文章编号: 1001-0920(2010)08-1169-04

基于鲁棒学习的最小二乘支持向量机及其应用

张淑宁[a], 王福利[a,b], 尤富强[a,b], 贾润达[a]

(东北大学 a. 信息科学与工程学院, b. 流程工业综合自动化教育部重点实验室, 沈阳 110819)

摘 要: 鉴于最小二乘支持向量机比标准支持向量机具有更高的计算效率和拟合精度, 但缺少标准支持向量机的鲁棒性, 即当采样数据存在奇异点或者误差变量的高斯分布假设不成立时, 会导致不稳健的估计结果, 提出了一种鲁棒最小二乘支持向量机方法. 该方法在最小二乘支持向量机基础上, 通过引入鲁棒学习方法来获得鲁棒估计. 仿真分析及某湿法冶金厂的应用实例验证了该方法的可行性和有效性.

关键词: 最小二乘支持向量机; 奇异点; 鲁棒学习; 鲁棒估计; 草酸钴合成过程

中图分类号: TP273 **文献标识码:** A

Robust least squares support vector machine based on robust learning algorithm and its application

ZHANG Shu-ning[a], *WANG Fu-li*[a,b], *YOU Fu-qiang*[a,b], *JIA Run-da*[a]

(a. School of Information Science and Engineering, b. Key Laboratory of Integrated Automation of Process Industry of Ministry of Education, Northeastern University, Shenyang 110819, China. Correspondent: ZHANG Shu-ning, E-mail: zhangshn0221@163.com)

Abstract: Least squares support vector machine(LS-SVM) is computationally more efficient than the standard SVM, but unfortunately the robustness of standard SVM is lost. LS-SVM might lead to estimates which are less robust with respect to outliers on the data or when the assumption of a Gaussian distribution for error variables is not realistic. Therefore, an approach based on the robust least squares support vector machine(RLS-SVM) is proposed, in which robust learning algorithm(RLA) is employed to enhance the robust capability of LS-SVM. Finally, simulation analysis and the modeling of a typical plant for hydrometallurgy illustrate the effectiveness and feasibility of the presented method.

Key words: Least squares support vector machine; Outliers; Robust learning; Robust estimation; Cobalt oxalate synthesis process

1 引 言

支持向量机(SVM)技术是在小样本统计学习理论和结构风险最小化理论的基础上发展起来的一种新型的机器学习方法[1]. 一方面, 通过核函数将输入空间映射到高维空间, 确定该高维空间的最优超平面能够使模型获得最大的泛化能力; 另一方面, 通过求解凸二次规划, SVM 能够得到唯一的全局最优解, 不会陷入局部极小点.

为了避开SVM中求解耗时费力的二次规划问题, Suykens 等[2]提出了一种标准SVM的重要扩展: 最小二乘支持向量机(LS-SVM). 它用等式约束代替不等式约束, 极大地降低了运算时间, 且相对于常用的线性不敏感函数, LS-SVM不再需要指定逼近精度. 但LS-SVM也存在一定的缺点, 它失去了标准支持向量机的鲁棒性, 当考虑奇异点或者误差变量的高斯分布假设不成立时, LS-SVM采用的误差平方和评价函数会导致不稳健的估计结果.

通常情况下, 由于现场干扰以及过失差错等原因, 训练样本均有一定的奇异点存在, 文献[3-6]对奇异点的影响进行了研究. 奇异点的影响使LS-SVM模型在实际过程中的应用受到很大限制, 针对此问题, [7]提出了加权LS-SVM(WLS-SVM), 通过对样本误差进行加权来减小奇异点的影响. [8]通过引入模糊聚类的思想, 对每一个样本点进行加权, 从而使LS-

收稿日期: 2009-09-01; **修回日期:** 2009-12-21.

基金项目: 国家863计划项目(2006AA060201).

作者简介: 张淑宁(1983–), 男, 山东济宁人, 博士生, 从事复杂工业过程建模与优化的研究; 王福利(1957–), 男, 辽宁辽阳人, 教授, 博士生导师, 从事复杂工业过程建模与优化、故障诊断等研究.

SVM具有一定的鲁棒性. 二者均通过加权减小奇异点的影响来获得鲁棒性, 但权重是一个常数, 且容易受到LS-SVM参数选择的影响. [9]通过引入鲁棒学习提出了鲁棒RBF网络, 但是存在初始权值选择以及结构确定困难等问题.

