文章编号：0258-8013 (2009) 08-0092-05　　中图分类号：TK 264　　文献标志码：A　　学科分类号：470·20

基于支持向量机的回归预测和异常数据检测

王雷[1]，张瑞青[1]，盛伟[1]，徐治皋[2]

(1. 沈阳工程学院动力工程系，辽宁省 沈阳市 110136;
2. 东南大学能源与环境学院，江苏省 南京市 210096)

Regression Forecast and Abnormal Data Detection Based on Support Vector Regression

WANG Lei[1], ZHANG Rui-qing[1], SHENG Wei[1], XU Zhi-gao[2]

(1. Thermal Power Engineering Department, Shenyang Institute of Engineering, Shenyang 110136, Liaoning Province, China;
2. School of Energy and Environment, Southeast University, Nanjing 210096, Jiangsu Province, China)

ABSTRACT: In order to solve the general phenomena that abnormal data existing in the real-time data, based on support vector regression (SVR), a method of detecting abnormal data was presented. The model does not intend to eliminate individual major error while consider the smooth of the regression curve entirety on the whole, and distinguishes the abnormal data by comparing the regression value with the operational data. Main steam flow rate in 300 MW thermal power plant is selected to be the studying object，simulation results show that the method can determine the location and separate the abnormal data from the real-time data correctly, which establishes the foundations for enhancing the precise for the performance of power plant calculation.

KEY WORDS: thermal power engineering; abnormal data; real-time data; support vector regression; main steam flow rate

摘要：针对电站实时数据中普遍存在异常数据的情况，提出一种基于支持向量回归的异常数据检测方法。该方法利用支持向量回归估计模型不倾向于消除个别大的回归误差，而是从整体上考虑回归曲线的平滑程度的特点，通过比较回归估计值与实测值之间的残差来识别测量数据中的异常数据。以某 300 MW机组的主蒸汽流量为具体对象，给出了该方法的算例和检测结果，计算结果表明该方法能有效地实现异常数据的定位与分离，为提高电站性能计算的精确性奠定基础。

关键词：热能动力工程；异常数据；实时数据；支持向量回归；主蒸汽流量

0　引言

目前，电站的运行性能计算和优化依赖于大量现场参数的测量值，一些分析优化算法都是以所需的参数能被准确测量为前提的。但是由于各种原因，运行过程中不可避免的会出现测量参数失效，而运行过程中只要有一个参数失效就可能使各种数学模型中的某些运算无法进行，使得各种运行优化性能计算系统的现场应用受到限制，从而影响对设备的运行状态的监测，进而影响操作指导、经济性分析的准确度，严重的甚至关系到安全性问题。研究表明，关键参数的测量数据即使发生 1%的漂移也可能造成非常显著的热耗计算值的变化[1]，而往往这样的偏差是很难被直观地检测出来的，这就是测量过程中所谓的异常数据现象。由于参数的随机性很强，异常数据很难避免，因此优化性能计算系统必须具有数据检测能力，即能对实时测量参数进行有效检测，从而达到剔除异常数据的目的。

异常数据检测方法已经有很多种，如基于机理模型的预测方法[2]、统计学方法[3-4]、基于距离的检测方法[5-6]、神经网络方法[7-8]以及最近刚刚兴起的支持向量机方法等。基于支持向量机的异常数据检测方法是一个新兴的研究方向，在电站实时数据检测中的研究刚刚展开。本文利用支持向量机算法中结构风险函数较好的平滑性，建立基于支持向量回归的数据估计模型，实现异常数据的检验和分离。

1 基于支持向量回归的异常数据检测

1.1　支持向量回归算法

支持向量机(support vector machine，SVM)是由Vapnik[9]最早提出的一种统计学习方法，近年来已经被成功地应用于语音识别、文字识别、时序数列预测等领域。该方法具有学习速度快、全局最优和泛化能力强的优点，其学习结果明显好于其他的回

归预测方法，并在很多领域得到了成功应用[10-17]。

将支持向量机用于解决回归问题即支持向量回归(support vector regression，SVR)。SVR用来解决回归预测的基本思想如下：对样本集为$(\boldsymbol{x}_1, y_1),\ldots,(\boldsymbol{x}_l,y_l)$的回归问题，首先寻找一个输入空间到输出空间的非线性映射$\boldsymbol{\Phi}$，通过这个映射，将样本集中的数据映射到高维空间$\boldsymbol{F}$，并在特征空间中用下述线性函数在这个空间中进行线性回归，即

$$f(\boldsymbol{x}) = \boldsymbol{w}\cdot\boldsymbol{\Phi}(\boldsymbol{x})+b,\ \boldsymbol{\Phi}: R^n \to \boldsymbol{F},\ \boldsymbol{w}\in\boldsymbol{F} \tag{1}$$

