋,彭会釜,等.基于 Logistic 回归的识别算法适应性研究[J].四川兵工学报,2011,32(10):131-133.
HUANG Bingbing, ZHAO Jiufen, DENG Huifu, et al. Research on adaptability of recognition algorithm based on logistic regression[J]. Journal of Sichuan Ordnance, 2011, 32(10):131-133.

[31] Hawkins D M. Identification of outliers[M]. London: Chapman and Hall,1980.

[32] 陈希孺,王松桂.近代回归分析:原理方法及应用[M].合肥:安徽教育出版社,1987.

[33] Vapnik V N. Statistical learning theory[M]. New York: John Wiley & Sons,1998.

[34] Zeng S H, Tang Y Y, Wei Y, et al. Algorithm of ε-SVR based on a large-scale sample set: step-by-step search[J]. International Journal of Wavelets, Multiresolution and Information Processing, 2011, 9(2):197-210.

[35] Frank A, Asuncion A. UCI machine learning repository [EB/OL]. [2010-12-05]. http://archive.ics.uci.edu/ml/citation_policy.html.

(编辑 侯 湘)