文章编号:1000-582X(2012)02-123-05

混合高斯过程回归模型在铁水硅含量预报中的应用

任江洪[1],陈　韬[2],曹长修[1]

(1.重庆大学 自动化学院,重庆市 400044;2.南洋理工大学 化学与生物医学工程学院,新加坡 637459)

摘　要:为了提高基于高斯过程回归的软测量模型的预测精度,提出了一种混合高斯过程回归模型。该模型将高斯过程回归模型预测输出值的方差及其分布作为主要考虑因素,对多个高斯过程回归模型的输出值进行组合输出,获得了比单个高斯过程回归模型更高的预测精度和更强的模型鲁棒性。将该模型实用于高炉铁水硅含量预报模型的建模,获得了比使用单个高斯过程回归模型建模时更好的应用效果。

关键词:高斯过程回归;Bootstrap;软传感器;参数估计;统计方法

中图分类号:TP212.6　　**文献标志码**:A

Composite gaussian process regression model and its application to prediction of silicon content in hot metal

REN Jiang-hong[1], *CHEN Tao*[2], *CAO Chang-xiu*[1]

(1. Chongqing University College of Automation, Chongqing 400044, P. R. China. 2. School of Chemical and Biomedical Engineering, Nanyang Technological University, Singapore 637459, Singapore)

Abstract: In order to increase the predictive precision of gaussian process regression based soft sensor, a composite gaussian process regression model is proposed. This model combines the outputs of several gaussian process models as the output according to the variances and the distribution of the outputs, which results in higher prediction accuracy and higher robustness than the single gaussian process model. The proposed composite gaussian process regression model is successfully applied to the prediction of silicon content in hot metal.

Key words: Gaussian process regression; Bootstrap; soft sensor; parameter estimation; statistic method

高斯过程回归最早是由 O'Hagan 作为一种人工神经网络的可选替代方法提出来的[1]。Neal 的研究指出一大类基于人工神经网络的贝叶斯回归模型[2],在神经元数量没有限制的情况下,都可看作是一种高斯过程回归模型。同时,高斯过程回归也可通过将高斯先验分布加于非参数回归函数空间,并推导预测目标的后验分布获得[3]。高斯过程回归模型的分析特征适合于作较深入的理论分析[4]。同时其在实际应用中也有较好的应用效果,常被应用于软传感器的建模[5-7]、过程控制参数的优化[8]等领域。

在实践中,笔者将高斯过程回归模型用于高炉铁水硅含量预报模型的建模。在高炉冶炼过程中,铁水硅含量是评定高炉炉况稳定性和生铁质量的重要指标。准确地预报铁水硅含量,有助于控制高炉热状态,保证高炉稳定运行。但硅含量是不可能从

收稿日期:2011-10-8

基金项目:重庆市科委自然科学基金资助项目(CSTC2008BB2324)

作者简介:任江洪(1976-),男,重庆大学博士,主要从事工业过程建模、机器学习、统计分析的研究,(Tel)13452351150;(E-mail)rjh@cqu.edu.cn。

生产过程中实时测量获得的,因此建模的关键就在于辨识易于测量的铁水量与硅含量之间的关系。那么推理模型在这里就相当于一个能够测量硅含量的软传感器[9-10]。

在建模过程中笔者发现,单个的高斯过程回归模型并不总能获得较好的预测效果。模型参数与预测性能常常受到训练数据微小变化与参数初始值的影响。受到 Bagging(bootstrap AGGregatING)方法[11-12]将多个模型进行组合,组合模型的鲁棒性和预测准确性都有提高的启示。笔者提出了一种通过 bootstrap 重采样方法[13]获得多个训练数据集,基于这些数据集建立多个高斯过程回归模型,并根据各个预测模型预测方差和预测均值分布概率,对多个模型的输出进行组合获得组合输出的混合高斯过程回归模型。

1 高斯过程回归

1.1 高斯过程回归模型

由 N 个数据点组成训练数据集$\{x_i,y_i\}$,$i=1,\cdots,N$,$x_i=(x_{i1},\cdots,x_{iD})$。设回归函数 $y(x)$符合如下均值为 0 的高斯先验分布

$$\boldsymbol{y}=(y_1,\cdots,y_N)^{\mathrm{T}}\sim N(0,\boldsymbol{C}_N),\tag{1}$$

