文章编号: 1000−8152(2011)03−0343−08

基于鲁棒相关向量机的转炉炼钢终点预报模型

韩 敏[1], 赵 耀[1], 杨溪林[2], 林 东[3]

(1. 大连理工大学 电子信息与电气工程学部, 辽宁 大连 116023; 2. 北京金自天正智能控制股份有限公司, 北京 100070;
3. 本钢板材股份有限公司炼钢厂, 辽宁 本溪 117021)

摘要: 针对传统相关向量机在训练过程中易受异常点影响的问题, 提出了一种鲁棒相关向量机模型, 并将其应用于转炉炼钢终点碳含量和温度的预报. 通过为每一个训练样本设定独立的噪声方差系数, 并使其在训练过程中随模型预测误差的增大而逐渐减小来降低异常点的影响, 同时依据贝叶斯证据框架给出了模型超参数的迭代计算公式, 进行参数的优化. 使用标准测试数据和转炉炼钢实际生产数据进行仿真, 结果表明本文模型具有较好的预报精度和鲁棒性.

关键词: 转炉炼钢; 终点预报; 相关向量机; 噪声方差系数

中图分类号: TF721 **文献标识码**: A

Endpoint prediction model of basic oxygen furnace steelmaking based on robust relevance-vector-machines

HAN Min[1], ZHAO Yao[1], YANG Xi-lin[2], LIN Dong[3]

(1. Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian Liaoning 116023, China;
2. Beijing AriTime Intelligent Control Co Ltd, Beijing 100070, China;
3. Steelmaking Plant of Benxi Steel Sheet Co Ltd, Benxi Liaoning 117021, China)

Abstract: To deal with the problem that the classical relevance vector machine is sensitive to outliers, we present a novel robust relevance vector machine. This machine is applied to predict the endpoint carbon content and temperature of the basic-oxygen-furnace(BOF) steelmaking. Each training sample is assumed to have its individual coefficient of noise variance. With the increase of the prediction error during training procedure, the coefficients of outliers gradually decrease, reducing the impact of outliers. In addition, the iterative formulas for the optimization of hyper-parameters are derived in the Bayesian evidence framework. Simulation results of benchmark test data and the BOF steelmaking data show that the proposed mode achieves high prediction accuracy and good robustness.

Key words: BOF steelmaking; endpoint prediction; relevance-vector-machines; coefficient of noise variance

1 引言(Introduction)

转炉炼钢是目前世界上最常用的炼钢方法之一, 通过向炉内铁水顶吹氧气达到降碳、升温和除杂的目的, 从而获得终点碳含量和温度都满足工艺要求的钢水. 在转炉炼钢生产过程中, 钢水终点碳含量和温度预报是最重要的环节之一. 因为预报结果直接关系到对吹氧量的调节以及炼钢终点的控制, 所以建立准确的终点预报模型对保证钢水质量、减少生产返工、缩短冶炼时间、降低冶炼成本等具有重要意义[1].

传统的终点碳含量和温度预报是根据吹氧降碳过程中的质量平衡和热量平衡建立机理模型, 但由于转炉炼钢的反应过程十分复杂, 而且易受炼钢原材料成分不稳定等因素的影响, 因此机理模型的预报精度通常较低[2]. 近几年来, 有学者采用神经网络(neural network, NN)[3,4]、支持向量机(support vector machine, SVM)[5,6]等智能方法建立炼钢过程的终点预报模型并且取得了不错的效果, 但是这类模型通常需要一定数量的数据样本进行训练. 随着样本规模的增加, 往往会造成模型参数(如NN的隐层节点个数、SVM的支持向量个数)逐渐增多、模型结构更加复杂等问题出现[7], 因此不适合在实际生产现场使用.

相关向量机(relevance vector machine, RVM)是由M. E. Tipping于2000年提出的一种非线性概率模型[7]. 它利用贝叶斯方法进行推理, 模型结构简单, 具有优于SVM的稀疏性, 而且使用的核函数不需满足Mercer条件, 选择范围更加广泛[8]. 目前, 该模型

收稿日期: 2009−08−20; 收修改稿日期: 2010−04−26.
基金项目: 国家高技术研究发展计划“863”计划资助项目(2007AA04Z158); 国家自然科学基金资助项目(60674073).

已在化工[9]、制造[10]等工业过程领域的建模方面有所应用. 虽然RVM相比SVM具有模型结构更加稀疏等优势, 但该方法的不足是假设所有样本数据的噪声均满足高斯分布, 当训练样本中包含异常点时会给模型造成影响, 降低预测精度[11]. 在工业生产中, 由于环境复杂、测量误差等因素的影响, 现场得到的数据不可避免会包含一些异常点. 这样就导致训练样本噪声不再满足高斯分布, 影响了相关向量机的训练过程, 降低模型鲁棒性并限制其进一步的应用.

针对该问题, 本文提出了一种鲁棒相关向量机(robust relevance vector machine, RRVM)并利用其建立转炉炼钢终点碳含量和温度预报模型. 通过给每个训练样本点设定独立的噪声方差系数来降低异常点的影响, 同时依据贝叶斯证据框架对模型超参数进行优化. 仿真结果表明, 该方法在训练样本包含异常点时仍能保持较高的预报精度, 具有较好的鲁棒性.

2 传统相关向量机的回归原理(Regression principle of classical RVM)

设给定输入–输出样本对$\{\boldsymbol{x}_i, t_i\}_{i=1}^N$作为训练数据集, 其中: $\boldsymbol{x}_i \in \mathbb{R}^n, t_i \in \mathbb{R}$, 同时假设输出变量$t_i$全部由带有噪声$\varepsilon_i$的回归模型产生, 如式(1)所示:

$$t_i = y(\boldsymbol{x}_i; \boldsymbol{w}) + \varepsilon_i, \tag{1}$$

其中: 样本噪声ε_i满足均值为0, 方差为σ^2的高斯分布, 即$\varepsilon_i \sim N(\varepsilon_i|0, \sigma^2)$. 类似于SVM, RVM也将回归函数$y(\boldsymbol{x}; \boldsymbol{w})$表示成一系列基函数$K(\boldsymbol{x}, \boldsymbol{x}_i)$的线性组合形式:

$$y(\boldsymbol{x}; \boldsymbol{w}) = \sum_{i=1}^{N} w_i K(\boldsymbol{x}, \boldsymbol{x}_i) + w_0. \tag{2}$$