针对上述问题, 提出了一种鲁棒LS-SVM(RLS-SVM), 该方法在LS-SVM的基础上, 通过引入鲁棒学习方法来获得鲁棒估计. 仿真分析及某湿法冶金厂的应用实例验证了该方法的可行性和有效性.

2 LS-SVM

SVM回归方法的主要思想是: 首先通过某种非线性变换将输入向量映射到高维特征空间; 然后在高维特征空间中构造线性最优决策函数, 使模型结构风险最小[10].

训练样本为$(x_i, y_i) \in R^m \times R$, $i = 1, 2, \cdots, N$. 其中: N为样本总数, m为样本空间的维数. 通过非线性映射$z_i = \varphi(x_i) \in R^d$, 将原$m$维输入空间映射到$d(d \gg m)$维特征空间. 在这个高维特征空间中采用线性函数$f(x) = w\varphi(x) + b$对其进行拟合, 并容许出现拟合误差, 目标是使回归模型在模型推广能力和经验风险之间找到最佳平衡点, 即结构风险最小.

在LS-SVM中确定最优超平面需要求解如下二次规划问题:

$$\min J(w, e) = \frac{1}{2} w^{\mathrm{T}} w + \frac{1}{2}\gamma \sum_{i=1}^{N} e_i^2,$$

$$\text{s.t. } y_i = w\varphi(x_i) + b + e_i. \tag{1}$$

其中: $J(w, e)$为结构风险; γ为经验风险调整因子; e_i为容许误差, $i = 1, 2, \cdots, N$. 由统计学习理论可知$w^{\mathrm{T}} w$控制模型的推广能力.

利用Lagrange法求解式(1)的优化问题, 根据KKT优化条件整理得到如下线性方程组:

$$\begin{bmatrix} 0 & \mathbf{1}_N^{\mathrm{T}} \\ \mathbf{1}_N & \Omega + \dfrac{1}{\gamma} I_N \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix}. \tag{2}$$

其中: $Y = [y_1\ y_2\ \cdots\ y_N]^{\mathrm{T}}$, $\mathbf{1}_N = [1\ 1\ \cdots\ 1]^{\mathrm{T}}$, I_N为单位阵, $\Omega_{i,j} = \varphi(x_i)^{\mathrm{T}}\varphi(x_j)$, $i, j = 1, 2, \cdots, N$, $\alpha = [\alpha_1\ \alpha_2\ \cdots\ \alpha_N]^{\mathrm{T}}$.

根据Mercer条件定义核函数为

$$K(x_i, x_j) = \varphi(x_i)^{\mathrm{T}}\varphi(x_j). \tag{3}$$

并可以得到LS-SVM回归模型为

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i) + b. \tag{4}$$

其中: N为输入样本的个数, x_i为第i个输入样本, x为某一个输入变量. 选择不同的核函数能够建立不同的LS-SVM模型, 模型的性能也不同[11]. 本文选用径向基核函数, 即

$$K(x, x_i) = \exp\Big(-\frac{\|x - x_i\|^2}{2\sigma^2}\Big), \tag{5}$$

其中σ为径向基宽度.

3 基于RLA的LS-SVM

LS-SVM因训练过程中惩罚项参数γ是固定的, 导致模型对某些特殊情形(如奇异点和噪声)过分敏感, 出现过拟合现象[12]. 工业现场数据因受到各种形式的干扰而存在奇异点, 如果算法不具备稳定性和鲁棒性, 则会造成算法适用范围较小. WLS-SVM虽然提高了LS-SVM的鲁棒性, 但其权值的设置是一个常数, 不能随着奇异点以及模型预测误差分布变化进行更新, 因此其应用范围也受到了很大限制. 针对上述问题, 本文提出了一种基于鲁棒学习的LS-SVM.