其中 $\boldsymbol{C}_N$ 是 $N\times N$ 的协方差矩阵,其元素定义为 $\boldsymbol{C}_{ij}=\boldsymbol{C}(x_i,x_j)$。一个协方差函数的例子如下

$$\boldsymbol{C}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\theta_0+\theta_1\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j+\theta_2\exp\{-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\}+\delta_{ij}\sigma^2。\tag{2}$$

当 $i=j$ 时,$\delta_{ij}=1$,其他情况下 $\delta_{ij}=0$。$\boldsymbol{\theta}=\{\theta_0,\theta_1,\theta_2,\sigma^2\}$为协方差函数的超越参数。高斯过程回归是一种非参数回归方法,超越参数的意义和作用与参数回归中待估计的参数不同。超越参数必须为非负,以保证协方差矩阵的正定。θ_0 与 $\theta_1\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$ 分别表示 2 个数据间的常数偏差和线性相关程度。$\theta_2\exp\{-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\}$同径向基函数的形式相同,是对相近输入间的输出的潜在强相关性的考察。σ^2 反应的是随机误差的效果。将线性项和非线性项目都包括在协方差函数中,使高斯过程回归既能处理线性又能处理非线性的数据结构。

根据高斯过程的定义,对于新的输入 $\boldsymbol{x}_{N+1}$,其输出 y_{N+1}与训练数据中的 N 个输出值 y_i,$i=1,\cdots,N$ 服从联合高斯分布,即

$$p(\boldsymbol{y}_{N+1})=N(\boldsymbol{y}_{N+1}\mid 0,\boldsymbol{C}_{N+1})。\tag{3}$$

其中

$$\boldsymbol{C}_{N+1}=\begin{pmatrix}\boldsymbol{C}_N & \boldsymbol{k}\\ \boldsymbol{k}^{\mathrm{T}} & c\end{pmatrix}。\tag{4}$$

以上矩阵中 $\boldsymbol{k}=[\boldsymbol{C}(\boldsymbol{x}_{N+1},\boldsymbol{x}_1),\cdots,\boldsymbol{C}(\boldsymbol{x}_{N+1},\boldsymbol{x}_N)]^{\mathrm{T}}$,$c=\boldsymbol{C}(\boldsymbol{x}_{N+1},\boldsymbol{x}_{N+1})$。那么 $p(\boldsymbol{y}_{N+1}\mid\boldsymbol{y}_N)$也服从高斯分布,其均值和方差分别为

$$\hat{y}_{N+1}=\boldsymbol{k}^{\mathrm{T}}\boldsymbol{C}_N^{-1}\boldsymbol{y}_N;\tag{5}$$

$$\sigma^2(y_{N+1})=c-\boldsymbol{k}^{\mathrm{T}}\boldsymbol{C}_N^{-1}\boldsymbol{k}。\tag{6}$$

1.2 超越参数的学习

高斯过程回归模型中的超越参数,控制着模型对数据的相关性尺度、噪声等重要特征。对超越参数的学习可采用极大似然方法。以超越参数为条件的极大似然函数

$$\ln p(\boldsymbol{y}\mid\theta)=-\frac{1}{2}\ln|\boldsymbol{C}_N|-\frac{1}{2}\boldsymbol{y}\boldsymbol{C}_N^{-1}-\frac{N}{2}\ln(2\pi)。\tag{7}$$

将上式对超越参数中的每个参数求导有

$$\frac{\partial}{\partial\theta_i}\ln p(\boldsymbol{y}\mid\theta)=-\frac{1}{2}tr\left(\boldsymbol{C}_N^{-1}\frac{\partial\boldsymbol{C}_N}{\partial\theta_i}\right)+\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{C}_N^{-1}\frac{\partial\boldsymbol{C}_N}{\partial\theta_i}\boldsymbol{C}_N^{-1}\boldsymbol{y}。\tag{8}$$