定义模型的权值向量$\boldsymbol{w} = [w_0 \ \ w_1 \ \ \cdots \ \ w_N]^{\mathrm{T}}$. 由于假设噪声$\varepsilon_i$满足高斯分布, 结合式(1)可知输出变量$t_i$也满足均值为$y(\boldsymbol{x}_i; \boldsymbol{w})$, 方差为$\sigma^2$的高斯分布, 即$p(t_i|\boldsymbol{x}_i) = N\left(t_i|y(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2\right)$. 为方便起见, 引入超参数$\beta = 1/\sigma^2$, 则整个训练数据集的似然函数可以表示为

$$p(\boldsymbol{t}|\boldsymbol{w}, \beta) = (\frac{\beta}{2\pi})^{N/2} \exp\{-\frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{w}\|^2\}, \tag{3}$$

其中: $\boldsymbol{t} = [t_1 \ \ t_2 \ \ \cdots \ \ t_N]^{\mathrm{T}}$, $\boldsymbol{\Phi} \in \mathbb{R}^{N\times(N+1)}$是设计矩阵, 定义式为$\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_1) \ \ \phi(\boldsymbol{x}_2) \ \ \cdots \ \ \phi(\boldsymbol{x}_N)]^{\mathrm{T}}$, 基函数向量$\phi(\boldsymbol{x}_i) = [1 \ \ K(\boldsymbol{x}_i \ \ \boldsymbol{x}_1) \ \ \cdots \ \ K(\boldsymbol{x}_i, \boldsymbol{x}_N)]^{\mathrm{T}}$, $i = 1, \cdots, N$.

RVM训练过程的最终目的是求出权值向量$\boldsymbol{w}$的后验分布, 为了保证模型稀疏性, 需要定义权值w_j $(j = 0, \cdots, N)$的先验分布. 假设其满足均值为0, 方差为α_j^{-1}的高斯分布, 因此$\boldsymbol{w}$的先验可以表示为

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{j=0}^{N} N(w_j|0, \alpha_j^{-1}), \tag{4}$$

其中: $\boldsymbol{\alpha} = [\alpha_0 \ \ \alpha_1 \ \ \cdots \ \ \alpha_N]^{\mathrm{T}}$, 每一个独立的超参数$\alpha_j$都只与其对应的权值$w_j$相关. 根据贝叶斯公式, 利用样本似然函数式(3)和$\boldsymbol{w}$先验分布式(4)可以得到$\boldsymbol{w}$的后验分布计算公式:

$$p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{\alpha}, \beta) = \frac{p(\boldsymbol{w}|\boldsymbol{\alpha})p(\boldsymbol{t}|\boldsymbol{w}, \beta)}{p(\boldsymbol{t}|\boldsymbol{\alpha}, \beta)}. \tag{5}$$

由于$p(\boldsymbol{w}|\boldsymbol{\alpha})$和$p(\boldsymbol{t}|\boldsymbol{w}, \beta)$均为高斯分布, 因此其乘积也满足高斯分布, 而$p(\boldsymbol{t}|\boldsymbol{\alpha}, \beta)$中不包含$\boldsymbol{w}$, 可以看作是归一化系数, 所以$\boldsymbol{w}$的后验能进一步表示成

$$p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{\alpha}, \beta) = N(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{6}$$

协方差矩阵和均值分别如公式(7)(8)所示:

$$\boldsymbol{\Sigma} = \left(\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \boldsymbol{A}\right)^{-1}, \tag{7}$$

$$\boldsymbol{\mu} = \beta\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{t}, \tag{8}$$

其中$\boldsymbol{A} = \mathrm{diag}\{\alpha_0, \alpha_1, \cdots, \alpha_N\}$. 超参数$\beta$和$\alpha_j$直接影响$\boldsymbol{w}$的后验分布, 需要对其优化从而得到$\boldsymbol{w}$的最大后验分布. 优化过程依据贝叶斯证据框架, 即通过最大化边缘似然$p(\boldsymbol{t}|\boldsymbol{\alpha}, \beta)$实现超参数的优化. 首先将$p(\boldsymbol{t}|\boldsymbol{\alpha}, \beta)$取负对数得到目标函数, 然后令目标函数分别对超参数$\alpha_j (j = 0, 1, \cdots, N)$和$\beta$求偏导并令导数式为0, 就可以得到迭代计算公式, 如式(9)(10)所示:

$$\alpha_j = \frac{1}{\mu_j^2 + \Sigma_{jj}} = \frac{\gamma_j}{\mu_j^2}, \ j = 0, 1, \cdots, N, \tag{9}$$

$$\beta = \frac{(N - \sum_{j=0}^{N} \gamma_j)}{\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{\mu}\|^2}, \tag{10}$$

其中: μ_j是权值后验均值向量$\boldsymbol{\mu}$的第j个元素, Σ_{jj}是协方差矩阵$\boldsymbol{\Sigma}$的第j个对角元, $\gamma_j = 1 - \alpha_j\Sigma_{jj}$. 在RVM模型的训练过程中, 公式(7)∼(10)要依次进行迭代计算, 直到所有参数都收敛或达到最大训练次数时为止, 此时RVM的建模过程完成.

3 鲁棒相关向量机模型(Robust relevance vector machine model)

RVM假设所有训练样本的噪声ε_i都满足具有相同方差σ^2的0均值高斯分布, 但是如果训练数据集包含异常点, 就会破坏该假设. 异常点可以看作是正常样本点夹杂了严重的噪声所致, 因此其噪声分布的方差可能大于σ^2, 这样就导致式(3)不成立, 利用RVM建立的回归模型就会受到异常点的影响, 致使其鲁棒性下降.

在训练数据包含异常点的情况下, 用相同方差σ^2(或超参数β)来描述所有训练样本的噪声分布显然不合适. 为解决该问题, 本文提出的鲁棒相关向量机通过为每个样本点引入独立的噪声方差系数来提高模型对异常点的抑制能力. 同时, 依据贝叶斯证

据框架推导出模型超参数和噪声方差系数的迭代计算公式, 实现各参数的优化. 具体过程如下: 借鉴贝叶斯加权线性回归的思想[12], 设第i个样本点所包含噪声ε_i的分布为

$$p(\varepsilon_i)=N(\varepsilon_i|0,\frac{\sigma^2}{\beta_i}),\tag{11}$$

其中: σ^2是所有样本的平均方差, β_i是引入的噪声方差系数, 并且假设其先验分布满足γ分布, 即$p(\beta_i)=\text{Gamma}(a_i,b_i)$. 定义向量$\boldsymbol{\beta}=[\beta_1\ \ \beta_2\ \ \cdots\ \ \beta_N]^{\mathrm{T}}$, 则训练样本的似然函数变为

$$p(\boldsymbol{t}|\boldsymbol{w},\boldsymbol{\beta},\sigma^2)=(2\pi\sigma^2)^{-N/2}|\boldsymbol{B}|^{1/2}\times\exp\{-\frac{1}{2\sigma^2}(\boldsymbol{t}-\boldsymbol{\Phi w})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi w})\}.\tag{12}$$

