基于 AR 模型思想的高斯过程多模型建模方法*

邓卫卫[1]，杨慧中[1,2]

(1. 江南大学 教育部轻工过程先进控制重点实验室，江苏 无锡 214122；2. 上海市电站自动化技术重点实验室，上海 200072)

摘 要：针对 K-近邻算法中难以确定 K 值的定量问题，提出一种基于 AR 模型思想的高斯过程多模型建模方法。该方法借鉴 AR 模型的思想，将前一时刻的输出值作为当前时刻输出值的一个影响因素放入输入集中，通过计算训练样本的平均最小距离从而得到一个搜索半径，根据搜索半径来确定 K 值和 K 个近邻样本的权重，采用加权输出的方式以得到组合模型的输出。将其建模方法应用到某双酚 A 反应釜出口苯酚含量的软测量建模中，仿真结果表明，该方法具有较高的精度和较好的模型推广能力。

关键词：K-近邻算法；AR 模型；高斯过程；多模型

中图分类号：TP273 **文献标志码**：A **文章编号**：1001-3695(2012)05-1628-03

doi:10.3969/j.issn.1001-3695.2012.05.006

Gaussian process multi-model modeling method based on AR model

DENG Wei-wei[1], YANG Hui-zhong[1,2]

(1. *Key Laboratory of Advanced Process Control for Light Industry of Ministry of Education, Jiangnan University, Wuxi Jiangsu* 214122, *China*; 2. *Shanghai Key Laboratory of Power Station Automation Technology, Shanghai* 200072, *China*)

Abstract: The value of K is difficult to be exactly determined in K-nearest neighbor algorithm. This paper proposed a Gaussian process multi-model modeling method based on the idea of AR model. The method introduced the model output value of previous moments into the input set of the current moment, calculated the mean minimum distance of the training samples to get a search radius. And it determined the value of K according to the radius and calculated the weights of the output according to the K neighbor samples. Finally it took the weighted output mode to get the output of combinational model. The method was used for the soft-sensor model to estimate the content of phenol at the outlet of a reaction vessel in a Bisphenol A production process. The simulation results show that the method has a higher accuracy and better model generalization ability.

Key words: K-nearest neighbor algorithm(KNN); AR model; Gaussian process; multi-models

流程工业往往涉及到复杂物系的加工和处理过程，其中有许多难以测量或者无法测量的重要质量变量。软测量技术作为解决这个问题的一种方法，已经在工业场合得到了广泛的应用。在双酚 A 缩聚反应过程中，由于受催化剂活性波动以及其他因素的影响，工况经常发生较大的变化，若采用单一模型描述这种复杂非线性的过程特性，往往会导致精度和泛化性能差等缺点，因此宜采用多模型建模的软测量估计方法[1]。

K-近邻(KNN)算法是一种基于实例学习的方法，它具有简单、有效和高鲁棒性等特点，广泛应用于文本分类、机器学习和数据挖掘等领域，以解决分类及多模型建模等问题[2]。在 KNN 算法中，K 值的确定一直是广大学者研究的内容之一，Xie 等人[3]提出用朴素贝叶斯方法来选择近邻(SNNB)，Jiang 等人[4]提出动态 K-近邻朴素贝叶斯属性加权(DKNAW)方法。这些方法提供了选择 K 值的依据，但其没有考虑到样本的分布特性，当样本数据较为密集或者较为稀疏时可能无法得到较好的结果。本文考虑到样本分布的复杂性，将平均最小距离引入 KNN 算法中，根据测试样本与各训练样本的距离来确定具体的 K 值，然后计算各子模型的输出权重，由子模型加权得到最终输出值；同时由于工业连续生产过程中的操作变量一般不会发生大突变，前后时刻的输出值之间存在连续性，因此考虑将 AR 模型的思想应用到建模过程中，将模型前一时刻的输出值引入到后一时刻的输入集，由高斯过程建模方法建立各子模型。仿真结果表明了该方法的有效性。