如果将超越参数的先验概率加入到式(7)的似然函数中,那么可对超越参数作极大后验估计。也可使用马尔科夫链蒙特卡诺方法产生随机样本[4],来获得超越参数的后验分布。但是马尔科夫链蒙特卡诺方法的计算负担较大,在实践中采用该方法这是个需要考虑的问题。

在算法实现中,采用了共轭梯度下降方法[14]来寻找使极大似然函数取得最大值的超越参数。

2 混合高斯过程回归模型

为获得训练多个高斯过程模型所需的训练数据,采用了 bootstrap 重采样方法。它将现有数据视为对总体的 1 个估计,或者看作 1 个伪总体。然后从这些数据中进行随机的、有放回的重复采样,从而形成多个数据集。Bagging 方法在 bootstrap 重采样获得的多个训练数据集的基础上,训练多个预测模型,并将这些模型进行组合。组合模型可以用下式表述

$$\boldsymbol{y}(\boldsymbol{x})=\frac{1}{K}\sum_{k=1}^{K}y_k(\boldsymbol{x}),\tag{9}$$

基于以上 Bagging 方法的启示,用$\widehat{y^*}_k,\sigma_k^2$,$k=1,\cdots,K$ 分别表示第 k 个模型对于输入 $\boldsymbol{x}^*$ 的预测均值和预测方差,用 w_k 表示混合模型参数,并且满足 $\sum_{k=1}^{K}w_k=1$,那么混合高斯过程回归模型的基本形式如下

$$p(y^*\mid\boldsymbol{w})=\sum_{k=1}^{K}w_kN(y^*\mid\widehat{y^*}_k,\sigma_k^2),\tag{10}$$

其中的 $\boldsymbol{w}=\{w_1,\cdots,w_K\}$。模型的关键是找到适合的 $\boldsymbol{w}$。

根据高斯过程回归的定义,$y_1,\cdots,y_N,y_k^*$ 符合

联合高斯分布，并且由式(5)可知预测均值为训练数据中输出值的线性组合，那么预测均值在各个高斯过程回归模型采用相同协方差函数的情况下，各个模型的预测均值$\widehat{y^*}_k$，$k=1,\cdots,K$ 可以近似看作服从高斯联合分布(其中 K 是模型的数量)

$$(\widehat{y^*}_1,\cdots,\widehat{y^*}_K)\sim N(\widehat{y^*}_k\mid\mu_{\text{com}},\sigma^2_{\text{com}})。\quad(11)$$

使用矩估计方法来估计以上的 $\mu_{\text{com}},\sigma^2_{\text{com}}$ 有

$$\mu_{\text{com}}=\frac{1}{K}\sum_{k=1}^{K}\widehat{y^*}_k,\quad(12)$$

$$\sigma^2_{\text{com}}=\frac{1}{K-1}\sum_{k=1}^{K}(\widehat{y^*_k}-\mu_{\text{com}})^2。\quad(13)$$

$\widehat{y^*}_k$ 在式(11)给定的高斯分布的概率高低反应了其预测值与预测均值的趋近程度。概率越大的预测值对最终预报值有更大的贡献。同时各个预测模型的预测方差 σ_k^2 反映了模型对预测的不确定程度，那么对于预测确定性高的模型，即预测输出方差小的模型，理应获得更大的权值。将以上 2 个影响因素考虑在内，给出如下的公式来确定混合高斯过程回归模型的参数 $\boldsymbol{w}$

$$\boldsymbol{w}_k=\frac{\frac{1}{\sigma_k^2}N(\widehat{y^*}_k\mid\mu_{\text{com}},\sigma^2_{\text{com}})}{\sum_{i=1}^{K}\frac{1}{\sigma_i^2}N(\widehat{y^*}_i\mid\mu_{\text{com}},\sigma^2_{\text{com}})}。\quad(14)$$

传统的 Bagging 方法以确定性的方式对多个模型进行组合，而以上策略是根据多个模型的输出自适应地确定组合参数，因此参数能更好地反应各个模型的变化情况，从而提高了组合模型的整体适应性。