3.1 鲁棒代价函数

传统学习方法采用最小二乘代价函数主要是为了减少训练模型的预测误差, 而鲁棒学习方法采用鲁棒代价函数的目的是尽量减小大误差点在建模过程中的作用, 有效减少“过拟合”现象. 最小二乘代价函数和鲁棒代价函数分别为

$$E_L(t) = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{2} e_j^2(t), \tag{6}$$

$$E_R(t) = \frac{1}{N}\sum_{j=1}^{N} \sigma(e_j(t)). \tag{7}$$

其中: t为迭代次数, N为训练样本数目, $e_j(t) = y_j - \hat{y}_j(t)$为第$t$次迭代模型预测误差, $\sigma(\cdot)$为鲁棒代价函数. 本文采用双曲正切估计函数为鲁棒代价函数[13], 即

$$\sigma(e_j(t)) = \begin{cases} \dfrac{1}{2} e_j^2(t),\ 0 \leqslant |e_j(t)| < a(t); \\ \dfrac{1}{2} a^2(t) + \dfrac{c_1}{c_2} \ln\Big[\dfrac{\cosh(c_2(b(t) - a(t)))}{\cosh(c_2(b(t) - |e_j(t)|))}\Big], \\ \qquad a(t) \leqslant |e_j(t)| \leqslant b(t); \\ \dfrac{1}{2} a^2(t) + \dfrac{c_1}{c_2} \ln[\cosh(c_2(b(t) - a(t)))], \\ \qquad b(t) < |e_j(t)|. \end{cases} \tag{8}$$

其中: $a(t)$和$b(t)$是区间端点; c_1和c_2是常数, 本文选定$c_1 = 1.73$, $c_2 = 0.93$. 最小二乘代价函数和鲁棒代价函数曲线如图1所示. 鲁棒代价函数的导数为

$$\phi(e_j(t)) = \begin{cases} e_j(t),\ 0 \leqslant |e_j(t)| < a(t); \\ c_1 \tanh[c_2(b(t) - |e_j(t)|)]\mathrm{sign}(e_j(t)), \\ \qquad a(t) \leqslant |e_j(t)| \leqslant b(t); \\ 0,\ b(t) < |e_j(t)|. \end{cases} \tag{9}$$

其中: $\mathrm{sign}(x)$ 为符号函数, 当 $x<0$ 时, $\mathrm{sign}(x)=-1$; 当 $x>0$ 时, $\mathrm{sign}(x)=1$.

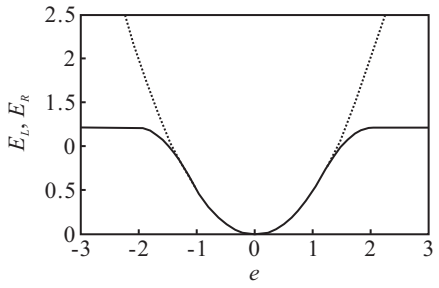

图 1　最小二乘代价函数与鲁棒代价函数

从图 1 可以看出, 鲁棒代价函数是对大误差点的影响有限幅作用的函数, 即在 $0 \leqslant |e_j(t)| < a(t)$ 范围内, 因为 LS-SVM 的权值已接近最优值, 所以其函数形式与最小二乘代价函数一致; 而在 $a(t) \leqslant |e_j(t)| \leqslant b(t)$ 范围内, 因误差较大, 对具有较大误差的点的影响进行一定程度的限幅; 在 $b(t) < |e_j(t)|$ 范围内, 因误差非常大, 不考虑其在建模过程的作用. 综上, 上述过程是一个大误差点的影响逐渐消除的过程.

3.2　RLS-SVM 算法步骤

RLS-SVM 是鲁棒学习与 LS-SVM 的结合, 在 LS-SVM 的基础上, 通过引入鲁棒学习算法来提升 LS-SVM 建模鲁棒性. 假定训练样本的置信度为 $1-q$, 算法具体步骤如下:

Step 1: 算法初始化. 确定 LS-SVM 的核函数类型及参数, 并给定区间端点初始值 $a(0), b(0)$ 和算法终止条件 ε_R.

Step 2: 计算模型预测误差. 训练 LS-SVM 模型, 计算模型预测值 $f(x_j,\alpha)$ 及预测误差 $e_j(t)=y_j(t)-f(x_j,\alpha),\ j=1,2,\cdots,N$.

Step 3: 区间端点更新[14]. 对 $e_j(t)$ 进行升序排序, 取 $j^*=(1-q)N$, 得到 $a(t)=e_{j^*}$ 和 $b(t)=2a(t)$.

