

一种多率采样的在线支持向量回归及应用

李大海¹, 李天石¹, 李宗斌²

(1. 西安交通大学机械工程学院, 710049, 西安; 2. 西安交通大学机械制造系统工程国家重点实验室, 710049, 西安)

摘要: 针对应用支持向量回归对不确定控制系统在线建模时精度受异常数据影响的问题, 通过分析不同样本分布情况下异常数据的影响, 指出增加异常数据邻域的样本密度可以有效地提高建模精度. 提出了多率采样的支持向量回归在线建模方法, 通过多率采样增加局部样本密度, 并利用支持向量回归在小样本学习时的良好性能, 构建一种局部样本密集的滚动时间窗, 用以减少训练样本数和在线剔除异常数据. 将该方法应用于多通道电液力伺服同步加载系统的负荷输出预测, 结果表明, 与传统单率采样的方法相比, 在训练样本只增加 2 个的情况下, 该方法具有更好的鲁棒性和预测精度, 预测平均绝对误差达到了 0.66%.

关键词: 支持向量回归; 多率采样; 异常数据

中图分类号: TP271.4 **文献标志码:** A **文章编号:** 0253-987X(2010)03-0037-05

On-Line Support Vector Regression by Multirate Sampling with Application

LI Dahai¹, LI Tianshi¹, LI Zongbin²

(1. School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China; 2. State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: To improve on-line modeling precision of support vector regression (SVR) in uncertain control systems, by analyzing the effects of training data sets distribution and outliers, it is demonstrated that the precision can be effectively improved by increasing data density at neighborhood of outliers. A modeling method using multirate sampling is presented based on on-line SVR. In this method, multirate sampling is chosen to increase local data density, and a local-data-intensive sliding time window is established to reduce the training data number and eliminate outliers by taking advantage of the good performance of SVR in small sample. The method is applied to a multichannel electrohydraulic force servo synchronous loading system to predict the load output. Compared with the traditional single rate sampling method, the results indicate the better robustness and prediction accuracy. The prediction mean absolute percentage error gets 0.66% while only two training samples are added.

Keywords: support vector regression; multirate sampling; outlier

利用实时数据估计当前工作条件下的精确在线模型, 可以有效地保证对不确定系统的实时控制精度. 由于支持向量回归(SVR)解决了估计中的小样本、高维数和非线性等问题^[1-2], 近年来被广泛采用^[3-5], 但实际建模时, 在实时数据中不可避免地存在异常数据, 因此建模精度往往难以保证.

根据具体的系统来构造一个损失函数, 使 SVR 对异常数据具有鲁棒性^[6], 其中线性 ϵ 不敏感损失函数可使 SVR 具有计算复杂度低的特点, 因此更适合在线建模, 但有时却不能保证鲁棒性, 其原因在于: 为了保证实时性和模型的准确性, 训练样本通常被限定在一个滚动时间窗内^[3], 而传统更新样本的

方法只考虑了样本时间上的相关性,采用先入先出的原则,并未考虑窗内样本的分布.异常数据在周围样本稀疏时会导致过拟合,并且线性 ϵ 不敏感损失函数隐含了假定偏差服从同一分布^[2],因此不能通过调整SVR的参数进行改善.在锅炉燃烧优化的在线函数估计中,文献[4]通过保证样本均匀分布获得了很好的效果,但对控制系统在线建模时,窗内样本的分布取决于历史采样时刻系统的输入输出值,所以很难保证均匀.本文分析了在不同样本分布情况下,异常数据对建模精度的影响,提出了一种多率采样的支持向量回归(MS-SVR)的在线建模方法,并应用于多通道电液力伺服同步加载系统的负荷输出预测中.