当然回归预测的主要目标是在新的输入情况下获得相应的输出值，以式(10)的数学期望作为预测输出

$$\overline{y^*}=E[p(y^*\mid\boldsymbol{w})]=\sum_{k=1}^{K}w_kE[N(y^*\mid\widehat{y^*}_k,\sigma_k^2)]$$

$$=\sum_{k=1}^{K}w_k\widehat{y^*}_k。\quad(14)$$

相应的方差为

$$\text{cov}[p(y^*\mid\boldsymbol{w})]=\sum_{k=1}^{K}w_k^2\text{cov}[N(y^*\mid\widehat{y^*}_k,\sigma_k^2)]$$

$$=\sum_{k=1}^{K}w_k^2\sigma_k^2。\quad(15)$$

该方差小于任何单个预测模型的预测方差，也即混合高斯过程回归模型相对于单个高斯过程回归模型有更好的模型鲁棒性。后面的实验数据分析将进一步说明这一点。

3　实验数据分析

在这一部分中，将混合高斯过程回归模型用于某工厂高炉铁水硅含量的预报。铁水的硅含量是影响铁水质量的重要因素，能准确及时地获得铁水硅含量参数对提高整个生产品质有很大帮助。但传统的通过实验室检测的方法，数据获取时间很长(2 h 或者以上)，最后获得的数据已失去了优化生产过程、提高生产质量的意义。因此需要一种对铁水的硅含量进行在线预报的方法。铁水硅含量可以通过对生产过程建立准确的机理模型，基于容易获取的生产数据，对其进行推理预测获得。但是机理模型的获得非常困难，在现阶段的技术条件下并不可行。因此只能通过基于数据驱动的建模方法来建立高炉铁水硅含量预报的数学模型。混合高斯过程回归模型就是这样一种方法。

影响铁水硅含量的因素很多，并且各种因素之间也存在着错综复杂的交互影响。众多影响因素导致高炉炉温变化具有复杂性、非线性、高维数等特点[15]。考虑到实际数据的采集条件及与铁水硅含量之间的相关程度，选取了 8 个主要因素作为模型的输入变量：透气性指数、全压差、风压、风量、风温、炉顶煤气温度、喷煤量、以及上一炉铁水硅含量。这些变量的单位和数值各不相同，数量级差别较大，为了不使输入的样本因为过大或者过小而使其他的样本分量失去调控作用，将所有输入数据都处理为零均值，单位方差的量。

从工厂生产数据库获取历史数据，在清除异常点等无效数据后共获得了 240 个有效数据点。为避免模型对数据出现过度拟合的情况，随机地将有效数据点分成训练数据和测试数据。

在图 1 所涉及的实验中，通过随机划分获得训练数据(200 个)和测试数据(40 个)，并对训练数据采用 bootstrap 方法进行重采样，获得 30 组训练数据，每组 200 个数据。在这 30 组训练数据上分别建立单个的高斯过程回归模型，其相应的协方差函数为式(2)所示，并采用共轭梯度下降的方法获取相应的超越参数值。这 30 个高斯过程回归模型在测试值上的预报误差，用均方根误差(root mean square error，RMSE)表示，如图 1 中的直方图所示。基于这 30 个高斯过程回归模型，采用提出的方法建立混合高斯过程回归模型，其相应的预报误差也用均方根误差表示，如图 1 中的直线所示。图 1 中 30 个模型的预测误差出现了较大的波动，反应出高斯过程回归模型的预测精度受模型数据和初始参数的影响

较大。但在这种情况下，混合高斯过程回归模型(RMSE:0.031，图1中横线所示)获得了比单个高斯过程回归模型中最优的模型(RMSE:0.034)更高的预测精度。从另一方面来看，也反映出混合高斯过程回归模型受初始参数等的影响较小，模型的鲁棒形更强。