1 基本原理

1.1 AR 模型

时间序列分析是统计学科的一个重要分支内容，它根据观测数据的特点为数据建立尽可能合理的统计模型，在信号处理、经济管理等领域得到了广泛应用[5]。AR 模型是时间序列分析中的一种简单而又常用的模型。

AR 模型[6,7]的表达式为

$$y(t)=a_1y(t-1)+a_2y(t-2)+\cdots+a_ny(t-n)+\xi(t) \tag{1}$$

其中：$y(t)$ 为输出变量；a_i 为自回归参数，表示 $t-i$ 时刻的输出值对 t 时刻输出值的影响程度；$\xi(t)$ 是均值为零的白噪声时间序列。

收稿日期：2011-10-30；**修回日期**：2011-12-08 **基金项目**：国家自然科学基金资助项目(60674092)；江苏省高技术研究项目(工业部分)(BG2006010)；上海市科学技术委员会资助项目(09DZ2273400)

作者简介：邓卫卫(1986-)，男，福建龙岩人，硕士研究生，主要研究方向为数据挖掘与工业过程建模(dww081@163.com)；杨慧中(1955-)，女，教授，博导，主要研究方向为工业过程建模与优化控制.

1.2 K-近邻算法

KNN 算法[8,9]的基本思想是将给定的待分类样本集和带类标签的训练样本集,通过计算新样本与各训练样本之间的距离,找出距离该新样本最近的 K 个近邻样本;根据这些近邻样本所属的类别来判定新样本的类别,将新样本分配到 K 个近邻样本的最公共类别中。在 KNN 算法中,K 值一般都是由经验确定的,需要通过不断调整 K 值寻找最佳的分类结果。

1.3 高斯过程建模方法

假设给定训练样本 $D=\{(x_i,y_i)\}(i=1,\cdots,n)$,其中,$x_i\in R^d$,$y_i\in R$,$n$ 为训练样本数,d 表示输入向量的维数。对于测试集中的新样本 x,预测分布就是 n 个训练样本的输出与该新样本之间形成的 $n+1$ 维联合高斯分布。高斯过程(Gaussian process, GP)模型预测的均值为

$$\hat{y}(x)=\boldsymbol{k}^{\mathrm{T}}(x)\boldsymbol{K}^{-1}\boldsymbol{y} \tag{2}$$

预测值的方差为

$$\sigma_{\hat{y}}^2(x)=k(x,x)-\boldsymbol{k}^{\mathrm{T}}(x)\boldsymbol{K}^{-1}\boldsymbol{k}(x) \tag{3}$$

其中:$\boldsymbol{k}(x)$为测试输入和训练样本输入值之间的 $n\times 1$ 维协方差向量,$\boldsymbol{k}(x)=[C(x,x_i)]^{\mathrm{T}}(i=1,\cdots,n)$,$C(x_i,x_j)$表示协方差函数;$\boldsymbol{K}$为 $n\times n$ 维训练样本间的协方差矩阵,$K_{ij}=C(x_i,x_j)$;$k(x,x)$为测试样本的输入和其自身的协方差;$y=[y_1,\cdots,y_n]^{\mathrm{T}}$。

GP 模型[10]的建立首先要进行模型的选择,即确定协方差函数。选择协方差函数的前提是要保证对于任一输入都能够产生一个对称正半定的协方差矩阵,同时希望相邻的输入产生相邻的输出。常用的协方差函数有平稳协方差函数和非平稳协方差函数,模型选择中要求协方差函数连续并且可导,因此通常采用径向基函数作为协方差函数:

$$C(x_i,x_j)=\upsilon_0\exp\left[-\frac{1}{2}\sum_{k=1}^{d}\omega_k(x_{ik}-x_{jk})^2\right]+\upsilon_1\delta_{ij} \tag{4}$$

该函数由两部分组成:前一部分表示相邻的输入有着高度相关的输出,其中 υ_0 表示先验知识的总体度量;后一部分为数据中的噪声,其中 υ_1 表示服从高斯分布的噪声方差,δ_{ij}是 Kronecker 算子。