Step 4: 权值更新. 应用梯度下降法更新 RLS-SVM 权值 (Lagrange 乘子), 即

$$\alpha(t+1)=\alpha(t)+\Delta\alpha(t), \tag{10}$$

其中 $\Delta\alpha(t)$ 为下降梯度, 且

$$\Delta\alpha(t)=\eta\frac{\partial E_R(t)}{\partial\alpha}=\frac{\eta}{N}\sum_{j=1}^{N}\varphi(e_j(t))\frac{\partial e_j(t)}{\partial\alpha}, \tag{11}$$

η 为学习因子, $\varphi(e_j(t))=\partial\sigma(e_j(t))/\partial e_j(t)$ 为鲁棒代价函数的导数, $\partial e_j(t)/\partial\alpha=-K(x_i,x_j)$.

Step 5: 终止判断. 若 $E_R \leqslant \varepsilon_R$ 不成立, 则转至 Step 2; 否则终止循环学习, 计算模型预测输出.

4　仿真分析

为了验证文中所提方法的有效性, 在 Matlab 环境下进行了仿真分析. 本文采用的仿真函数为

$$f(x)=\sin(2x)/2x,\ x\in[-3,3]. \tag{12}$$

其中: 100 个训练样本和 45 个测试样本分别在区间均匀产生, 且训练样本期望输出加上了高斯分布误差, 5 个奇异点是人为将某些点偏离正常位置所产生的.

为了便于比较 RLS-SVM 与其他支持向量机的性能, 本文针对不考虑奇异点和考虑奇异点两种情况进行了仿真分析, 如图 2 和图 3 所示.

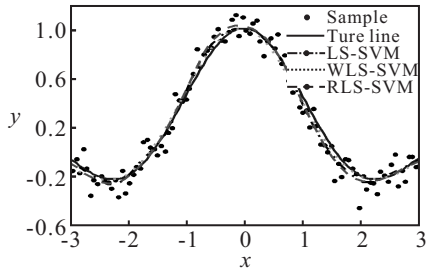

图 2　不考虑奇异点时仿真结果

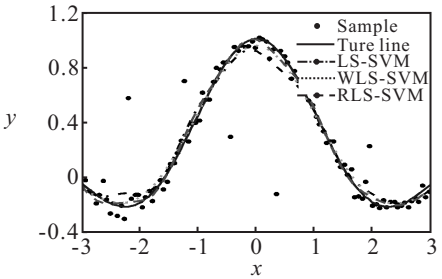

图 3　考虑奇异点时仿真结果

由图 2 和图 3 可以看出, 在不考虑奇异点情况下, 3 种模型均获得了较好的拟合效果, 而当训练数据受到奇异点影响时, RLS-SVM 的预测精度优于其他两种方法, 这说明本文所提出的 RLS-SVM 方法鲁棒性较强.

为了便于比较, 模型的最大绝对值误差 (MAE) 和均方根误差 (RMSE) 如表 1 所示.

表 1　预测误差比较

		LS-SVM	WLS-SVM	RLS-SVM
with outliers	MAE	0.1298	0.0521	0.0460
	RMSE	0.0545	0.0315	0.0234
without outliers	MAE	0.0845	0.0651	0.0874
	RMSE	0.0364	0.0334	0.0441

5　RLS-SVM应用

5.1　合成过程分析及辅助变量选择

草酸钴作为制备钴粉的主要原料可用于硬质合金, 随着硬质合金工业的发展, 对钴粉粒度提出了更高要求, 但是合成过程的重要控制指标草酸钴粒度却很难进行在线检测. 因此, 如何建立预报性能较好的草酸钴粒度软测量模型是目前需要解决的问题. 草酸钴合成工艺现场比较简陋, 由于人为因素或检测仪表故障使采样数据受到污染, 从而影响模型的预测精度.

针对上述问题, 将本文提出的RLS-SVM应用到草酸钴粒度分布预测模型.

根据机理分析, 选取搅拌电机转速、草酸铵流速和合成釜温度作为辅助变量. 草酸铵溶液浓度和氯化钴溶液浓度虽然也是草酸钴粒度分布的重要影响因素, 但鉴于现场实际情况, 二者在合成过程中基本是一个常数, 此处不将二者作为模型的辅助变量.