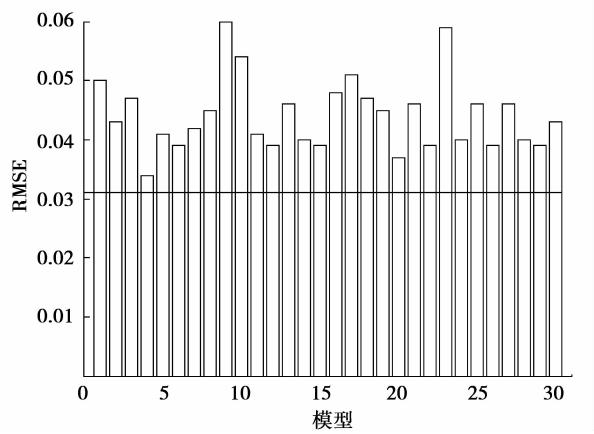

图1 单个高斯过程回归模型与混合高斯过程

模型预测精度对比

在图2的实验中，按照以上所述的随机划分方法对历史数据进行40次随机划分。在每次获得的训练数据上，按照图1的实验方法建立30个单独的高斯过程回归模型，以及混合高斯过程回归模型，并记录单个高斯过程回归模型中对测试数据的最小预测误差(RMSE)和混合高斯过程回归模型对测试数据的预测误差(RMSE)。40次重复实验获得了两组数据(每组40个)，见图2中的折线1和折线2，分别为最小单个高斯过程回归模型的最小预测误差和混合高斯过程回归模型的预测误差。该实验的目的是重点考察混合高斯过程回归模型的鲁棒性。由图可见折线1表现出了比折线2更大的波动性，也即相对于单个高斯过程回归模型，相对于不同的训练数据，混合高斯过程回归模型的预测值波动性明显要小。所以混合高斯过程回归模型有比单个高斯过程模型更强的鲁棒性。

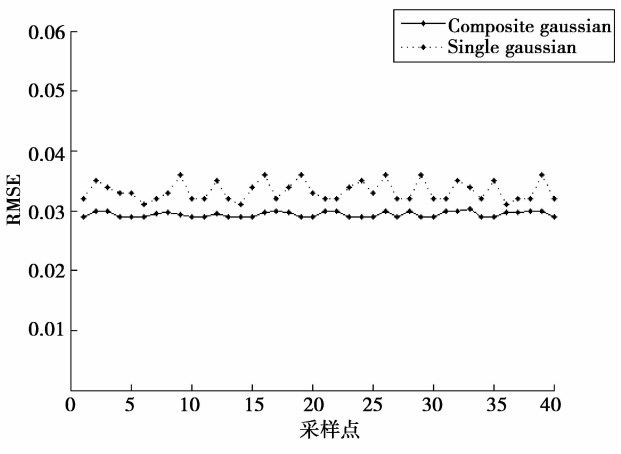

图2 最优的单个模型预测误差和混合模型的预测误差

图3的实验同图2的实验相同，对历史数据进行40次随机划分。在每次划分数据的基础上建立30个高斯过程回归模型。依次以1～30个单独高斯过程回归模型为基础建立30个基数不同的混合高斯过程回归模型，并取得这些混合高斯过程回归模型对测试数据的预测误差(RMSE)。40次实验后，不同基数的混合高斯过程回归模型对测试数据的预测误差平均值见图3。该实验目的是讨论混合高斯过程回归模型的预测误差，相对于不同基数的预测误差是否会收敛。高斯从图3可以看出，在混合高斯过程回归模型由8个及以上的单个高斯过程回归模型组成时，对测试数据的预测误差将稳定到一个基本恒定的水平上。因此在建立最终的铁水硅含量预报模型时，采用了8个单独的高斯过程回归模型来组成混合高斯过程回归模型。

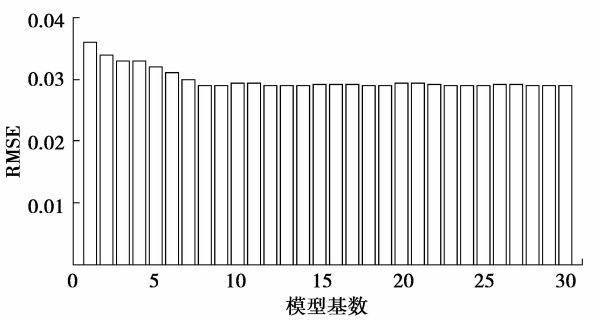

图3 在不同模型基数下的混合模型的RMSE误差

利用Matlab作为建模工具，在Intel酷睿2双核2.2GHz处理器、2G内存的硬件平台，Ubuntu Linux8.10的软件平台上，选择8个高斯过程回归模型来组成混合高斯过程回归模型的情况下，30次重复实验所得的混合高斯过程回归模型的模型训练平均时间为38.3 s。这个训练时间能够满足实际建模的需要。在训练数据集更大的情况下，高斯过程的协方差矩阵的求逆运算的运算成本会大大增加。这时可采用高斯过程的稀疏训练算法来降低模型的训练时间。笔者也在尝试使用Matlab的并行计算工具箱，通过利用处理器的多核能力来进一步提高模型的训练时间。