其中: $\boldsymbol{B}=\text{diag}\{\beta_1,\beta_2,\cdots,\beta_N\}$, $|\boldsymbol{B}|$表示矩阵$\boldsymbol{B}$的行列式, $\boldsymbol{t},\boldsymbol{w}$和$\boldsymbol{\Phi}$的定义与前面相同, $\boldsymbol{w}$的先验分布不变, 仍为式(4), 因此由贝叶斯公式可得权值的后验分布为

$$p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{\alpha},\boldsymbol{\beta},\sigma^2)=\frac{p(\boldsymbol{w}|\boldsymbol{\alpha})p(\boldsymbol{t}|\boldsymbol{w},\boldsymbol{\beta},\sigma^2)}{p(\boldsymbol{t}|\boldsymbol{\alpha},\boldsymbol{\beta},\sigma^2)}=N(\boldsymbol{w}|\boldsymbol{\mu},\boldsymbol{\Sigma}),\tag{13}$$

其中$\boldsymbol{\Sigma}$和$\boldsymbol{\mu}$的计算公式分别为式(14)和(15):

$$\boldsymbol{\Sigma}=(\boldsymbol{A}+\sigma^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Phi})^{-1}=(\boldsymbol{A}+\sigma^{-2}\sum_{i=1}^{N}\beta_i\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}})^{-1},\tag{14}$$

$$\boldsymbol{\mu}=\sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{t}=\sigma^{-2}\boldsymbol{\Sigma}(\sum_{i=1}^{N}\beta_i\phi(\boldsymbol{x}_i)t_i).\tag{15}$$

由于均值和协方差矩阵的计算公式中都含有超参数$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$和σ^2, 依据最大似然的思想, 需要对其进行优化, 从而得到模型权值$\boldsymbol{w}$的最大后验估计. 优化方法仍然根据贝叶斯证据框架, 通过最大化边缘似然函数$p(\boldsymbol{t}|\boldsymbol{\alpha},\boldsymbol{\beta},\sigma^2)$和噪声方差系数$\boldsymbol{\beta}$的先验分布$p(\boldsymbol{\beta})$的乘积实现[13,14]. 与传统RVM不同, 鲁棒相关向量机模型中待优化的超参数除了$\boldsymbol{\alpha}$和σ^2外, 还增加了噪声方差系数向量$\boldsymbol{\beta}$. 前面已经定义了$p(\beta_i)$的先验分布, 因此$p(\boldsymbol{\beta})$可以表示为

$$p(\boldsymbol{\beta})=\prod_{i=1}^{N}p(\beta_i).\tag{16}$$

边缘似然$p(\boldsymbol{t}|\boldsymbol{\alpha},\boldsymbol{\beta},\sigma^2)$的计算公式如式(17)所示:

$$p(\boldsymbol{t}|\boldsymbol{\alpha},\boldsymbol{\beta},\sigma^2)=\int p(\boldsymbol{t}|\boldsymbol{w},\boldsymbol{\beta},\sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})\mathrm{d}\boldsymbol{w}=(2\pi)^{-N/2}|\boldsymbol{C}|^{-1/2}\exp\{-\frac{1}{2}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{t}\},\tag{17}$$

其中$\boldsymbol{C}=\sigma^2\boldsymbol{B}^{-1}+\boldsymbol{\Phi}\boldsymbol{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}$.

最大化边缘似然函数$p(\boldsymbol{t}|\boldsymbol{\alpha},\boldsymbol{\beta},\sigma^2)$和$\boldsymbol{\beta}$先验分布$p(\boldsymbol{\beta})$的乘积等价于最小化其负对数. 另外, 为了便于计算, 一般不直接关于$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$和σ^2进行优化, 而是关于其对数形式$\log\boldsymbol{\alpha}$, $\log\boldsymbol{\beta}$和$\log\sigma^2$进行优化[7]. 对$p(\boldsymbol{t}|\boldsymbol{\alpha},\boldsymbol{\beta},\sigma^2)$和$p(\boldsymbol{\beta})$的乘积取负对数并去掉与$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$和σ^2无关的项后, 得到待优化的目标函数:

$$L=-\frac{1}{2}[-\log|\boldsymbol{\Sigma}|-\log|\boldsymbol{A}|+N\log\sigma^2-\log|\boldsymbol{B}|+\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}+\sigma^{-2}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu})]+\sum_{i=1}^{N}(a_i\log\beta_i-b_i\beta_i).\tag{18}$$

令式(18)分别对$\log\alpha_j(j=0,\cdots,N)$, $\log\beta_i(i=1,\cdots,N)$和$\log\sigma^2$求偏导并令导数式等于0, 整理后得到超参数的迭代公式, 如式(19)~(21)所示(详细推导过程请参见附录):

$$\alpha_j=\frac{1}{\mu_j^2+\Sigma_{jj}}=\frac{\gamma_j}{\mu_j^2},\tag{19}$$

$$\beta_i=(2a_i+1)/[2b_i+\sigma^{-2}(t_i-\phi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\mu})^2+\sigma^{-2}\text{tr}(\boldsymbol{\Sigma}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}})],\tag{20}$$

$$\sigma^2=\frac{(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu})}{N-\sum_{j=0}^{N}\gamma_j}.\tag{21}$$

其中: $j=0,\cdots,N$, $i=1,\cdots,N$, μ_j是均值向量$\boldsymbol{\mu}$的第j个元素, Σ_{jj}是协方差矩阵$\boldsymbol{\Sigma}$的第j个对角元, $\gamma_j=1-\alpha_j\Sigma_{jj}$, $\text{tr}(\cdot)$表示矩阵的迹. 引入γ_j的目的是为了通过分析γ_j与α_j的关系来表征w_j适应训练样本数据的程度. α_j是w_j先验分布方差的倒数, 如果α_j较大, 由式(14)易知$\Sigma_{jj}\approx\alpha_j^{-1}$, 此时$\gamma_j\approx0$, 表明$w_j$受先验分布的影响较大, 对训练数据的适应程度较差. 反之当α_j较小时, 有$\gamma_j\approx1$, 表明w_j对训练数据的适应程度较好. 由此就得到了鲁棒相关向量机权值后验的协方差矩阵、均值以及超参数的迭代公式, 分别为式(14)(15)(19)~(21).