5.2 数据选择与规范化

首先建立训练和测试数据集, 但由于人为记录和生产故障, 数据库中存在被污染(即某些数据值不在期望的值域内或具有与期望不同的类型)数据, 如由于人为原因导致数据记录不完整或数据异常等情况. 如果将这些奇异数据用于建模, 必将严重影响模型的性能. 为此, 采用如下方法对数据库中的数据进行预处理是十分必要的: 考虑到z-score规范化无需预先知道属性的最大值和最小值, 且可以显著地减小噪声点对规范化的影响, 本文选择z-score规范化作为数据预处理方法. 对于任何有效数据, 采用z-score规范化处理$\hat{X} = X - \bar{X}/S$, 其作用是将属性X的值基于平均值$\bar{X}$和标准差S规范化.

5.3 模型比较与分析

经数据选择和规范化处理, 选择122组数据作为训练样本, 45组数据作为测试样本. 利用训练样本分别训练基于LS-SVM, WLS-SVM和RLS-SVM的模型, 然后应用训练好的模型进行草酸钴粒度分布预测. 预测结果如图4所示, 预测误差比较如表2所示.

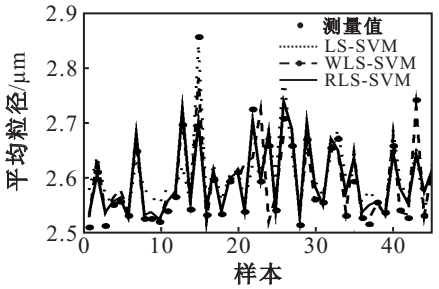

图 4 预测结果比较

表 2 预测误差比较

	LS-SVM	WLS-SVM	RLS-SVM
MAE	0. 1007	0. 1382	0. 1549
RMSE	0. 0393	0. 0388	0. 0343

由图4可以看出, 将本文提出的方法应用到草酸钴合成过程中会取得满意的效果, 其预测精度较LS-SVM和WLS-SVM都高. 特别是对于第15个样本点, 表面上LS-SVM计算精度高, 实际上由于LS-SVM无法辨别奇异点而造成过拟合, 导致整体预测精度较低. 虽然WLS-SVM整体预测精度有所提高, 但也明显低于本文方法. 由表2的预测误差比较也可以验证这一结论.

6 结 论

本文提出了一种RLS-SVM方法, 通过采用鲁棒学习算法来提高LS-SVM的鲁棒性, 避免了LS-SVM模型训练中的过拟合现象, 使模型可以避免奇异点影响. 通过仿真例子验证了本文方法拟合精度高、鲁棒性强, 将其应用于草酸钴合成过程中进行草酸钴粒度预测, 提高了模型的精度, 并取得了满意的效果.

参考文献(References)

[1] Vapnik V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.

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

[3] Pell R J. Multiple outlier detection for multivariate calibration using robust statistical techniques[J]. Chemometrics and Intelligent Laboratory Systems, 2000, 52(1): 87-104.

[4] Daszykowski M, Kaczmarek K, Heyden Y V, et al. Robust statistics in data analysis—A review basic concepts[J]. Chemometrics and Intelligent Laboratory Systems, 2007, 85(1): 203-219.

[5] Mølle S F, Frese J V, Bro R. Robust methods for multivariate data analysis[J]. J of Chemometrics, 2005, 19(10): 549-563.

[6] Kruger U, Zhou Y, Wang X, et al. Robust partial least squares regression: Algorithmic developments[J]. J of Chemometrics, 2008, 22(1): 1-13.

[7] Suykens J A K, Brabanter J D, Lukas L, et al. Weighted least squares support vector machines: Robustness and sparse approximation[J]. Neurocomputing, 2002, 48(1): 85-105.

[8] Shim J, Hwang C, Nau S. Robust LS-SVM regression using fuzzy C-means clustering[J]. Advances in Natural Computer Science, 2006, 42(21): 157-166.

[9] Sánchez A V D. Robustization of a learning method for RBF networks[J]. Neurocomputing, 1995, 9(1): 85-94.

[10] 常玉清, 王福利, 王小刚, 等. 基于支持向量机的软测量方法及其在生化过程中的应用[J]. 仪器仪表学报, 2006, 27(3): 241-244.
(Chang Y Q, Wang F L, Wang X G, et al. Soft sensor modeling based on support vector machines and its applications to fermentation process[J]. Chinese J of Scientific Instrument, 2006, 27(3): 241-244.)

(下转第1177页)