4 总　结

提出了一种混合高斯过程回归模型：该模型通过 bootstrap 方法对训练数据进行采样获得多个训练数据集，在这些数据集上建立多个高斯过程回归模型；在综合考虑高斯过程回归模型预测输出值的分布及各个模型的预测方差的基础上，建立了一种自适应的确定模型组合参数的策略；基于该策略获取的模型组合参数，对多个高斯过程回归模型的输出进行组合，获得了比单个模型更高的预测精度，并且组合模型比单个高斯过程回归模型的鲁棒性更强。

将该模型应用于铁水硅含量预报应用，获得了比单个高斯过程回归模型更好的预报效果。同时混合高斯过程回归模型也是一种贝叶斯方法，能够给出预测的分布概率。这对于将预测结果用于其他统计推理应用，例如故障诊断、参数优化，有很大的实用价值。该方法对在实践中应用高斯过程回归模型有一定的借鉴意义。

参考文献：

[1] OHAGAN A. Curve fitting and optimal design for prediction [J]. Journal of the Royal Statistical Society Series B: Methodological, 1978, 40(1):1-42.

[2] RADFORD M N. Bayesian learning for neural networks[M]. New York: Springer, 1996.

[3] BISHOP C M. Pattern recognition and machine learning[M]. New York: Springer, 2006.

[4] RASMUSSEN C E, WILLIAMS C K I. Gaussian processes for machine learning[M]. Cambridge: MIT Press, 2006.

[5] CHEN T, MORRIS J, MARTIN E. Gaussian processes regression for multivariate spectroscopic calibration [J]. Chemometrics and Intelligent Laboratory Systems, 2007, 87:59-67.

[6] 熊志化，黄国宏，邵惠鹤. 基于高斯过程和支持向量机的软测量建模比较及应用研究[J]. 信息与控制，2004，33(6)：754-757.
XIONG ZHI-HUA, HUANG GUO-HONG, SHAO HUI-HE. Comparison and application research on soft sensor modeling based on gaussian processes and support vector machines[J]. Information and Control, 2004, 33(6):754-757.

[7] 潘立登，李大宇，马俊英. 软测量技术原理与应用[M]. 北京：中国电力出版社，2009.

[8] YUAN J, WANG K, YU T, et al. Reliable multi-objective optimization of highspeed WEDM process based on Gaussian process regression [J]. International Journal of Machine Tools and Manufacture, 2008, 48(1):47-60.

[9] FORTUNA L, GRAZIANI S, RIZZO A, et al. Soft sensors for monitoring and control of industrial processes [M]. London: Springer, 2007.

[10] 严爱军，岳恒，赵大勇，等. 一类复杂工业过程的智能预报模型及其应用[J]. 控制与决策，2005，20(7)：794-797.
YAN AI-JUN, YUE HENG, ZHAO DA-YONG, et al. Intelligent prediction model for a class of complex industrial process and its application[J]. Control and Decision, 2005, 20(7):794-797.

[11] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(2):123-140.

[12] CHEN T, REN J H. Bagging for Gaussian process regression [J]. Neurocomputing, 2009, 72 (7-9): 1605-1610.

[13] EFRON B. Bootstrap methods: another look at the jackknife[J]. Annals of Statistics, 1979, 7(1):1-26.

[14] HAYKIN S. Neural networks and learning machine (3rd edition)[M]. New Jersey: Prentice Hall, 2008.

[15] TANG X L, REN J H, ZHUANG L, et al. Application of neural network trained by chaos particle swarm optimization to prediction of silicon content in hot metal [C]//Proceedings of the 7th World Congress on Intelligent Control and Automation, June 25-27, 2008. Chongqing, China: IEEE, 2008: 2446-2449.

（编辑　侯　湘）