综上所述, 鲁棒RVM的训练过程如下:

Step 1　初始化模型的超参数$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, σ^2以及a_i, b_i, $i=1,\cdots,N$;

Step 2　利用式(14)(15)计算权值$\boldsymbol{w}$后验分布的协方差矩阵$\boldsymbol{\Sigma}$和均值$\boldsymbol{\mu}$;

Step 3　利用式(19)~(21)更新超参数$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$和σ^2的值, 部分α_j在迭代过程中会趋于正无穷(通常判断其是否大于某一阈值, 如10^9, 对应的μ_j逐渐趋近于0, 实现“剪枝”, 从而保证模型稀疏性[8];

Step 4　判断所有参数是否收敛, 若收敛则停止迭代, 否则返回Step 2. 此时后验均值$\boldsymbol{\mu}$中剩余非0元素对应的基函数即为“相关向量”.

在模型训练过程中, 式(14)(15)(19)~(21)几个公式需要交替使用, 直到所有参数都收敛或是达到最大迭代次数时为止. 从式(20)可以看出, 样本点$\{\boldsymbol{x}_i,$

$t_i\}$对应的预测误差平方$(t_i-\phi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\mu})^2$在分母上, 如果其值很大, 那么对应的噪声方差系数β_i就会很小. 当预测误差趋近于正无穷时, 相应的β_i值就会趋于0. 通过分析式(14)(15)可知, $\boldsymbol{w}$后验均值$\boldsymbol{\mu}$和协方差矩阵$\boldsymbol{\Sigma}$的计算公式中包含了所有样本点的加权求和项. 其中$\boldsymbol{\mu}$的计算公式中包含的加权求和项为$\sum\limits_{i=1}^{N}\beta_i\phi(\boldsymbol{x}_i)t_i$, $\boldsymbol{\Sigma}$的计算公式包含的求和项为$\sum\limits_{i=1}^{N}\beta_i\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}}$, 加权系数值恰好都为$\beta_i$. 因此, 在利用公式(14)(15)迭代计算$\boldsymbol{w}$的后验均值和协方差矩阵时, 如果一个样本点对应的噪声方差系数β_i较小, 那么该点对均值$\boldsymbol{\mu}$和协方差矩阵$\boldsymbol{\Sigma}$的贡献率就会下降. 这样当样本中包含异常点时, 由于训练过程中该点的预测误差较大, 因此对应的β_i就会逐渐减小, 相当于在建模过程中自动实现对异常点的检测和剔除, 降低了其对模型的影响. 模型训练完成后即可使用测试样本进行检验, 若输入向量为$\boldsymbol{x}^*$, 则输出为$y^*=\phi(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{\mu}$.

下面分析算法的时间复杂度. 模型训练时要利用式(14)求出$\boldsymbol{w}$后验分布的协方差矩阵$\boldsymbol{\Sigma}$, 这一过程中要进行矩阵求逆操作, 通过Cholesky分解实现, 时间复杂度为O(N^3), 其中N是训练样本个数. 当N过大时, 计算$\boldsymbol{\Sigma}$所需的时间会增加, 但是由于鲁棒RVM不需要使用交叉检验确定模型参数, 所以在模型参数选择方面只需要较少的计算时间即可, 从而补偿了计算$\boldsymbol{\Sigma}$所消耗的时间, 这样在样本规模较小时, 模型总的训练时间并不会增加很多.

4 仿真实验(Simulations)

为了验证鲁棒RVM模型的性能, 使用两组数据进行仿真实验: 一组是sinc函数逼近, 另一组是转炉炼钢的实际生产数据. 核函数选择常用的高斯型核函数$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\{-||\boldsymbol{x}_i-\boldsymbol{x}_j||^2/c^2\}$, 超参数初始值$\alpha_j=N/\mathrm{var}(\boldsymbol{t})$, $\sigma^2=\mathrm{var}(\boldsymbol{t})$, 其中var$(\boldsymbol{t})$为$\boldsymbol{t}$的方差. 在$\beta_i$的先验分布Gamma$(a_i,b_i)$中令$a_i=1$, $b_i=1$, 这样β_i先验分布的均值为$a_i/b_i=1$, 方差为$a_i/b_i^2=1$[12], 表示最初假设所有训练样本点的噪声都满足相同的高斯分布, 即认为样本中没有异常点存在. 在迭代过程中每个样本对应的β_i会发生变化, 异常点的β_i逐渐减小, 由于设定了$a_i=1$和$b_i=1$的先验信息, 由式(20)可知迭代过程中β_i的变化范围是$0<\beta_i<1.5$. 这组先验参数值在大多数情况下是适用的.

4.1 sinc函数逼近(Approximation of sinc function)

$y=\mathrm{sinc}\,x=\sin|x|/|x|$是检验算法回归性能的标准测试函数, x取值范围是$[-10,10]$. 利用该函数随机生成100个采样点并在输出目标值上加入均值为0, 方差为0.01的高斯噪声作为训练样本. 训练结束后需进行测试, 为了能更好地表现出函数曲线的真实性, 从$x\in[-10,10]$的区间内均匀产生1000个采样点带入训练好的模型中计算. 测试点的数量较多, 所以点之间的距离较小, 从而能更真实地体现出函数曲线形状的局部特性, 更好地反映模型的逼近效果. 分别采用RVM和本文提出的RRVM模型进行训练及检验, 高斯核宽度都选择为2.5(经过反复多次实验, 发现核宽度在2.5附近时逼近效果最好), 结果如图1所示. 在训练样本不包含异常点的情况下两种算法的逼近结果相差不多, 逼近值与实际值比较接近, 图中的小方框代表RRVM模型的相关向量.

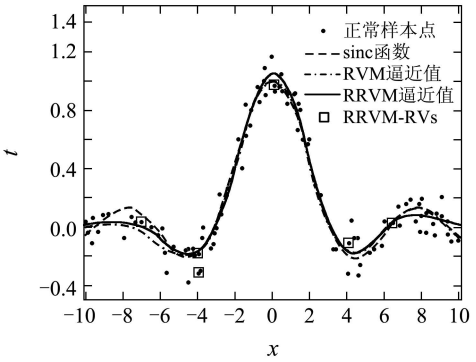

图 1 无异常点时的sinc逼近结果对比

Fig. 1 Comparison of approximation to sinc without outliers

为了检验模型的鲁棒性, 在原有100个训练样本中增加一定数量的异常点, 加入方式分为两种: 第1种是加入25个输出为恒值的异常点$\{x_{\mathrm{o}},t_{\mathrm{o}}\}$, 其中$t_{\mathrm{o}}=0.8$; 第2种是加入25个满足标准高斯分布的异常点. 使用这125个样本点训练模型, 并仍然采用相同的方法检验模型对sinc函数的逼近效果, 结果分别如图2, 3所示. 由于RVM假设所有训练样本的噪声都满足同一高斯分布, 所以在异常点的干扰下函数逼近效果较差. 本文提出的RRVM模型由于引入了噪声方差系数, 因此在训练过程中能很好地抑制异常点的影响, 具有较好的鲁棒性, 逼近效果好于传统RVM模型.