式中b为阈值。由于$\boldsymbol{\Phi}$固定不变，因此影响$\boldsymbol{w}$的有经验风险的总和以及使其在高维空间平坦的$\|\boldsymbol{w}\|^2$，即

$$R(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{l}\varepsilon(f(\boldsymbol{x}_i)-y_i) \tag{2}$$

式中：l 表示样本的数目；$\varepsilon()$是损失函数，损失函数只计算$[-\varepsilon, \varepsilon]$以外的样本点，其定义如下：

$$\varepsilon(f(\boldsymbol{x}_i)-y_i)=\begin{cases}0, & |f(\boldsymbol{x}_i)-y_i|<\varepsilon\\ f(\boldsymbol{x}_i)-y_i|-\varepsilon, & \text{其他}\end{cases} \tag{3}$$

为控制函数的复杂性应使线性回归函数尽量平坦，亦即使$\boldsymbol{w}$的欧拉范数$\|\boldsymbol{w}\|^2$最小，并考虑可能超出精度的拟合误差，引入松弛因子$\xi_i,\xi_i^*\ge 0$，以处理未满足式(3)的数据点。根据统计学习理论的结构风险化准则，支持向量回归方法通过极小化目标函数来确定式(1)中的 $\boldsymbol{w}$ 和 b，即求下式最小：

$$R(\boldsymbol{w}, \xi_i, \xi_i^*) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l}(\xi_i+\xi_i^*)$$
$$\text{s.t.}\begin{cases} y_i - \boldsymbol{w}\cdot\boldsymbol{\Phi}(\boldsymbol{x}_i) - b \le \varepsilon+\xi_i\\ \boldsymbol{w}\cdot\boldsymbol{\Phi}(\boldsymbol{x}_i)+b-y_i\le\varepsilon+\xi_i^*\\ \xi_i,\xi_i^*\ge 0\end{cases} \tag{4}$$

式(4)中第 1 项使回归函数更为平坦，泛化能力更好，第 2 项则为减少误差，常数 C>0 用来平衡回归函数 f 的平坦程度和偏差大于ε样本点的个数。

引入拉格朗日乘子 $\alpha_i,\alpha_i^*,\lambda_i,\lambda_i^*$，建立拉格朗日方程：

$$L(\boldsymbol{w},b,\alpha,\xi) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l}(\xi_i+\xi_i^*) -$$
$$\sum_{i=1}^{l}\alpha_i(\varepsilon+\xi_i-y_i+\boldsymbol{w}\cdot\boldsymbol{\Phi}(\boldsymbol{x}_i)+b) -$$
$$\sum_{i=1}^{l}\alpha_i^*(\varepsilon+\xi_i^*+y_i-\boldsymbol{w}\cdot\boldsymbol{\Phi}(\boldsymbol{x}_i)-b) -$$
$$\sum_{i=1}^{l}(\lambda_i\cdot\xi_i+\lambda_i^*\cdot\xi_i^*) \tag{5}$$

要使式(5)取得最小值，对于参数 $\boldsymbol{w},b,\xi_i,\xi_i^*$ 的偏导都应等于零，即

$$\begin{cases}\dfrac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\cdot\boldsymbol{\Phi}(\boldsymbol{x}_i) = 0\\ \dfrac{\partial L}{\partial b} = \sum_{i=1}^{l}(\alpha_i-\alpha_i^*) = 0\\ \dfrac{\partial L}{\partial \xi_i} = C-\alpha_i-\lambda_i = 0\\ \dfrac{\partial L}{\partial \xi_i^*} = C-\alpha_i^*-\lambda_i^* = 0\end{cases} \tag{6}$$

将式(6)代入式(5)，可以得到对偶优化问题：

$$\min\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(\boldsymbol{\Phi}(\boldsymbol{x}_i)\cdot\boldsymbol{\Phi}(\boldsymbol{x}_j)) -$$
$$\sum_{i=1}^{l}\alpha_i(\varepsilon-y_i) - \sum_{i=1}^{l}\alpha_i^*(\varepsilon+y_i)$$
$$\text{s.t.}\begin{cases}\sum_{i=1}^{l}(\alpha_i-\alpha_i^*) = 0\\ \alpha_i,\alpha_i^*\in[0,C]\end{cases} \tag{7}$$