1 基本理论

1.1 支持向量回归及其求解分析

给定训练样本集 $\{(\mathbf{x}_i, y_i), i=1, 2, \dots, l\}$,其中 $\mathbf{x}_i \in \mathbf{R}^n, y_i \in \mathbf{R}$.SVR通过把输入空间映射到高维特征空间,将非线性问题转化为线性问题.建立线性回归方程

$$f(\mathbf{x}) = \mathbf{W}^T \boldsymbol{\phi}(\mathbf{x}) + b \quad (1)$$

式中: $\boldsymbol{\phi}(\cdot)$ 是 $\mathbf{x}$ 到高维特征空间的映射; $\mathbf{W}$ 是高维特征空间中的权向量; b 是回归方程的偏置项. $\mathbf{W}$ 、 b 通过最小结构化风险^[2]求解,即

$$R[f]: = \frac{1}{l} \sum_{i=1}^l \hat{c}(\xi_i) + \frac{\lambda}{2} \|\mathbf{W}\|^2 \quad (2)$$

式中: λ 为正规化参数; $\hat{c}(\xi)$ 为损失函数; l 为样本数.SVR可描述为优化问题

$$\begin{aligned} \min_{\mathbf{W}, b, \xi_i, \xi_i^*} & \frac{1}{2} \|\mathbf{W}\|^2 + C \sum_{i=1}^l (\hat{c}(\xi_i) + \hat{c}(\xi_i^*)) \\ \text{s. t.} & y_i - \mathbf{W}^T \boldsymbol{\phi}(\mathbf{x}_i) - b \leq \epsilon + \xi_i \\ & \mathbf{W}^T \boldsymbol{\phi}(\mathbf{x}_i) + b - y_i \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, \quad i = 1, 2, \dots, l \\ & C = 1/\lambda l \end{aligned} \quad (3)$$

式中: ξ_i 为松弛变量.式(3)可以通过对偶形式求解,即

$$\begin{aligned} \min_{\alpha, \alpha^*} & \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\mathbf{x}_i, \mathbf{x}_j) + \\ & \sum_{i=1}^l [\epsilon(\alpha_i + \alpha_i^*) - (y_i - b)(\alpha_i - \alpha_i^*) - \\ & C(T(\xi_i) + T(\xi_i^*))] \\ \text{s. t.} & 0 \leq \alpha_i^{(*)} \leq C \frac{\partial \hat{c}(\xi_i^{(*)})}{\partial \xi_i^{(*)}}, \quad i = 1, 2, \dots, l \end{aligned} \quad (4)$$

$$T(\xi) = \hat{c}(\xi) - \xi \frac{\partial \hat{c}(\xi)}{\partial \xi}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}^T(\mathbf{x}_i) \boldsymbol{\phi}(\mathbf{x}_j)$$

式中: α, α^* 为拉格朗日乘子; $K(\mathbf{x}_i, \mathbf{x}_j)$ 为核函数.

对于具体的回归问题,求解的复杂性由损失函数决定.虽然希望损失函数符合实际问题,但复杂的损失函数却导致优化问题变得复杂且难以求解.采用 ϵ 不敏感损失函数 $\hat{c}(\xi) = |\xi|_P^P$,当 $P=1$ 时, $\frac{\partial \hat{c}(\xi)}{\partial \xi} = 1, T(\xi) = 0, \alpha \in [0, C]$;当 $P>1$ 时, α 的上界随 $\frac{\partial \hat{c}(\xi)}{\partial \xi}$ 增长;当 $P<1$ 时,式(3)非凸^[2].因此,采用线性 ϵ 不敏感损失函数求解的复杂度最低,并且由于 α 上界限定为 C ,所以对异常数据有一定的鲁棒性.

1.2 在线支持向量回归

令 $\theta_i (i=1, 2, \dots, c) = \alpha_i - \alpha_i^*, h(\mathbf{x}_i) = f(\mathbf{x}_i) - y_i$,根据式(4)的KKT(Karush-Kuhn-Tucker)条件,可将训练样本分为3类^[7],即

$$\begin{aligned} E &= \{i \mid h(\mathbf{x}_i) > \epsilon, \theta_i = -C\} \cup \{i \mid h(\mathbf{x}_i) < -\epsilon, \theta_i = C\} \\ S &= \{i \mid h(\mathbf{x}_i) = \epsilon, -C \leq \theta_i \leq 0\} \cup \{i \mid h(\mathbf{x}_i) = -\epsilon, 0 \leq \theta_i \leq C\} \\ R &= \{i \mid -\epsilon < h(\mathbf{x}_i) < \epsilon, \theta_i = 0\} \end{aligned}$$