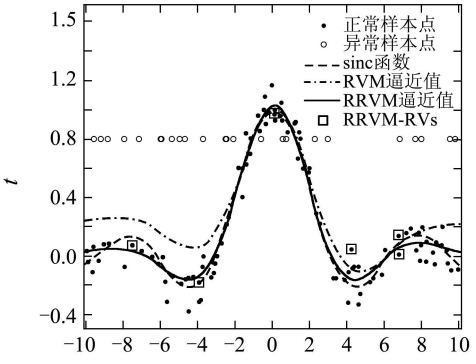

图 2 加入恒值异常点时的sinc逼近结果对比

Fig. 2 Comparison of approximation to sinc with outliers the outputs of which are constant

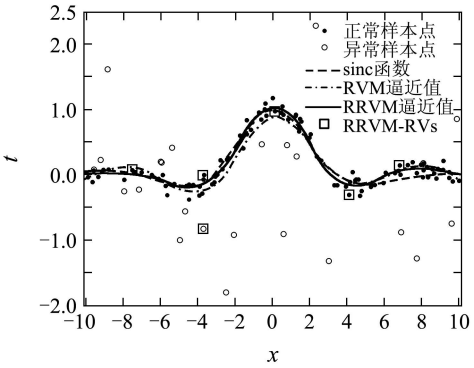

图 3 加入标准高斯分布异常点时的sinc逼近结果对比

Fig. 3 Comparison of approximation to sinc with outliers which are generated from standard Gaussian distribution

除传统RVM外, 本文还将RRVM与文献[15]提出的另一种提高RVM鲁棒性的改进模型TRVM进行了对比. TRVM通过剔除异常点、约简样本集的方法来增强模型鲁棒性. 在训练过程中首先根据模型预测误差来检测异常点并将其从训练样本中剔除, 进行样本集的约简, 然后利用去除异常点后的样本重新训练模型, 直到所有异常点都被检测到为止. 表1给出了训练样本有、无异常点时, 3种模型逼近值与函数实际值均方根误差(root mean square error, RMSE)的对比. 可以看出, RRVM的逼近精度要高于其他两种模型, 尤其是在训练样本包含异常点的情况下, 仍然能保持较高的精度.

表 1 不同模型逼近结果的RMSE对比

Table 1 RMSE comparison of different models

模型	RMSE(不同训练样本情况下)		
	无异常点	恒值异常点	高斯异常点
RVM	0.0517	0.1454	0.1145
TRVM[15]	0.0520	0.0565	0.0575
RRVM	0.0428	0.0424	0.0517

4.2 转炉炼钢终点预报(Endpoint prediction of BOF steelmaking)

本文使用转炉炼钢的实际生产数据建立终点碳含量和温度预报模型. 炼钢过程中, 当吹氧进行到80%左右时, 炉前操作人员会降下副枪第一次测量钢水温度和碳含量, 然后根据测量信息调整吹氧量, 因此用副枪测得的过程碳含量和温度对终点预报有影响. 另外, 冶炼过程中熔池内的降碳和升温主要靠吹氧实现, 所以从第一次副枪测量到冶炼结束这段时间内(工艺上称作二吹阶段)的吹氧量也是影响终点的重要因素. 通过以上分析, 选择终点预报模型的输入变量为: 副枪测量的过程碳含量$[C]_{\rm S}$、过程温度$T_{\rm S}$和二吹阶段氧气量$V_{\rm O_2}$, 输出变量为钢水从第一次副枪测量到吹炼结束时的降碳量$\Delta[C]$和升温值ΔT. 利用本文提出的RRVM分别建立钢水终点降碳量和升温值预报模型:

$$\Delta[C] = f_{[C]}([C]_{\rm S}, T_{\rm S}, V_{\rm O_2}), \tag{22}$$

$$\Delta T = f_{\rm T}([C]_{\rm S}, T_{\rm S}, V_{\rm O_2}). \tag{23}$$

根据降碳量$\Delta[C]$和升温值ΔT的预报结果就可以求出钢水终点的碳含量和温度. 因为转炉炼钢二吹阶段的主要目的是降低钢水碳含量并升高温度, 因此终点碳含量应该由过程碳含量$[C]_{\rm S}$减去降碳量$\Delta[C]$得到, 而终点温度则应由过程温度$T_{\rm S}$加上升温值ΔT后得到. 相应的钢水终点碳含量、温度计算公式分别如式(24)(25)所示:

$$[C]_{\rm End} = [C]_{\rm S} - \Delta[C], \tag{24}$$

$$T_{\rm End} = T_{\rm S} + \Delta T. \tag{25}$$

仿真实验数据来源于某钢厂一座 150 t 转炉, 共计300组. 选择数据集中的前200组用于模型训练, 后100组用于测试. 为了进行对比, 分别使用SVM、RVM、前面介绍过的文献[15]中的 TRVM 和本文提出的RRVM进行仿真. SVM 程序选择目前常用的LIBSVM软件包[16], 4种模型的核函数都使用高斯核. 经多次实验后降碳量预报模型的核宽度选择为2.8, 升温值预报模型的核宽度选择为0.5, SVM模型的正则项系数C和ε不敏感损失函数的参数均通过交叉检验确定. 仿真结果比较的指标除了均方根误差RMSE外, 还增加了在转炉炼钢领域经常使用的炼钢终点碳含量和温度的命中率, 定义如下:

$$\text{命中率} = \frac{\text{终点命中的炉次数量}}{\text{炉次的总数量}} \times 100\%, \tag{26}$$

其中: 终点命中是指钢水终点碳含量或温度的预报值$y_{\rm predict}$与实际值$y_{\rm real}$之间的误差绝对值在允许范围e之内, 即满足$|y_{\rm predict} - y_{\rm real}| < e$, 通常转炉炼钢终点碳含量的允许误差绝对值为0.02%, 终点温度的允许误差绝对值为12 ℃.

仿真结果对比如表 2 所示, 除温度命中率外, RRVM在其他指标上都好于其他3种方法. 此外还可以看出, 各模型预报终点温度的命中率普遍低于碳含量命中率. 这是由于在转炉炼钢冶炼后期, 钢水中的磷、硅等杂质元素均已脱除, 影响钢水降碳的因素较少, 降碳过程比较平稳, 因此4种模型预报结果的命中率都普遍较高. 而钢水温度在吹炼过程中所受的影响因素则较多, 如喷溅、炉壁传热、铁元素氧化以及炉气带走热量等, 这些都使钢水升温过程变得复杂, 从而导致终点温度预报的命中率普遍偏低, 但是在允许误差范围内, RRVM还是取得了优于其他方法的预报结果. 图4,5分别是使用RRVM模型时, 终点碳含量和温度预报值与实际值的比较.