协方差函数中的超参数 $\theta=(\upsilon_0,\omega_1,\cdots,\omega_d,\upsilon_1)$采用极大似然法来获得最优超参数[11]。对于上述给定的协方差函数,其超参数的对数似然函数为

$$L(\theta)=-\frac{1}{2}\lg(|\boldsymbol{K}|)-\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{y}-\frac{n}{2}\lg(2\pi) \tag{5}$$

其中,$|\boldsymbol{K}|$表示 $\boldsymbol{K}$的行列式。通过不断调整超参数 θ 使训练样本的对数似然函数最大。优化过程中需要计算似然函数 $L(\theta)$ 对各参数的偏导数:

$$\frac{\partial L}{\partial\theta_i}=-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}^{-1}\frac{\partial\boldsymbol{K}}{\partial\theta_i}\right)+\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}^{-1}\frac{\partial\boldsymbol{K}}{\partial\theta_i}\boldsymbol{K}^{-1}\boldsymbol{y}=\frac{1}{2}\mathrm{tr}\left\{\left[(\boldsymbol{K}^{-1}\boldsymbol{y})(\boldsymbol{K}^{-1}\boldsymbol{y})^{\mathrm{T}}-\boldsymbol{K}^{-1}\right]\frac{\partial\boldsymbol{K}}{\partial\theta_i}\right\} \tag{6}$$

令$\frac{\partial L}{\partial\theta_i}=0,i=1,\cdots,d+2$,通过求解方程组即可得到超参数的极大似然估计值 $\hat{\theta}$。在获得最优的超参数后,对于测试集中的样本 x,可以将它们代入式(2)(3)中得到预测的均值 $\hat{y}(x)$ 和方差 $\sigma_{\hat{y}}^2(x)$。

2 基于 AR 模型思想的高斯过程多模型建模方法

2.1 基于 AR 模型思想的建模方法

在 AR 模型的表达式(1)中,如果将步长 n 定为 1,则表达式变为

$$y(t)=a_1y(t-1)+\xi(t) \tag{7}$$

如果将 $a_1y(t-1)$看成是前一时刻输出值对当前时刻输出值的影响,将 $\xi(t)$看成输入变量对当前时刻输出值的影响,则可以考虑将这种思路引入软测量建模中,将前一时刻的输出值引入当前时刻的输入集中作为影响当前时刻输出值的一维输入变量。这种建模方法由于考虑到了前一时刻输出值对当前时刻输出值的影响,可以使建立的模型具有更高的精度。

2.2 基于 MMD 的 K-近邻算法

KNN 算法中 K 值的选取对分类的结果有很大的影响,也会对后续的建模过程产生一定的影响。本文根据平均最小距离(mean minimum distance,MMD)来确定 K 值的取值,在样本分布密集或者稀疏的情况下均能得到较为精确的结果。

MMD 是衡量相似性测度的一个量,即在 d 维空间中,给定具有 n 个数据对象的集合 $A:\{X_1,\cdots,X_n\}$,其中 $X_i=(x_{i1},\cdots,x_{id})$。MMD 是该集合中数据对象到其最近的邻居点的平均距离,定义如下[12]:

$$\mathrm{MMD}=\frac{1}{n}\sum_{i=1}^{n}\min_{j\neq i}\left[\left(\sum_{k=1}^{d}(x_{ik}-x_{jk})^2\right)^{\frac{1}{2}}\right] \tag{8}$$

令搜索半径 r 的值为求取的 MMD 的 a 倍,某一新数据点 X 到集合 A 中各元素的距离用 $d_i(i=1,\cdots,n)$来表示。若 $d_i<r$,则将该元素挑选出来,挑选出来的元素个数即为该新数据点对应的 K 值。

2.3 基于 AR 模型思想的高斯过程多模型建模方法

结合 AR 模型的建模思想和 MMD 的 KNN 算法,将输出序列中 $t-1$ 时刻的输出值引入 t 时刻的输入中,输入变量由原来的 d 维变成 $d+1$ 维。用简单的聚类算法将训练样本分为 m 类,并且用高斯过程建模方法建立 m 个子模型。对测试样本使用 KNN 算法,寻找在半径 r 内距离测试样本最近的 K 个训练样本,然后分别计算属于每个类别的距离的倒数和以及 K 个距离的倒数总和,最后将这两个距离倒数和的比值作为各子模型的输出权重,经过加权组合得到模型的最终输出值。由于距离近的样本具有更大的相似性,应该赋予更大的权重,所以将距离的倒数作为输出权重。