回归方程式(1)可表示为

$$f(\mathbf{x}) = \sum_{j \in \text{SUF}} \theta_j \boldsymbol{\phi}^T(\mathbf{x}) \boldsymbol{\phi}(\mathbf{x}_j) + b \quad (5)$$

在线算法^[8]的思想是在保持训练样本集中样本始终满足KKT的条件(即具有类别属性,始终对应于 E, S, R 中的任何一个)下,将一个新样本加入或将旧样本移除.加入新样本 $c(x_c, y_c)$ 时,对应的 θ_c 的初值设为0,根据 $-h(x_c)$ 的符号逐步增加或减少 θ_c ,直到 c 也具有类别属性. θ_c 的改变会引起训练样本集中样本类别属性的改变,或使 c 具有类别属性,因此每次更新 θ_c 时, $\Delta\theta_c$ 的取值为引起这些变化的最小增量.剔除旧样本 i 的方法与此类似,不同之处为 θ_i 的初值为当前值,并视为无类别属性的样本.

2 多率采样的在线支持向量回归

2.1 不同样本分布时的估计性能

基于结构风险的最小化,SVR解决了小样本的学习问题,但在样本分布稀疏时,异常数据会使估计效果变差,为此,以 $y = \cos(x)$ 的估计为例来分析异常数据的影响.在训练样本个数为 n 、对 y 的叠加均

值为0、标准差为0.05的正态分布噪声中,选用高斯径向基核函数,宽度系数 $\sigma=1$,设置SVR的正则化参数为10.

2.1.1 样本分布不均匀时的估计性能 在相同的噪声水平下,无异常数据时,样本分布不均匀也能取得较好的估计效果.样本分布稀疏的区间出现异常数据时,容易导致过拟合,通过保证样本的均匀分布,可在一定程度上保证估计的精度^[4],但样本整体分布比较稀疏时,却难以消除异常数据的影响.

2.1.2 样本分布稀疏时的估计性能 样本分布稀疏时的表现为:①输入空间的范围小,样本数少,由式(2)可见异常数据对结构风险影响比较大;②输入空间范围大,但样本距离远,由式(5)可见 θ 的变化会引起 $\mathbf{W}$ 较大的变化.因此,在样本分布稀疏的情况下,异常数据会导致泛化性能的恶化(见图1).

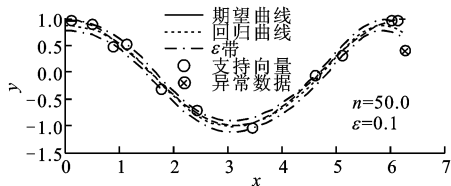


图1 样本分布稀疏时的估计结果

在图1中,样本在区间 $[0, 2\pi]$ 上均匀分布,可见异常数据引起了过拟合.由于异常数据一般产生于不同的机制,参数的选择互相关联并且与噪声相关,而且过于严格的参数设置会增加支持向量个数,也容易导致过拟合,因此通过调整SVR的参数来改善精度实际上是比较困难的.

由图2可见,减小 ϵ 不仅不能消除异常数据的影响,而且还会大幅增加支持向量的个数.

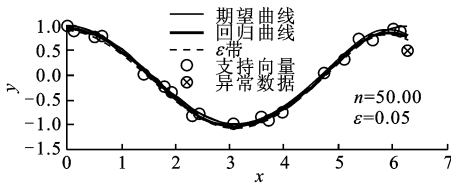


图2 减小 ϵ 后的估计结果

2.1.3 样本分布局部密集时的估计性能 由式(1)和式(5)可以看出, $\mathbf{W}$ 是映射空间中支持向量的线性组合,可以通过增加样本密度来降低异常数据的影响,但需要增加的样本数则很大.由于多变量函数的复杂性随输入空间的维数呈指数增长,因此整体增加密度是不可行的,实际上只要异常数据邻域的样本密度足够高,就可以得到比较好的估计结果,如图3所示.