表 2 不同模型的终点预报结果对比

Table 2 Endpoint prediction result comparison of different models

模型	终点碳含量		终点温度	
	命中率/%	RMSE/(0.01 %)	命中率/%	RMSE/℃
SVM	90	1.2230	70	12.37
RVM	92	1.1648	71	12.55
TRVM[15]	92	1.1648	71	12.70
RRVM	94	1.1027	71	12.02

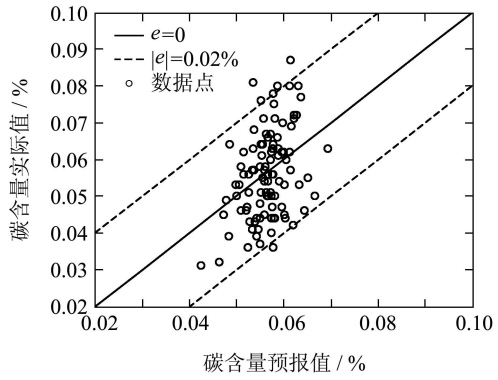

图 4 终点碳含量预报值与实际值的比较

Fig. 4 Comparison of predicted value and actual value of endpoint carbon content

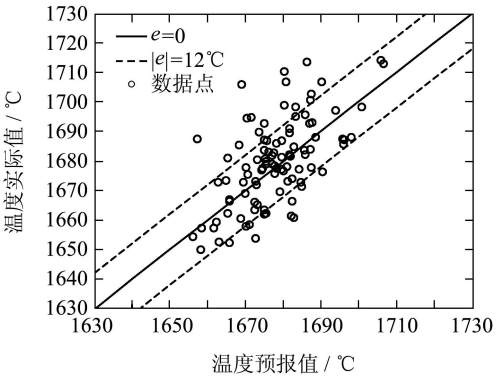

图 5 终点温度预报值与实际值的比较

Fig. 5 Comparison of predicted value and actual value of endpoint temperature

5 结论(Conclusions)

本文提出了一种鲁棒相关向量机模型RRVM, 通过为每个训练样本设定独立的噪声方差系数来提高模型对于异常点的抑制能力, 从而解决传统RVM易受异常点影响的问题. 使用标准测试数据sinc函数的仿真结果表明, 在训练样本包含异常点的情况下, RRVM仍能保持较高的逼近精度. 在此基础上, 利用RRVM分别建立了转炉炼钢终点碳含量和温度预报模型并使用实际生产数据进行仿真, 均取得了较好的命中率和预报精度, 说明该方法在冶炼现场条件复杂、测量数据噪声较大的转炉炼钢终点预报方面具有一定的应用前景.

参考文献(References):

[1] 王勇, 杨宁川, 王承宽. 我国转炉炼钢的现状和发展[J]. 特殊钢, 2005, 26(4): 1 – 5.
(WANG Yong, YANG Ningchuan, WANG Chengkuan. Present situation and development of converter steelmaking in China[J]. *Special Steel*, 2005, 26(4): 1 – 5.)

[2] 陶钧, 谢书明, 柴天佑. 转炉炼钢控制模型的研究与展望[J]. 钢铁, 1999, 34(8): 69 – 72.
(TAO Jun, XIE Shuming, CHAI Tianyou. Study and prospect of BOF steelmaking model[J]. *Iron and Steel*, 1999, 34(8): 69 – 72.)

[3] 谢书明, 陶钧, 柴天佑. 基于神经网络的转炉炼钢终点控制[J]. 控制理论与应用, 2003, 20(6): 903 – 907.
(XIE Shuming, TAO Jun, CHAI Tianyou. BOF steelmaking endpoint control based on neural network[J]. *Control Theory & Applications*, 2003, 20(6): 903 – 907.)

[4] FILETI A M F, PACIANOTTO T A, CUNHA A P. Neural modeling helps the BOS process to achieve aimed end-point conditions in liquid steel[J]. *Engineering Applications of Artificial Intelligence*, 2006, 19(1): 9 – 17.

[5] YUAN P, MAO Z Z, WANG F L. Endpoint prediction of EAF based on multiple support vector machines[J]. *Journal of Iron and Steel Research, International*, 2007, 14(2): 20 – 24, 29.

[6] 张绍德, 毛雪菲, 毛雪芹, 等. 基于grey Markov-支持向量机的电弧炉终点参数预报[J]. 控制理论与应用, 2009, 26(12): 1443 – 1448.
(ZHANG Shaode, MAO Xuefei, MAO xueqin, et al. End-point parameter prediction for electric are curnace based on grey Markov-support-vector-machines[J]. *Control Theory & Applications*, 2009, 26(12): 1443 – 1448.)

[7] TIPPING M E. The relevance vector machine[M] //SOLLA S A, LEEN T K, MULLER K R. *Advances in Neural Information Processing Systems 12*. Cambridge, United States: MIT Press, 2000: 652 – 658.

[8] TIPPING M E. Sparse Bayesian learning and the relevance vector machine[J]. *Journal of Machine Learning Research*, 2001, 1(3): 211 – 244.

[9] SUN Z H, SUN Y X. Soft sensor based on relevance vector machines for microbiological fermentation[J]. *Developments in Chemical Engineering and Mineral Processing*, 2005, 13(3/4): 243 – 248.

[10] YUAN J, WANG K S, YU T, et al. Integrating relevance vector machines and genetic algorithms for optimization of seed-separating process[J]. *Engineering Applications of Artificial Intelligence*, 2007, 20(7): 970 – 979.

[11] FAUL A C, TIPPING M E. A variational approach to robust regression[C] //*Proceedings of the 2001 International Conference on Artificial Neural Networks*. Vienna, Austria: Springer, 2001: 95 – 102.

[12] TING J A, D'SOUZA A, SCHAAL S. Automatic outlier detection: a Bayesian approach[C] //*Proceedings of the 2007 IEEE International Conference on Robotics and Automation*. New York: IEEE, 2007: 2489 – 2494.

[13] MACKAY D J C. Bayesian interpolation[J]. *Neural Computation*, 1992, 4(3): 415 – 447.

[14] MACKAY D J C. The evidence framework applied to classification networks[J]. *Neural Computation*, 1992, 4(5): 720 – 736.