具体步骤如下:

a)借鉴 AR 模型的思想,将样本输出序列中 $t-1$ 时刻的输出值放入 t 时刻的输入中一同作为输入变量,样本的输入变量由 d 维增加到 $d+1$ 维。

b)将输入样本进行归一化处理,新增加的输出序列也同样进行归一化处理,然后将样本数据分为训练集和测试集。

c)采用 K-均值聚类算法将训练样本分为 m 类,对训练样本标记对应的类标签,并分别用高斯过程方法训练各类样本得到 m 个子模型。

d)根据式(8)计算训练样本的 MMD。

e)对于新的测试样本,取 $r=a\times\mathrm{MMD}$,计算该测试样本与各训练样本之间的距离 $d_i(i=1,\cdots,n)$,并与 r 进行比较。若 $d_i<r$,则将该训练样本挑选出来。

f)对于挑选出来的 K 个训练样本,分别查看其类标签,并计算各类别中距离的倒数和 D_i,计算各权重系数 $w_i=D_i/D$,其

中 $D=\sum_{i=1}^{m}D_i$。

g)将测试样本送入各子模型,得到各子模型的输出 Y_i,然后分别乘上各权重系数,即得到最终的模型输出值 $Y=\sum_{i=1}^{m}w_i\cdot Y_i$。

相应的模型结构如图 1 所示。

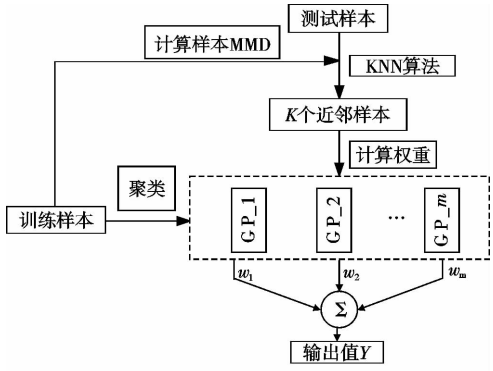

图1 多模型结构

3 仿真实例

双酚 A 生产过程较为复杂,采集的生产数据有的在某些工况点附近聚集,有的则分布较为分散。因此将本文提出的多模型建模方法应用于该生产过程的质量指标软测量建模中,实现对反应釜出口苯酚含量的实时估计。根据对生产过程的工艺和现场条件的分析,选择反应釜入口丙酮进料流量、苯酚进料流量、入口温度以及反应釜上的温度测量点 A 和温度测量点 F 这五个操作变量作为输入变量,反应釜出口的苯酚含量作为输出变量。首先对样本数据进行异常样本数据的剔除,然后对各输入变量进行归一化处理以消除变量间量纲不同对建模效果的影响,由此得到 451 组样本数据,取其中的 394 组数据进行训练建立模型,剩下的 57 组数据作为测试样本检验所建模型的效果。按照本文方法建立多模型软测量模型,将训练样本分为三类,分别建立三个高斯过程子模型,计算出训练样本的 MMD,对测试样本分别计算相应的输出权重,由各子模型输出加权组合得到最终的输出估计值。

为了验证本文方法的有效性,还采用单模型建模方法、传统的多模型建模方法建立模型,并与本文方法进行比较。其中,传统的多模型方法是对测试样本直接使用 KNN 算法,根据挑选出来的 K 个近邻样本的共同类别来确定该测试样本的类别,然后将测试样本送入对应的子模型得到输出估计值。使用均方根误差(RMSE)、平均相对误差(MRE)和最大相对误差(MAXE)来评价模型的性能。它们的定义分别如下:

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y(x_i))^2} \tag{9}$$

$$\mathrm{MRE}=\mathrm{mean}\left(\sum_{i=1}^{n}\frac{|f(x_i)-y(x_i)|}{y(x_i)}\right) \tag{10}$$

$$\mathrm{MAXE}=\max_{i=1}^{n}\left(\frac{|f(x_i)-y(x_i)|}{y(x_i)}\right) \tag{11}$$

其中,$f(x_i)$ 和 $y(x_i)$ 分别为模型的输出值和真实值。

表 1 分别列出了三种方法的测试误差。

表 1 模型测试结果比较

方法	RMSE	MRE	MAXE
单模型方法	0.877 6	0.009 5	0.033 4
传统多模型方法	0.763 7	0.008 5	0.028 5
本文方法	0.613 0	0.007 0	0.021 3

从表 1 中可以看出,无论是均方根误差、平均相对误差还是最大相对误差,本文方法得到的模型误差均较小。三种方法的模型测试结果如图 2 所示。

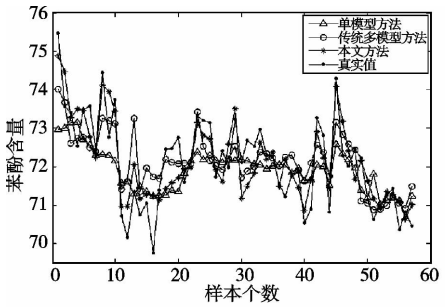

图2 双酚A生产过程中反应釜出口苯酚含量的模型估计值和真实值

4 结束语

针对 KNN 算法中 K 值较难确定对最终结果的影响问题,本文将 MMD 引入 KNN 算法中,根据距离来确定具体的 K 值,同时将 AR 模型的思想应用到建模过程中。将该方法用于实现双酚 A 生产过程中反应釜出口苯酚含量的软测量建模中,仿真结果表明,该方法建立的模型具有较高的精度和较好的跟踪效果。

参考文献:

[1] 陈定三, 杨慧中. 粗糙分类器的多模型软测量建模方法[J]. 计算机与应用化学, 2010,27(4):457-460.

[2] 王龙, 李晓光, 钟绍春. 基于 K 近邻法及移动 agent 技术的垃圾邮件检测系统研究[J]. 计算机应用研究, 2009,26(7): 2630-2632.

[3] XIE Zhi-peng, HSU W, LIU Zong-tian, *et al*. SNNB: a selective neighborhood based naive Bayes for lazy learning[C]//Proc of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. 2002:104-114.

[4] JIANG Liang-xiao, ZHANG H, CAI Zhi-hua. Dynamic K-nearest-neighbor naive Bayes with attribute weighted[C]//Proc of the 3rd International Conference on Fuzzy Systems and Knowledge Discovery. Berlin:Spinger,2006:365-368.

[5] 刘建. 汽车道路谱的 AR 模型重构方法研究及实现[D]. 镇江: 江苏大学, 2010.

[6] CHISCI L, MAVINO A, PERFERI G, *et al*. Real-time epileptic seizure prediction using AR models and support vector machines[J]. IEEE Trans on Biomedical Engineering, 2010,57(5): 1124-1132.

[7] 张伟, 胡昌华, 焦李成. 最小二乘 AR 模型的惯性器件故障预测[J]. 仪器仪表学报, 2006,27(6):1755-1757.

[8] WANG Ling, FU Dong-mei. Estimation of missing values using a weighted K- nearest neighbors algorithm[C]//Proc of International Conference on Environmental Science and Information Application Technology. 2009:660-663.

[9] 李欢, 焦建民. 简化的粒子群优化快速 KNN 分类算法[J]. 计算机工程与应用, 2008,44(32): 57-59.

[10] 申倩倩, 孙宗海. 基于自适应自然梯度法的在线高斯过程建模[J]. 计算机应用研究, 2011, 28(1): 95-97.

[11] 王华忠. 高斯过程及其在软测量建模中的应用[J]. 化工学报, 2007,58(11): 2840-2845.

[12] 罗健旭, 邵惠鹤. 软测量建模数据的过失误差侦破——一种基于聚类分析的方法[J]. 仪器仪表学报, 2005,26(3): 238-241.