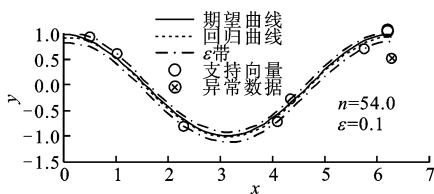


图3 增加异常数据邻域内样本密度的估计结果

与图1相比,在 $x=1.98\pi, 1.985\pi, 1.99\pi, 1.995\pi$ 处增加4个样本,由图3可见,所估计的精度比图1有较大的提高,支持向量的个数也有所减少.显然,对于样本分布不均匀的情况,增加局部样本密度是可行的方法.在线回归的速度依赖于训练样本个数和支持向量个数,因此增加局部样本密度的方法可以有效地提高计算速度.

2.2 局部样本密集的滚动时间窗

考虑不确定的慢时变非线性系统,并假定系统满足局部Lipschitz条件,用NARMAX模型描述^[5]系统

$$y(k+1) = f(\mathbf{x}(k)) \quad (6)$$

$$\mathbf{x}(k) = [y(k), \dots, y(k-n+1), u(k), \dots, u(k-m+1)]$$

式中: $u(k)$ 、 $y(k)$ 分别是系统在第 k 个采样时刻的输入输出; $f(\cdot)$ 是非线性函数.考虑参数的时变和系统工作域随时间迁移的因素,为了更准确地反映系统的当前状态,可构建一个滚动时间窗,将样本限制在一定时间范围内,利用窗内样本建模,模型随窗内样本在线更新^[3].由式(6)可见, $u(k)$ 、 $y(k)$ 是系统的固有特性、控制目标和控制律的反映,它决定了窗内样本分布一般是不均匀的,当采样率比较低时甚至可能是稀疏的.为了降低异常数据对建模精度的影响,需要足够高的采样率,以保证足够的样本密度.

由于异常数据出现的不确定性,单率采样时需要保证窗内样本分布最稀疏区间的样本密度,因此对采样率的要求比较高.实际上对无异常数据的区间,提高采样率是没有必要的,而且在实际应用中,采样率的提高也受到多种因素的限制.如果输出采样率不受限制,可考虑增加输出采样率来提高局部样本密度,即在控制输入刷新周期 $[kT, kT+T]$ 内,在 $kT+t_i$ 时刻对输出采样(其中 T 为帧采样周期, $i=1, 2, \dots, m, 0=t_1 < \dots < t_m < T$ 为 kT 时刻的密集样本).单率采样时,采用单入单出的方式构建窗内样本个数为 a 的时间窗,而多率采样时,每次出入时

间窗的样本为 m 个. 如果在窗内保留每个帧周期的密集样本, 则窗内样本个数为 am , 在线计算量将比较大. 在无异常数据时, 样本稀疏对回归估计精度的影响很小, 而增加局部样本密度时 SVR 对异常数据有很好的鲁棒性, 考虑到靠近 ϵ 带的样本最有可能成为下一次估计时的支持向量, 因此加入 m 个最新时刻的密集样本时, 对于窗内的上一个历史帧周期的样本, 只保留其中 $j(j < m)$ 个 $\|h(\mathbf{x})| - \epsilon\|$ 值较小的样本, 同时去掉 j 个时间最早的样本, 窗内样本个数为 $aj + m - j$. 窗内样本为

$$\begin{aligned} & \{(x_{(k-a)T}^1, y_{(k-a)T}^1), \dots, (x_{(k-a)T}^j, y_{(k-a)T}^j), \dots, \\ & (x_{(k-1)T}^1, y_{(k-1)T}^1), \dots, (x_{(k-1)T}^j, y_{(k-1)T}^j), \\ & (x_{kT}^1, y_{kT}^1), \dots, (x_{kT}^m, y_{kT}^m)\} \end{aligned} \quad (7)$$

由式(7)可见 kT 时刻的样本密度高于窗内历史时刻的样本密度, 因此称该时间窗是局部样本密集的滚动时间窗. 一方面新加入的样本是密集的, 可以保证模型不受异常数据的干扰, 另一方面删除了历史数据中潜在的异常数据, 但并不破坏样本在时间上的分布.