[15] YANG B, ZHANG Z K, SUN Z S. Robust relevance vector regression with trimmed likelihood function[J]. *IEEE Signal Processing Letters*, 2007, 14(10): 746 – 749.

[16] CHANG C C, LIN C J. LIBSVM: a library for support vector machines[EB/OL]. 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[17] ROWEIS S. Matrix identities[EB/OL]. 1999. Document available at http://www.cs.nyu.edu/∼roweis/notes/matrixid.pdf.

附录　超参数迭代公式(19)∼(21)的推导过程 (Appendix　Derivation process of iteration formulas (19)∼(21) on hyperparameters)

超参数$\boldsymbol{\alpha},\boldsymbol{\beta}$和$\sigma^2$优化过程的目标函数是文中的公式(18), 即

$$L=-\frac{1}{2}[-\log|\boldsymbol{\Sigma}|-\log|\boldsymbol{A}|+N\log\sigma^2-\log|\boldsymbol{B}|+\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}+\sigma^{-2}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu})]+\sum_{i=1}^{N}(a_i\log\beta_i-b_i\beta_i).$$

1) $\alpha_j(j=0,1,\cdots,N)$计算公式(19)的推导过程.

首先令目标函数L分别对$\log\alpha_j(j=0,1,\cdots,N)$求偏导, 去掉与$\log\alpha_j$无关的项后可以得到

$$\frac{\partial L}{\partial\log\alpha_j}=-\frac{1}{2}[\frac{\partial}{\partial\log\alpha_j}(-\log|\boldsymbol{A}|+\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}-\log|\boldsymbol{\Sigma}|)]=-\frac{1}{2}[-\frac{\partial\log|\boldsymbol{A}|}{\partial\log\alpha_j}+\frac{\partial}{\partial\alpha_j}(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}-\log|\boldsymbol{\Sigma}|)\frac{\partial\alpha_j}{\partial\log\alpha_j}].\tag{A1}$$

因为$\boldsymbol{A}=\mathrm{diag}\{\alpha_0,\cdots,\alpha_N\}$且$\boldsymbol{\mu}=[\mu_0\ \ \cdots\ \ \mu_N]^{\mathrm{T}}$, 所以

$$\frac{\partial\log|\boldsymbol{A}|}{\partial\log\alpha_j}=\frac{\partial}{\partial\log\alpha_j}(\log\prod_{j=0}^{N}\alpha_j)=\frac{\partial}{\partial\log\alpha_j}(\sum_{j=0}^{N}\log\alpha_j)=1,\tag{A2}$$

$$\frac{\partial\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}}{\partial\log\alpha_j}=\frac{\partial}{\partial\alpha_j}(\boldsymbol{\mu}^{\mathrm{T}}\mathrm{diag}\{\alpha_0,\cdots,\alpha_N\}\boldsymbol{\mu})=\mu_j^2,\tag{A3}$$

其中μ_j是$\boldsymbol{\mu}$的第j个元素. 根据矩阵行列式对数求偏导的计算公式[17]:

$$\frac{\partial\log|\boldsymbol{X}(z)|}{\partial z}=\mathrm{tr}(\boldsymbol{X}^{-1}\frac{\partial\boldsymbol{X}}{\partial z}),$$

可以得到下式:

$$\frac{-\partial\log|\boldsymbol{\Sigma}|}{\partial\alpha_j}=\frac{\partial\log|\boldsymbol{\Sigma}^{-1}|}{\partial\alpha_j}=\mathrm{tr}(\boldsymbol{\Sigma}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\alpha_j}),$$

其中tr(·)表示矩阵的迹. 根据文中的公式(14)可知$\boldsymbol{\Sigma}^{-1}=(\boldsymbol{A}+\sigma^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Phi})$, 所以$\dfrac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\alpha_j}$是一个绝大多数元素为0的稀疏矩阵, 只有第$j$个对角元素为1, 这样就可以得到

$$\frac{-\partial\log|\boldsymbol{\Sigma}|}{\partial\alpha_j}=\mathrm{tr}(\boldsymbol{\Sigma}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\alpha_j})=\Sigma_{jj},\tag{A4}$$

其中Σ_{jj}是$\boldsymbol{\Sigma}$的第j个对角元. 又因为$\dfrac{\partial\alpha_j}{\partial\log\alpha_j}$的值为$\alpha_j$, 将此结果与式(A2)∼(A4)代入公式(A1)中就可以得到

$$\frac{\partial L}{\partial\log\alpha_j}=-\frac{1}{2}[-1+\alpha_j(\mu_j^2+\Sigma_{jj})].$$

令上式等于0, 整理后即求出α_j的迭代公式:

$$\alpha_j=\frac{1}{\mu_j^2+\Sigma_{jj}},\ j=0,1,\cdots,N.$$

如前文所述, 引入$\gamma_j=1-\alpha_j\Sigma_{jj}$用以表征$w_j$对训练样本数据的适应程度. γ_j的值越大, 表明w_j越适应训练数据, γ_j的值越小, 则说明w_j受先验分布的影响较大, 对训练数据的适应程度较差. 引入γ_j后就得到文中的式(19):

$$\alpha_j=\frac{1}{\mu_j^2+\Sigma_{jj}}=\frac{\gamma_j}{\mu_j^2},\ j=0,1,\cdots,N.$$

2) $\beta_i(i=1,2,\cdots,N)$计算公式(20)的推导过程.

令目标函数L分别对$\log\beta_i(i=1,2,\cdots,N)$求偏导, 去掉与$\log\beta_i$无关的项后, 得到

$$\frac{\partial L}{\partial\log\beta_i}=-\frac{1}{2}\cdot\frac{\partial}{\partial\log\beta_i}[-\log|\boldsymbol{B}|+\log|\boldsymbol{\Sigma}^{-1}|+\sigma^{-2}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu})]+\frac{\partial}{\partial\log\beta_i}\sum_{i=1}^{N}(a_i\log\beta_i-b_i\beta_i).\tag{A5}$$

又因为$\boldsymbol{B}=\mathrm{diag}\{\beta_1,\beta_2,\cdots,\beta_N\}$, 所以上式第1项为

$$\frac{\partial\log|\boldsymbol{B}|}{\partial\log\beta_i}=\frac{\partial}{\partial\log\beta_i}(\log\prod_{i=1}^{N}\beta_i)=\frac{\partial}{\partial\log\beta_i}(\sum_{i=1}^{N}\log\beta_i)=1.\tag{A6}$$