由此，支持向量机的函数回归问题就可以归结为二次规划问题。求解该二次规划问题，可以得到用数据点表示的 $\boldsymbol{w}$：

$$\boldsymbol{w} = \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\boldsymbol{\Phi}(\boldsymbol{x}_i) \tag{8}$$

式中 α_i 和 α_i^* 是最小化 $R(\boldsymbol{w})$ 的解。由此可求得线性回归函数：

$$f(\boldsymbol{x}) = \boldsymbol{w}\cdot\boldsymbol{\Phi}(\boldsymbol{x})+b = \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)(\boldsymbol{\Phi}(\boldsymbol{x}_i)\cdot\boldsymbol{\Phi}(\boldsymbol{x}))+b =$$
$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(\boldsymbol{x},\boldsymbol{x}_i)+b \tag{9}$$

式中 $K(\boldsymbol{x}_i,\boldsymbol{x}) = \boldsymbol{\Phi}(\boldsymbol{x}_i)\cdot\boldsymbol{\Phi}(\boldsymbol{x})$，$K(\boldsymbol{x}_i,\boldsymbol{x})$ 称为核函数。选择不同形式的核函数就可以生成不同的支持向量机，常用的核函数有：径向基函数、多项式函数、感知器(Sigmoid)函数、线性函数等。

1.2　基于支持向量回归的异常数据检测

数据中含有异常值，是造成回归估计不理想的主要原因。用含有异常值的数据建立回归模型，通常会在进行回归估计时使回归曲线拉向异常数据，不仅极大地影响回归曲线拟合的效果，而且此时观测值(异常数据)和回归值之间残差反而小，难以区分回归估计值和观测值的孰真孰伪，因而易造成对设备运行状态的判断的失误。

支持向量回归算法中的结构风险函数具有较好的平滑性，它并不倾向于消除个别大的回归误差，而是通过ε不敏感损失函数(式(3))的作用来控制异常数据对回归模型的影响，从整体上考虑回归曲线的平滑性，这样就使异常数据与回归值之间的残

差(距离)十分显著，容易观测。

图 1 给出了在含有异常数据时支持向量方法与其他方法的回归估计比较。

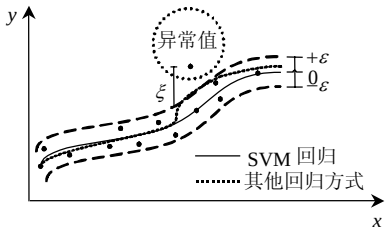

图 1 含有异常数据时的回归比较

Fig. 1 Regression comparison with the samples including abnormal data

支持向量回归将数据划分成支持向量和非支持向量，而支持向量又被分成边界支持向量和非边界支持向量。非支持向量和边界支持向量都位于损失函数的[−ε, ε]区域(ε-不敏感区域)以内，只有非边界支持向量才可能位于[−ε, ε]以外。

由于异常数据与其他观测值不是来自同一模型，因此，它最可能处在[−ε, ε]区域以外。根据式(3)~(7)可知，[−ε, ε]区域以外的样本点所对应的 α_i 及 α_i^* 参数一定满足 $\alpha_i = C$ 或者 $\alpha_i^* = C$ 。

结合对支持向量回归算法原理的分析，可以给出基于支持向量回归的异常数据检测步骤如下[18]：

1）利用实时数据，建立支持向量回归估计模型。

2）在利用模型进行回归估计时，若满足条件 $\alpha_i = C$ 或者 $\alpha_i^* = C$ 时，计算该样本点的残差，即计算 $E_i = |y_i - \hat{y}_i|$，y_i为回归估计值，$\hat{y}_i$ 为实际测量值，i 为满足条件的样本点序号。