2.3 建模步骤

基于在线支持向量的回归算法, 使用局部样本密集的滚动时间窗建立模型的步骤为: ①初始化窗长度为 $aj - j$, 加入 kT 时刻的密集样本(m 个), 建立初始模型; ②预测 $kT + T$ 时刻的系统输出; ③剔除 $m - j$ 个 kT 时刻的密集样本中 $\|h(\mathbf{x})| - \epsilon\|$ 值比较大的样本, 更新模型; ④剔除时间最早的 j 个的样本, 更新模型; ⑤采样数据有效时, 逐个加入 $kT + T$ 时刻的密集样本, 更新模型; ⑥返回②.

3 应用研究

在分析多通道电液力伺服同步加载系统时, 通常将其简化为 1 个 3 阶线性系统^[9], 受多种因素影响, 简化模型与实际的数学模型差异很大, 因此不能直接用于控制的主要原因在于: ①加载点发生变化影响系统的模型及参数; ②加载力的分布及变化使试件发生严重的交叉耦合变形, 从而产生复杂的内力干扰; ③试件结构、尺寸及材料的不同, 将影响系统的模型及参数; ④系统固有的非线性因素; ⑤安装误差产生的附加力和附加力矩. 由于试验过程的特点, 加载力的控制量不可能快速变化, 而系统的复杂性和系统内的耦合会导致异常数据的产生. 因此, 采用 MS-SVR 对系统负荷输出进行预测, 并作为控制律调整的依据.

MS-SVR 选用高斯径向基核函数, 选取 $C=10$,

$\sigma=30, \epsilon=0.1, m=3, j=1, a=35$ (窗内样本数为 37), 对 $kT + T_0$ 时刻的负荷进行预测, $k=1, 2, \dots, 115, T_0$ 为初始化完成后的采样时刻. 采样设置为 $T=900\text{ ms}, t_1=0\text{ ms}, t_2=300\text{ ms}, t_3=600\text{ ms}$. 系统模型为

$$\begin{aligned} & y(kT + T) = f(\mathbf{x}(kT)) \quad (8) \\ & \mathbf{x}(kT) = [y(kT + t_3), \dots, y(kT - T + t_1), u(kT + t_3), \dots, u(kT - T + t_1)] \\ & u((k - q)T + t_i) = u((k - q)T) \\ & q = 0, 1, \quad i = 1, 2, 3 \end{aligned}$$

对 $kT + T_0$ 时刻负荷的预测, 需对 MS-SVR 进行 6 次样本更新($k=1$ 时为 3 次)后计算预测值, 最大运算时间为 0.584 s, 最小运算时间为 0.121 s, 累计运算时间为 27.792 s, 平均运算时间为 0.241 s. 其中, 第 6 次更新和预测值的计算在 $kT + T_0 + t_3$ 时刻开始, 平均运算时间为 0.032 s, 可见 MS-SVR 有比较好的实时性. 作为比较, 采用相同参数设置, 使用单采样率的支持向量回归(SS-SVR)进行预测, 窗内样本数为 35. 将输入输出数据分别进行归一化处理, 结果如图 4 所示. MS-SVR 和 SS-SVR 预测的平均绝对误差分别为 0.66% 和 1.12%, 可见 MS-SVR 的预测精度高于 SS-SVR. 平均绝对百分比误差的计算公式为

$$\delta = \frac{1}{l} \sum_{i=1}^l \left| \frac{(y_i - \hat{y}_i)}{y_i} \right| \times 100\% \quad (9)$$

式中: y_i 为实际值; $\hat{y}_i$ 为估计值.

由图 4a 可见, 当输出曲线斜率出现变化时, 样本密度改变, MS-SVR 由于样本中增加了当前时刻的动态特性, 预测精度高于 SS-SVR. 由图 4b 可以

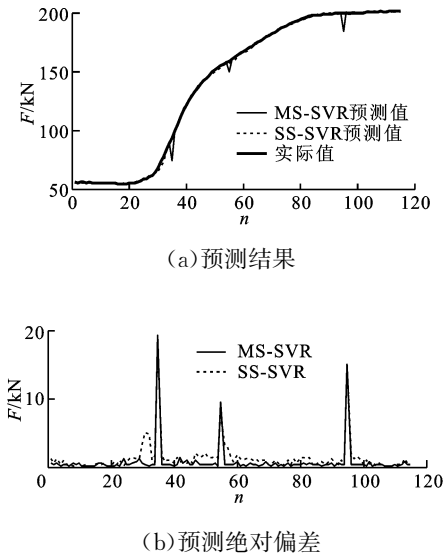


图 4 2 种方法预测结果的比较

看出,出现异常数据时 SS-SVR 的预测偏差比 MS-SVR 小,出现异常数据后,预测误差增加,说明异常数据使其回归方程的泛化性能变差. 由于 MS-SVR 采用了局部样本密集的时间窗,因此对异常数据具有很好的鲁棒性.