第2项为

$$\frac{\partial\log|\boldsymbol{\Sigma}^{-1}|}{\partial\log\beta_i}=\frac{\partial\log|\boldsymbol{\Sigma}^{-1}|}{\partial\beta_i}\frac{\partial\beta_i}{\partial\log\beta_i}=\mathrm{tr}(\boldsymbol{\Sigma}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\beta_i})\beta_i=\mathrm{tr}(\boldsymbol{\Sigma}\frac{\partial}{\partial\beta_i}(\boldsymbol{A}+\sigma^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Phi}))\beta_i=\sigma^{-2}\mathrm{tr}(\boldsymbol{\Sigma}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}})\beta_i.\tag{A7}$$

第3项为

$$\frac{\partial}{\partial\log\beta_i}(\sigma^{-2}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu}))=\frac{\partial}{\partial\beta_i}(\sigma^{-2}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu}))\frac{\partial\beta_i}{\partial\log\beta_i}=\sigma^{-2}(t_i-\phi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\mu})^2\beta_i.\tag{A8}$$

第4项为

$$\frac{\partial}{\partial\log\beta_i}\sum_{i=1}^{N}(a_i\log\beta_i-b_i\beta_i)=a_i-b_i\beta_i.\tag{A9}$$

将以上式(A6)∼(A9)的计算结果代入$\dfrac{\partial L}{\partial\log\beta_i}$的计算公式(A5)中可得

$$\frac{\partial L}{\partial\log\beta_i}=-\frac{1}{2}[-1+\sigma^{-2}\mathrm{tr}(\boldsymbol{\Sigma}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}})\beta_i+\sigma^{-2}(t_i-\phi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\mu})^2\beta_i]+a_i-b_i\beta_i.$$

令上式等于0, 整理后即可求出$\beta_i(i=1,2,\cdots,N)$的迭代计算公式, 即文中的式(20):

$$\beta_i=\frac{2a_i+1}{2b_i+\sigma^{-2}(t_i-\phi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\mu})^2+\sigma^{-2}\mathrm{tr}(\boldsymbol{\Sigma}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}})}.$$

3) σ^2计算公式(21)的推导过程.

令目标函数L对$\log\sigma^2$求偏导, 并去掉与$\log\sigma^2$无关的项后可以得到

$$\frac{\partial L}{\partial\log\sigma^2}=$$
$$-\frac{1}{2}[N+\frac{\partial(\log|\boldsymbol{\Sigma}^{-1}|+\sigma^{-2}(\boldsymbol{t}-\boldsymbol{\Phi\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi\mu}))}{\partial\log\sigma^2}], \tag{A10}$$

其中

$$\frac{\partial}{\partial\log\sigma^2}(\log|\boldsymbol{\Sigma}^{-1}|+\sigma^{-2}(\boldsymbol{t}-\boldsymbol{\Phi\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi\mu}))=$$
$$\frac{\partial}{\partial\sigma^2}(\log|\boldsymbol{\Sigma}^{-1}|+\sigma^{-2}(\boldsymbol{t}-\boldsymbol{\Phi\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi\mu}))\frac{\partial\sigma^2}{\partial\log\sigma^2}=$$
$$(\frac{\partial}{\partial\sigma^2}\log|\boldsymbol{\Sigma}^{-1}|-\sigma^{-4}(\boldsymbol{t}-\boldsymbol{\Phi\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi\mu}))\sigma^2=$$
$$\sigma^2\frac{\partial}{\partial\sigma^2}\log|\boldsymbol{\Sigma}^{-1}|-\sigma^{-2}(\boldsymbol{t}-\boldsymbol{\Phi\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi\mu}). \tag{A11}$$

上式第1项又可以表示为

$$\sigma^2\frac{\partial}{\partial\sigma^2}\log|\boldsymbol{\Sigma}^{-1}|=\sigma^2\mathrm{tr}(\boldsymbol{\Sigma}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\sigma^2})=$$
$$\sigma^2\mathrm{tr}(\boldsymbol{\Sigma}\frac{\partial}{\partial\sigma^2}(\boldsymbol{A}+\sigma^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{B\Phi}))=$$
$$\sigma^2\mathrm{tr}(-\sigma^{-4}\boldsymbol{\Sigma\Phi}^{\mathrm{T}}\boldsymbol{B\Phi})=-\mathrm{tr}(\boldsymbol{\Sigma}(\sigma^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{B\Phi}))=$$
$$-\mathrm{tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1}-A))=-\mathrm{tr}(\boldsymbol{I}-\boldsymbol{\Sigma A}).$$

前面已经定义$\gamma_j=1-\alpha_j\Sigma_{jj}$, 所以$\sigma^2\dfrac{\partial}{\partial\sigma^2}\log|\boldsymbol{\Sigma}^{-1}|$可以进一步表示为

$$\sigma^2\frac{\partial}{\partial\sigma^2}\log|\boldsymbol{\Sigma}^{-1}|=-\mathrm{tr}(\boldsymbol{I}-\boldsymbol{\Sigma A})=$$
$$-(N-\mathrm{tr}(\boldsymbol{\Sigma A}))=-(N-\sum_{j=0}^{N}(\Sigma_{jj}\alpha_j))=$$
$$-\sum_{j=0}^{N}(1-\Sigma_{jj}\alpha_j)=-\sum_{j=0}^{N}\gamma_j. \tag{A12}$$

将式(A11)(12)代入公式(A10)中可得

$$\frac{\partial L}{\partial\log\sigma^2}=$$
$$-\frac{1}{2}[N+(-\sum_{j=0}^{N}\gamma_j-\sigma^{-2}(\boldsymbol{t}-\boldsymbol{\Phi\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi\mu}))].$$

令上式等于0, 整理后就得到了σ^2的迭代计算公式, 即文中的式(21):

$$\sigma^2=\frac{(\boldsymbol{t}-\boldsymbol{\Phi\mu})^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{t}-\boldsymbol{\Phi\mu})}{N-\sum_{j=0}^{N}\gamma_j}.$$

以上就是RRVM模型超参数$\boldsymbol{\alpha},\boldsymbol{\beta}$和$\sigma^2$迭代计算公式的详细推导过程.

作者简介:

韩 敏 (1959—), 女, 教授, 博士生导师, 研究方向为复杂工业过程建模与控制、神经网络、混沌时间序列分析、专家系统等, E-mail: minhan@dlut.edu.cn;

赵 耀 (1984—), 男, 硕士研究生, 研究方向为复杂工业过程建模与控制, E-mail: zhaoyao@mail.dlut.edu.cn;

杨溪林 (1958—), 男, 教授, 博士生导师, 研究方向为冶金生产自动化技术及自动化系统集成, E-mail: aritime007@126.com;

林 东 (1970—), 男, 高级工程师, 研究方向为炼钢技术及系统集成, E-mail: lindong1970@sina.com.