3）根据实际需要以及所要测量数据的精度要求，定义一个常数 σ，如果 $E_i > \sigma$，则认为第 i 个样本 $\hat{y}_i$ 为异常数据。

由上面步骤得到计算流程图如图 2 所示。

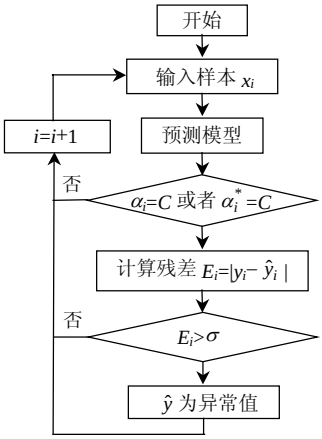

图 2 基于 SVR 的数据异常数据检测流程图

Fig. 2 Detection flow chart of abnormal data based-on SVR

2 主蒸汽流量异常数据检测

在火力发电厂机组运行过程中，主蒸汽流量测量不准确将影响煤耗、热耗计算的准确性，因此，准确测量主蒸汽流量对于提高机组经济性在线分析的可靠性具有重要的意义。目前对于大容量机组测量主蒸汽流量的方法主要是利用汽轮机调节级后的压力等参数计算求得。但由于在机组变工况过程中，调节级后压力与流量的关系呈非线性，并且当机组通流部分发生结垢时，由调节级后压力等参数计算得到的主蒸汽流量存在较大的偏差，从而产生异常数据。目前一种可行办法是利用回归方法建立各运行工况下主蒸汽流量与有关参数之间的关系模型，运行中，将该工况下的有关参数代入建立好的模型中，实现主蒸汽流量的回归测量。

首先利用SVR建立主蒸汽流量回归估计模型，以某电厂 300 MW机组为例，根据主蒸汽测量的原理，选择机组负荷、主蒸汽压力、主蒸汽温度、调节级后压力为模型输入参数，输出为主蒸汽流量[19-20]，在机组的正常运行范围内通过DCS采集了各测点对应的运行参数，采样间隔 1 min，共 180 个样本，其中 90 个作为训练样本，90 个作为测试样本集，在训练前对数据进行了归一化。通过择优选择模型参数C=1000，ε=0.001，核函数取径向基函数，其宽度系数σ=2。图 3 显示了模型的预测结果。可以看出，预测值(回归值)和实际值的吻合程度是比较好的，误差满足工程误差允许范围。

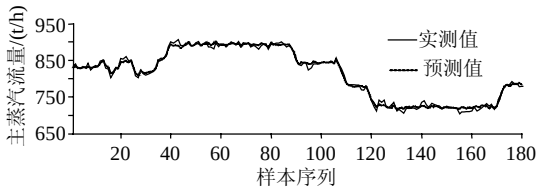

图 3 样本没有异常数据时主蒸汽流量的 SVR 预测

Fig. 3 Main steam flow rate forecast based-on SVR with the normal train samples

为了验证基于 SVR 的异常数据检测方法，在所有训练样本(主蒸汽流量)中加入故障异常数据代替原来的训练样本，用来模拟现场测量过程时出现的异常数据。异常数据为

$$y_i = y_i + a\delta_i \tag{10}$$

式中：a 为定值；δ_i 为单位函数，即

$$\delta_i = \begin{cases} 0, & \text{样本序列号} \neq i \\ 1, & \text{样本序列号} = i \end{cases} \tag{11}$$

令增量a=50 t/h，使带有异常数据的主蒸汽流量样本序列为

$$y(10i) = y(10i) + 50, \quad i = 1,2,\cdots,18 \qquad (12)$$

显然，这 18 个样本点可视为测量异常数据，其中有 9 个包含在训练样本中，9 个包含在预测样本中。利用带有异常数据的训练样本重新建立SVR预测模型，通过择优选择模型参数C=1 000，ε=0.001，核函数取径向基函数，其宽度系数σ=1，然后利用预测样本进行预测。同时采用RBF神经网络回归方法进行回归预测，选定RBF神经网络的参数σ=5，e=0.01。图 4 显示了 2 种模型的预测结果。

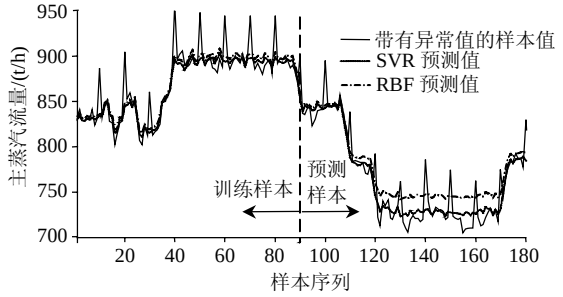

图 4　样本含有异常数据时 SVR 和 RBF 回归结果比较

Fig. 4　Comparison of regression between SVR and RBF with the train samples included abnormal data