4 结 论

在不同样本的分布情况下,本文通过分析异常数据对建模精度的影响,针对控制系统在线建模时样本分布不均匀或稀疏的特点,提出了一种多率采样的支持向量回归在线建模方法. 该方法通过多率采样来增加局部样本密度,提高了建模精度. 综合考虑样本的时间相关性和期望的样本分布,构建了一种局部样本密集的滚动时间窗,在线剔除了异常数据,减少了训练样本数目. 将该方法应用于多通道电液力伺服同步加载系统中,对负荷输出进行了预测,结果表明,与传统单率采样的方法相比,在训练样本只增加 2 个的情况下,该方法有更好的鲁棒性和预测精度.

参考文献:

- [1] VAPNIK V N, GOLOWICH S E, SMOLA A. Support vector method for function approximation regression estimation and signal processing [C]//MOZER M C. Advances in Neural Information Processing Systems: 9. Cambridge, MA, USA: MIT Press, 1997: 281-287.
- [2] SMOLA A J, SCHOLKOPF B. A tutorial on support vector regression [J]. Statistics and Computing, 2004, 14: 199-222.
- [3] 阎威武,常俊林,邵惠鹤. 基于滚动时间窗的最小二乘支持向量机回归估计方法及仿真 [J]. 上海交通大学学报, 2004, 38(4): 524-532.
YAN Weiwu, CHANG Junlin, SHAO Huihe. Least square SVM regression method based on sliding time window and its simulation [J]. Journal of Shanghai Jiaotong University, 2004, 38(4): 524-532.
- [4] SI Fengqi, ROMERO C E, YAO Zheng, et al. A new approach for function approximation in boiler combustion optimization based on modified structural AOSVR

[J]. Expert Systems with Applications, 2009, 36(4): 8691-8704.

- [5] ZHANG Haoran, WANG Xiaodong, ZHANG Changjiang, et al. Using least square SVM for nonlinear systems modeling and control [J]. Asian Journal of Control, 2007, 9(2): 215-220.
- [6] ZHAN Y, CHENG H Z. A robust support vector algorithm for harmonic and interharmonic analysis of electric power system [J]. Electric Power Systems Research, 2005, 73(3): 393-400.
- [7] HASTIE T, ROSSET S, TIBSHIRANI R, et al. The entire regularization path for the support vector machine [J]. Journal of Machine Learning Research, 2004(5): 1391-1415.
- [8] MA Junshui, THEILER J, PERKINS S. Accurate on-line support vector regression [J]. Neural Computation, 2003, 15(11): 2683-2703.
- [9] 曹阳,李天石. 电液伺服协调加载系统特性分析及其自适应统一预测控制策略的试验研究 [J]. 机械科学与技术, 2002, 21(4): 620-622.
CAO Yang, LI Tianshi. Research on electrohydraulic servo harmony loading system and adaptive unified predictive control [J]. Mechanical Science and Technology, 2002, 21(4): 620-622.

[本刊相关文献链接]

加权支持向量机求解路径算法研究. 西安交通大学学报, 2008, 42(10): 1226-1229.
 ϵ 不敏感损失函数支持向量机分类性能研究. 西安交通大学学报, 2007, 41(11): 1315-1320.
 考虑损伤程度的设备运行可靠性研究. 西安交通大学学报, 2010, 44(1): 36-40.
 动态概率模型跟踪性能退化的实时可靠性评估方法. 西安交通大学学报, 2010, 44(1): 46-50.
 一种自适应局部线性嵌入与谱聚类融合的故障诊断方法. 西安交通大学学报, 2010, 44(1): 77-82.
 自适应滤波实时网络流量异常检测方法. 西安交通大学学报, 2009, 43(12): 1-5.

(编辑 管咏梅)