由图 4 可以看出，当用含有异常值的数据建立回归模型时，SVR 和 RBF 模型的预测结果有很大差别，RBF 预测模型中，回归预测值受异常数据的影响，使回归曲线偏向异常数据，而 SVR 模型的预测值基本不受异常数据的影响，还能保持很好的预测精度。

图 5 是 SVR 和 RBF 模型预测结果残差值比较曲线。可以看出，RBF 回归预测值的总体残差远大于 SVR 预测值的残差，但异常点的预测残差反而比 SVR 预测值的残差小，所以此时利用 RBF 回归残差法检测异常数据已不能应用。但 SVR 预测模型中异常数据样本点的残差明显大于其他点的残差，所以很容易检测出异常数据，而且其他样本点的残差不受异常点的影响，保持很小，这也进一步说明了 SVR 模型受实际数据的随机波动影响较小，具有较好的鲁棒性。

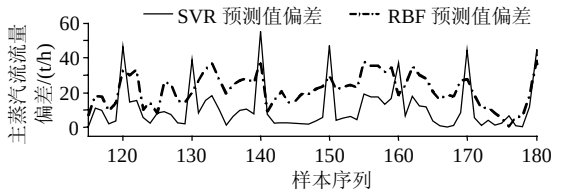

图 5　样本含有异常数据时 SVR 和 RBF 回归残差比较

Fig. 5　Comparison of regression residual error between SVR and RBF with the train samples included abnormal data

3　结论

测量数据的准确性对电站性能参数的实时计算是至关重要的，离开了数据的准确性，任何研究与监测算法都是不切实际的。基于支持向量回归的异常数据检测方法，在输入输出数据存在异常数据干扰的情况下，也能对测量数据中异常值进行正确检测与定位，同时能够对含有噪声干扰的数据进行正确回归估计，这对提高电站在线性能计算的有效性，具有重要的实际意义。

参考文献

[1] 杨勇平．火电机组性能在线监测系统测量误差定量分析[J]．电站系统工程，1999，15(5)：57-63．
Yang Yongping．Quantitative analysis on measure error of on-line performance monitoring system for thermal power unit[J]．Power System Engineering，1999，15(5)：57-63(in Chinese)．

[2] 张兴民，毛玉华，朱剑峰，等．利用图论方法进行多不良数据检测与辨识[J]．中国电机工程学报，1996，16(1)：69-72．
Zhang Xingmin，Mao Yuhua，Zhu Jianfeng，et al．Detection and identification of multi-bad data using graph theory[J]．Proceedings of the CSEE，1996，16(1)：69-72(in Chinese)．

[3] 李碧君，薛禹胜，顾锦汶，等．基于残差符号信息识别残差污染[J]．电力系统自动化，2000，24(5)：5-8.
Li Bijun，Xue Yusheng，Gu Jinwen，et al．Identification of residuals contamination based on signs of measurement residuals [J]．Automation of Electric Power System，2000，24(5)：5-8(in Chinese)．

[4] 司风琪，洪军，徐治皋．基于向量投影的数据检验 PCA 方法[J]．中国电机工程学报，2002，22(10)：157-160．
Si Fengqi，Hong Jun，Xu Zhigao．Data validation method based on PCA with vector projection[J]．Proceedings of the CSEE，2002，22(10)：157-160(in Chinese)．

[5] Knorr E M，Ng R T．Algorithms for mining distance-based outliers in large datasets[C]．Proceedings of the 24th Int. Conference on Very Large Databases，New York，1998：392-403．

[6] 董永贵，孙照焱，贾惠波．时间序列中异常值检测的负向选择算法[J]．机械工程学报，2004，40(10)：30-34．
Dong Yonggui，Sun Zhaoyan，Jia Huibo．Negative selected algorithm for anomaly detection in time series data[J]．Chinese Journal of Mechanical Engineering，2004，40(10)：30-34 (in Chinese)．

[7] 潘吉铮，周传光，钱宇．基于神经网络的化工过程测量数据在线校正技术的研究[J]．高校化学工程学报，2003，17(3)：319-324．
Pan Jizheng，Zhou Chuanguang，Qian Yu．On-line data rectification for chemical process measurements based on artificial neural networks[J]．Journal of Chemical Engineering of Chinese Universities，2003，17(3)：319-324(in Chinese)．

[8] 司风琪，徐治皋．基于自联想神经网络的测量数据自校正检验方法[J]．中国电机工程学报，2002，22(6)：152-155．
Si Fengqi，Xu Zhigao．Self-verifying data validation method based on the autoassociative neural network(AANN)[J]．Proceedings of the CSEE，2002，22(6)：152-155(in Chinese)．

[9] Vapnik V．The nature of statistical learning theory[M]．New York：Springer，1995：138-145．

[10] 王雷，徐治皋，司风琪．基于支持向量回归的凝汽器清洁系数时间序列预测[J]．中国电机工程学报，2007，27(14)：62-66．
Wang Lei，Xu Zhigao，Si Fengqi．Time series prediction of condenser cleanness coefficient based on support vector regression [J]．Proceedings of the CSEE，2007，27(14)：62-66(in Chinese)．

[11] 杨金芳，翟永杰，王东风，等．基于支持向量回归的时间序列预测[J]．中国电机工程学报，2005，25(17)：110-114．
Yang Jinfang，Zhai Yongjie，Wang Dongfeng，et al．Time series prediction based on support vector regression[J]．Proceedings of the CSEE，2005，25(17)：110-114(in Chinese)．

[12] 王春林，周昊，周樟华，等．基于支持向量机的大型电厂锅炉飞灰含碳量建模[J]．中国电机工程学报，2005，25(20)：72-76．
Wang Chunlin，Zhou Hao，Zhou Zhanghua，et al．Support vector machine modeling on the unburned carbon in fly ash[J]．Proceedings of the CSEE，2005，25(20)：72-76(in Chinese)．

[13] 黄景涛，马龙华，茅建波，等．基于支持向量回归的 300 MW电站锅炉再热气温建模[J]．中国电机工程学报，2006，26(7)：19-24．
Huang Jingtao，Ma Longhua，Mao Jianbo，et al．Modeling research of the reheat steam temperature of 300 MW boiler based on support vector regression[J]．Proceedings of the CSEE，2006，26(7)：19-24(in Chinese)．

[14] 孙斌，周云龙．基于支持向量机和小波包能量特征的气液两相流流型识别方法[J]．中国电机工程学报，2005，25(17)：93-99．
Sun Bin，Zhou Yunlong．Identification method of gas-liquid two-phase flow regime based on support vector machine and wavelet packet energe feature[J]．Proceedings of the CSEE，2005，25(17)：93-99(in Chinese)．

[15] 王雷，张欣刚，王洪跃，等．基于支持向量回归算法的汽轮机热耗率建模[J]．动力工程，2007，27(1)：19-23．
Wang Lei，Zhang Xingang，Wang Hongyue，et al．Model for the turbine heat rate based on the support vector regression[J]．Journal of Power Engineering，2007，27(1)：19-23(in Chinese)．

[16] 冯志刚，信太克规，王祈．基于最小二乘支持向量机预测器的传感器故障检测与数据恢复[J]．仪器仪表学报，2007，28(2)：193-197．
Feng Zhigang，Shida Katsunori，Wang Qi．Sensor fault detection and data recovery based on LS-SVM predictor[J]．Chinese Journal of Scientific Instrument，2007，28(2)：193-197(in Chinese)．

[17] 李元诚，方廷健，于尔铿．短期负荷预测的支持向量机方法研究[J]．中国电机工程学报，2003，23(6)：55-59
Li Yuancheng，Fang Tingjian，Yu Erkeng．Study of support vector machines for short-term load forecasting[J]．Proceedings of the CSEE，2003，23(6)：55-59(in Chinese)．

[18] 吴今培，孙德山．现代数据分析[M]．北京：机械工业出版社，2006：277-279．

[19] 周建新，王雷，吴海姬，等．基于支持向量回归的大容量机组主蒸汽流量建模[J]．动力工程，2008，23(2)：122-126．
Zhou Jianxin，Wang Lei，Wu Haiji，et al．A modeling of the main steam flow rate of a high-power steam turbine based on a support vector regression[J]．Journal of Engineering for Thermal Energy and Power，2008，23(2)：122-126(in Chinese)．

[20] 吴海姬，王雷，司风琪，等．基于 BP 神经网络的主蒸汽流量计算模型[J]．汽轮机技术，2007，49(4)：269-271，304．
Wu Haiji，Wang Lei，Si Fengqi，et al．The calculating model of main steam flowrate based on BP natural network[J]．Turbine Technology，2007，49(4)：269-271，304(in Chinese)．

王雷

收稿日期：2008-08-15。

作者简介：
王雷(1978—)，男，博士，讲师，主要研究方向为电厂节能、性能监测和故障诊断，zrqwl2003@126.com。

(责任编辑　车德竞)